Relative importance analysis for count regression models

Joseph N. Luchman

Fors Marsh

Author Note

Joseph N. Luchman https://orcid.org/0000-0002-8886-9717. The author has read and approved the final version of the manuscript, has no conflicts of interest with respect to his authorship or the publication of this article, and did not receive or benefit from fundings of this article.

Abstract

Count variables are common in behavioral science as an outcome. Count regression models, such as Poisson regression, are recommended when analyzing count variables but can be challenging to interpret given their non-linear functional form. I recommend relative importance analysis as a method to use in interpreting count regression model results. This work extends on past research by describing an approach to determining the importance of independent variables in count regression models using dominance analysis (DA). In this manuscript, I review DA as a relative importance method, recommend a pseudo-R^2 to use with count regression model-based DA, and outline the results of an analysis with simulated data that uses the recommended methodology. This work contributes to the literature by extending DA to count regression models and provides a thoroughly documented example analysis that researchers can use to implement the methodology in their research.

Keywords: Dominance Analysis, Relative Importance, Poisson Regression, R-square, Negative Binomial Regression, Count Data

Relative importance analysis for count regression models

Behavioral scientists often use how many times a behavior is observed as an outcome to answer research questions. Such behavior counts arise from many different sources among aggreates or collectives of people and individuals over time. Aggregate behavior counts used in the literature include the number of organizations adopting a specific practice in a week (Naumovska et al., 2021) and number of divestitures organizations make in a year (Bettinazzi & Feldman, 2021). Individual-level behavior counts used in the literature include the number of scientific articles published in a year among scholars (Rotolo & Messeni Petruzzelli, 2013) or number of errors that resulted in an accident in the last three months among medical doctors (Naveh et al., 2015). Behavior counts such as the above examples are valuable outcomes given that behavioral science concepts are often defined in terms of behavior (e.g., job performance; Motowidlo, 2003) and strategies to validate outcomes often use observable behavior as an outcome (e.g., criterion-oriented validity; Cronbach & Meehl, 1955).

However valuable as outcomes, behavior counts as a dependent variable in data analysis require the use of specialized tools. The recommended data analysis strategy with behavior counts uses regression models designed for non-negative integer or count distributions (e.g., Blevins et al., 2015). These count regression models include the Poisson regression and negative Binomial regression models. Both Poisson regression and negative Binomial regression are commonly used in the behavioral science literature and implemented in many data analytic software environments.

Count regression models are generalized linear models that transform the predictive equation to ensure that predicted values stay in the range of the depemdent variable. Count regression models use an exponential or log-linear transformation that has the form $y = e^{\beta X}$. Thus, the predictive equation parameterized by β requires back-transformation using a natural logarithm in order to obtain a predicted count value. This log-linear transformation ensures that the linear predicted values can take on any real number yet, when back-transformed into predicted counts, will have a lower bound of 0.

Count regression model results are challenging to interpret directly relative to linear regression. This is because count regression model coefficients describe how the natural logarithm of the dependent variable changes given a 1 unit change to an independent variable. When back-translated through an exponential function, count regression model coefficients are known as incidence rate ratios (IRRs) and describe the percentage change in the dependent variable per unit change to the independent variable. Because IRRs describe percentage change, count regression model coefficients produce predicted values that are relative or change in their magnitude over the continuum of the dependent variable. For example, a count regression model coefficient does not differentiate between a change from 1 predicted behavior to 2 and 5 predicted behaviors to 10. Despite the noteworthy difference in the absolute number of behaviors in both examples, each describe a 100% increase.

Model post-estimation methods such as graphing estimated marginal means are useful interpretive tools for log-linear count regression models as they help to contextualize the count regression models' predicted values. Another increasingly common model post-estimation tool used to contextualize model predictions is relative importance analysis (Tonidandel & LeBreton, 2011). Relative importance analysis is used to compare how each independent variable in the model contributes to a model fit metric such as the R^2 and is commonly implemented using dominance analysis (DA; Azen & Budescu, 2003).

Published methodological work has extended DA from the linear regression model on which it was originally developed, to other linear models including binary (Azen & Traxel, 2009), ordered, and multinomial logit models (Luchman, 2014). This work extends DA to count regression models and, in so doing, makes two contributions to the literature. This work first reviews DA as a relative importance methodology, recommends using a specific pseudo-R^2 statistic for count regression model-based DA, and implements a data analytic example of DA.

Second, this paper extends on the work of Blevins, Tsang, and Spain (2015) who review and offer multiple recommendations for the application of count regression models to research questions in behavioral science. Blevins et al. describe model and analytic details about the Poisson regression and negative Binomial regression models and provide a flowchart that researchers can use to identify which count regression model might be best to choose for their data analysis. In this work, I extend on their review to add an in-depth discussion of DA and its role as a post-estimation methodology. DA extends on the interpretation of coefficients to describe how the coefficients, when applied to the observed data, improve model-to-data fit in predicting the count dependent variable.

I begin the manuscript with a discussion of the conceptual background of DA. The conceptual discussion of DA focuses on outlining three different levels of dominance between independent variables, how these levels of dominance are determined in the data, and what each level of dominance means in terms of independent variable importance. The next section recommends a fit metric to apply to count regression models for the purpose of determinining importance using DA. In this count regression model focused section, I draw parallels between the explained variance R^2 used by the linear regression model as a fit metric and an analogous fit metric for count regression models that is particularly useful for DA. Finally, I describe an extensive data analytic example. This final section uses simulated data to estimate a Poisson regression model and a Poisson regression model-based DA where the methods described in the previous two sections are applied.

Dominance Analysis

Behavioral scientists have used many methods over the years to determine how important an independent variable is in a linear regression model (see reviews in Grömping, 2007; Johnson & LeBreton, 2004). Methods for determining the importance of an independent variable have ranged from the use of a correlation coefficient between the independent variable and dependent variable, to the independent variable's standardized regression coefficient, to the increment the independent variable makes to the R^2 over and above other independent variables. These methods, however informative in specific circumstances, make assumptions about an independent variable's contribution to prediction. These methods assume in some cases that an independent variable's contribution should not be adjusted for other independent variables (i.e., the correlation coefficient) and in other cases that the adjustment should occur only after all other independent variables are included (i.e., incremental R^2, standardized coefficient). In most situations, independent variable inclusion ordering is arbitrary and as opposed to choosing one approach or the other, recommended importance methods should account for different bi- and multivariate relationships with the dependent variable (Johnson & LeBreton, 2004). The most conceptually useful importance methods then produce determinations that are independent of independent variable inclusion order.

DA is a method for determining independent variable importance independent of inclusion order that was originally developed for the linear regression model by Budescu (1993). DA extended on previously proposed methods by defining importance in terms of pairwise independent variable comparisons across R^2 values associated with multiple sub-models. A linear regression model with p independent variables produces a total of 2^p sub-models corresponding to all possible combinations of independent variables included or excluded from estimation. DA achieves order independence in the importance determiniations it makes by comparing the R^2 values associated with the two focal, comparison independent variables' across all sub-models that include combinations of non-focal, other independent variables. A non-focal independent variable sub-model is one that contains a distinct subset of the $p - 2$ independent variables that are not the focus of the comparison. The $\emptyset$ or a sub-model that includes no non-focal independent variables, is also a possible sub-model used in these comparisons. In this way, the dominance comparisons do not depend on the order of independent variable inclusion and require that one of the focal, comparison independent variables obtain a higher R^2 value than the other focal independent variable across all sub-models—that is, irrespective of the order in which the independent variable is included in the model.

As an example of how the dominance comparison is implemented, consider a model with 4 independent variables: V, W, X and Z predicting Y. If I am comparing X and Z, there would be a total of four possible R^2 comparisons, each of which is reported below in Table 1. In each comparison, the subscripted model indicates the linear regression model prediction equation in symbolic form. The terms in braces include the other, non-focal independent variable subset across which both of the focal, comparison independent variables are being evaluated. X dominates Z only when all the R^2 values for sub-models that include X are greater than the R^2 values for sub-models that include Z. The dominance comparisons described in this section are known as complete dominance and are recognized as the most stringent, or hardest to achieve, dominance designation that an independent variable can have over another independent variable (Azen & Budescu, 2003).

Complete Dominance

Complete dominance is the most stringent of the dominance designations as it is a difficult designation for an independent variable to achieve over another. Complete dominance is difficult to achieve as it involves direct comparisons between independent variable pairs across multiple sub-model R^2 values and is non-compensatory; all sub-model R^2 value comparisons must show that one independent variable has a larger value than the other independent variable or the designation will fail to be achieved.

The process for determining complete dominance between a pair of independent variables, for example X and Z, with an arbitrary number of non-focal independent variables in the model $(p-2)$ proceeds as:

$$X\,\mathbf{D_f}\,Z \quad if \quad 2^{p-2} = \sum_{j=1}^{2^{p-2}} \begin{cases} if\ R^2_{Y\sim X+\{u_j\}} > R^2_{Y\sim Z+\{u_j\}}\ then\ 1 \\ else\ 0 \end{cases} \tag{1}$$

Where u_j is a distinct subset of the other $p-2$ independent variables. As in Table 1, the braces surrounding u_j indicate that it is a subset of non-focal independent variables. The $\mathbf{D_f}$ in this case is a designation indicating complete (or full/$_\mathbf{f}$) dominance of X over Z.

Ultimately, if X completely dominates Z as is outlined in Equation 1, X is clearly and unconditionally better than Z in terms of explaining variance in Y given the linear regression model from which all sub-models were derived.

Because complete dominance is a difficult criterion to achieve in comparing independent variable pairs, alternative and more compensatory, dominance designations have been proposed to provide more ways to compare the predictive usefulness of independent variables against one another. As I will discuss, the alternative dominance designations involve averaging the R^2 increment values associated with each independent variable and determining importance by comparing those average values.

Conditional Dominance

A less stringent dominance designation between independent variable pairs than complete dominance is called conditional dominance. Conditional dominance relaxes the stringency of the comparisons across pairs of independent variables by evaluating how each independent variable contributes to the R^2, on average, when they are included in every relative position in the model.

By comparing averages of R^2 increments by relative position, conditional dominance allows some sub-models with higher R^2 increment values at a specific inclusion order for X compared to Z to compensate for sub-models with lower R^2 increment values for X compared to Z at that same order. This compensatory property of the averages by inclusion order makes conditional dominance a less stringent criterion and makes it more likely to obtain a dominance designation.

The average increments to the R^2 used to determine conditional dominance are known as conditional dominance statistics. Because each independent variable can be included at any relative position in the model, in a model with p independent variables, each independent variable will have p conditional dominance statistics to compare to another independent variable. Extending on Table 1, determining conditional dominance between X and Z would involve all four different conditional dominance statistics, the computation of which are outlined below in Table 2.

The Δ used in Table 2 indicates that the R^2 value is an increment made by the focal independent variable beyond the subset of non-focal independent variables in braces. Note that the conditional dominance statistics use R^2 increments from subsets that include all other non-focal independent variables. Hence, conditional dominance comparisons between two independent variables will include increments from the independent variable against which they are being compared in their average value. This is a noteworthy difference from complete dominance designations that do not use sub-model R^2 values that include the independent variable against which the focal independent variable is being compared.

The process of computing conditional dominance statistics for X with i independent variables in the sub-model is defined as in Equation 2 below.

$$C_X^i = \frac{\sum_{g=1}^{k_i} \Delta R^2_{X+\{o_g\}}}{k_g} \tag{2}$$

Where k_i is the number of combinations of size i given p independent variables and o_g is a distinct subset of the $p-1$ independent variables of size $i-1$ that are included in the sub-model. Determining conditional dominance between X and Z proceeds as in Equation 3.

$$X \; \mathbf{D_c} \; Z \quad if \quad p = \sum_{i=1}^{p} \begin{cases} if \; C_X^i > C_Z^i \; then \; 1 \\ else \; 0 \end{cases} \tag{3}$$

Where $\mathbf{D_c}$ is a designation indicating conditional dominance of X over Z. When X does not completely but does conditionally dominate Z, X is generally better than Z for explaining variance in Y given the underlying linear regression model irrespective of inclusion order in the model. Conditional dominance thus suggests that X's explanatory value is generally higher than Z when considering their values at the same inclusion orders.

Conditional dominance between an independent variable pair is less stringent than complete dominance but can still be a difficult designation to meet in models with a great deal of between-independent variable overlap. As a result, a third dominance designation was developed that involves yet another averaging step as is described in the next section.

General Dominance

The least stringent dominance designation between independent variable pairs is called general dominance. General dominance further relaxes the stringency of the comparisons between independent variable pairs by changing the focus from comparing average increments grouped by the number of independent variables in a sub-model to the arithmetic average of these averages. General dominance is then the average of the conditional dominance statistics for each independent variable. By averaging over conditional dominance statistics, general dominance allows higher contributions at specific numbers of independent variables in the sub-model to compensate for lower contributions at other numbers of independent variables in the sub-model. The values generated by general dominance will, in almost all cases, produce a dominance designation between the pair of independent variables.

The averaged conditional dominance statistics computed for determining general dominance are known as general dominance statistics. Table 3 shows the general dominance statistic computation for X and Z. This computation incorporates all the values in Table 2 but summed into a single statistic.

As is implied by the computations in Table 3, general dominance statistics are a weighted average of the individual increments to the R^2s and is defined for X in Equation 4.

$$C_X = \frac{\sum_{i=1}^{p} C_X^i}{p} \tag{4}$$

Using the general dominance statistics computed in Equation 4, determining whether X generally dominates Z is proceeds as in Equation 5.

$$X \ \mathbf{D_g} \ Z \quad if \quad C_X > C_Z \tag{5}$$

Where $\mathbf{D_g}$ is a designation indicating general dominance of X over Z. When X does not completely or conditionally but does generally dominate Z, X is generally better than Z for explaining variance in Y given the underlying linear regression model but is sensitive to inclusion order in the model. General dominance thus suggests that X's predictive usefulness is, on average, better than Z and that it is more important when not directly considering the effects of independent variable inclusion order.

Note that the general dominance statistic values, when summed across the p independent variables, equals the sub-model R^2 when all p independent variables are included. This is a useful feature of the general dominance statistics that, as has been discussed in other reviews (Grömping, 2007; Johnson & LeBreton, 2004), ties this method to earlier work on independent variable importance which focused on the decomposition of the R^2 statistic. It is worth noting that, in this previous work, the closely related partial η^2 metric was used to develop a metric to decompose the R^2 that developed into the general dominance statistic (see Budescu, 1993; Lindeman et al., 1980). The partial η^2, also known as the squared semi-partial correlation, is an alternative formulation of the ΔR^2 from the linear regression model. By itself, the partial η^2 for an independent variable is tantamount to the conditional dominance statistic for when the subset size is p. When combined with all combinations of covariates for a specific independent variable, the partial η^2 values can be used in the same way as ΔR^2 to generate all dominance statistics and comparisons.

In the sections above, I have provided a brief discussion of the conceptual development of the DA method, reviewed how DA statistics are computed, and reviewed how importance for independent variables is determined. In the section below, I transition from a broad outline of DA to a more targeted discussion on application of DA to count regression models. The focus of the next section is on a discussion of count regression model-based DA with attention to considering which fit statistic should be used when applying DA to count regression models.

Applying Dominance Analysis to Count Regression Models

A complication of applying DA to count regression models is that the literature on DA has generally focused on how to apply the method to the linear regression model with the variance explained R^2 as a fit statistic. Given the semi-continuous nature of count dependent variables and that the predicted values from count regression models are typically in the form of a non-negative rational number, it is possible to use the explained variance R^2 when determining importance with a count regression model.

Although the variance explained R^2 could be applied to count regression models, there are good conceptual reasons to choose another fit statistic. In the section below, I provide a rationale for the choice of a different metric, the deviance R^2 or R^2_{DEV}, that better reflects the criteria underlying how count regression models fit to data.

Count Regression Fit Statistic: Deviance R^2

Statistical models are fit using information about the data as applied to a probability distribution to find their most likely parameter values. A fit statistic applied to evaluating the fit of a statistical model to data is also most useful when the computation of the statistic is conceptually aligned with the model that it is being used to evaluate. Thus, choosing a fit statistic that matches the underlying fitting criterion of statistical model's probability distribution will best reflect how the model fits to data.

Consider that the explained variance R^2 is computed as $\frac{\sum(\bar{Y}-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{SS_{model}}{SS_{total}}$ or the ratio of the variance of the predicted values (i.e., the model sums of squares; SS_{model}) over the variance of the predicted values (i.e., the total sums of squares; SS_{total}). The linear regression model is based on a Normal probability distribution and uses least-squares as its fitting criterion. Least squares seeks to minimize $SS_{residual} = \sum(Y - \hat{Y})^2$ or the residual sums of squares between the predicted values from the linear regression model and the observed dependent variable. The $SS_{residual} = SS_{total} - SS_{model}$ and thus the explained variance R^2 can also be computed as $1 - \frac{SS_{residual}}{SS_{total}}$. The explained variance R^2 is then closely tied to the linear regression model and its fitting criterion. It is important to note that the computation used to obtain the $SS_{residual}$ is also known as the deviance (DEV) for the Normal distribution as it describes how the model's predictions deviate from observed values (McCullagh & Nelder, 2019). The $SS_{residual}$ is thus a DEV_{model} or a model deviance for the linear regression model.

The Poisson regression follows the Poisson distribution and the negative Binomial regression follows the negative Binomial distribution–both of which are probability distributions meant for discrete data like non-negative counts. For example, the deviance for Poisson regression and a special case of the negative Binomial regression[1] is $\sum Y \ln \frac{Y}{\hat{Y}} - (Y - \hat{Y})$. Note that this Poisson regression-focused deviance value differs from the Normal distribution deviance in that it tends to penalize underprediction more than overprediction. The extra penalties assigned to underprediction are consistent with the truncated, semi-continuous nature of count dependent variables in that they cannot go below 0 and, thus, tend to be penalized more heavily toward the conceptual lower bound of the distribution. By contrast, Normal distribution-based deviance has no such constraint and penalizes discrepancies from observed values equally in either direction.

In considering a reasonable fit statistic to apply to count regression models given the differences between the underlying fitting criteria for the linear regression model and count regression models, Cameron and Windmeijer (1996) devised the R^2_{DEV} or deviance R^2 outlined in Equation 6 below.

$$R^2_{DEV} = 1 - \frac{DEV_{model}}{DEV_{null}} \tag{6}$$

Where DEV_{null} is the model deviance for an intercept- or mean-only model and is equivalent to SS_{total} for an linear regression model. The R^2_{DEV} is then a direct extension of the explained variance R^2 but is more flexible in that it can be applied to count regression models when using their deviance computations. I recommend the use of R^2_{DEV} with DEV_{model} and DEV_{null} values that are based on each count regression model's fitting criteria as a fit statistic for DA.

The R^2_{DEV} is a conceptually more reasonable choice than the explained variance R^2 for count regression model-based DA as, if a researcher applies the explained variance R^2

[1] This special case is the negative Binomial regression estimated using a quasi-likelihood method. Maximum likelihood methods require a more complex form given the estimation of the α/δ parameter.

to a count regression model, they are applying a fit statistic that uses a deviance computation intended for the linear regression model to count regression models with very different fitting criteria. Having outlined a rationale for the choice of a specific fit statistic to use when applying DA to count regression models, in the sections to follow I transition to an example application of the proposed methodology. The sections below briefly outline the data generated for this illustration, describe the models estimated from the data, and also describe how DA designations were determined.

Count Regression Model-based Dominance Analysis: An Analytic Example

The goal of this section was to provide an example analysis using the recommended R^2_{DEV} count regression model-based DA methodology. This example was intended to be useful as a guide for researchers and analysts interested in applying this method to count regression models. In the section below, I began by describing how I generated data for the example based on a Poisson regression model[2]. The methods used to generate these data were not directly relevant to the goals of this manuscript and, as such, have been included in an online supplement. Please note that the online supplement outlines the code to generate the data and development perspective behind the generation of these data in great detail. The section below was focused primarily on describing the conceptual nature of the data so that the reader can follow along prior to transitioning to reporting on the primary analysis.

Data Generation

The fabricated data simulated and discussed in this work is related to a fabricated study on the number of solutions a student provides to personal relationship problem given 20 minutes to respond that met minimum criteria for quality of reasoning. The data generated for this study were collected from 6,780 simulated students. The study researchers expected that the number of personal relationship problem solutions reported

[2] An example using a negative Binomial regression is available in the online supplement.

by each simulated student would be Poisson distributed[3].

The four independent variables used in the fabricated data were four survey scales that were normed such that the mean of each was roughly 0 and standard deviation was roughly 1. The first scale used to predict the number of solutions reported was a cognitive ability, or intelligence, scale reflecting the respondent's cognitive capacity and skills. This scale is denoted *ability* in the results. The second survey scale was solution motivation or the respondent's motivation to try to resolve the personal relationship problem. This scale is denoted *motivation* in the results. The third survey scale was tactfulness or the extent to which the respondent is sensitive to social considerations. This scale is denoted *tact* in the results. The fourth survey scale was rhetorical skill or the skill with which the respondent can construct a convincing argument This scale is denoted *skill* in the results.

The means, standard deviations, and correlations between all four independent variables and two dependent variables are reported below in Table 4.

Note that, on average, each of the simulated respondents produced a solution in the 20-minute period under study. Table 4 also shows that the variance of the number of solutions produced was consistent with its expected underlying Poisson distribution. Specifically, solution production had a variance of 1 which closely matches the mean as is assumed of the Poisson distribution.

Regression Results

The solution production dependent variable was Poisson distributed by design and could be modeled using a Poisson regression. The Poisson regression results using the four survey scale independent variables to predict solution counts were reported in Table 5. Table 5 included coefficients (β), standard errors (σ_β), 95% confidence intervals, and exponentiated coefficients or IRRs (e^β).

[3] The data used in this section were simulated from a single set random number draws with a single random seed for reproducibility. The methodology underlying the simulation is described further in the online supplement.

Table 5 showed that each of the independent variables had a positive effect on the number of solutions produced and each appeared to be statistically significant (at the $p < .05$ level) as was implied by the confidence intervals not including the value of 0.

In terms of coefficient magnitude, *ability* had the largest effect on the rate of solution production. Each standard deviation increase in the *ability* scale led to a 27.5% increase in the solutions produced given the IRR value. *skill* obtained the second-largest effect on solution production rates. Each standard deviation increase on the *skill* scale led to a 12.3% increase in the number of solutions produced given its IRR. *motivation* and *tact* obtained the smallest increases in solution production. *motivation* resulted in a 11.4% increase and *tact* a 10.5% increase in the number of solutions produced given their IRRs.

The IRR value for the model intercept represented the expected number of solutions produced when all independent variables were at their means of 0. Because each independent variable had a mean of 0, the model intercept represented the mean number of solutions produced. Note that the .8601 value obtained was similar to the overall mean of solutions in Table 4. This .8601 value could also used as the baseline rate of solutions produced. For instance, the expected number of solutions produced for respondents who had an *ability* score that was one standard deviation above the mean, but had mean value on all other independent variables, was $.8601 * 1.2753 = 1.0969$ or just above 1 solution produced. As opposed to reporting on single values as I have above, plotting of marginal means across levels of different independent variables adds value to the model interpretation (see Rönkkö et al., 2022, for a similar perspective) and more clearly depicts the multiplicative effects produced by count regression models. Although I did not include marginal means plots in this manuscript for brevity, such plots were reported in the online supplement for interested readers.

The Poisson regression modeling result reported in Table 5 showed that all four independent variables had non-trivial predictive effects on solution production and, in addition, obtained different coefficient magnitudes. These different coefficient magnitudes, in combination with the different variances observed in Table 4, indicated likely differences in the importance of each independent variable for explaining solution production.

In the section below, I transition to the focal analysis of this work where I determined the importance of each independent variable for predicting solution production using DA. In this way, the section below provides an empirical example that applied the designation and computational formulas in Equations 1, 2, 3, 4, and 5 to the Poisson regression model in Table 5.

Dominance Analysis Results

The DA designations and statistics used in this manuscript were computed using the collection of model fit statistics representing all possible combinations of independent variables included and excluded as sub-models. The four survey scale independent variables used in this manuscript resulted in a total of $2^4 = 16$ sub-models estimated from the data. The results from all sub-models were reported below in Table 6 omitting the sub-model with no predictors as it was not informative (i.e., produced a value of 0).

Table 6 showed that *ability* tended to obtain larger R^2_{DEV} values. Thus, *ability* was likely to be an important independent variable consistent with its large coefficient size. The R^2_{DEV} values associated with the three other independent variables showed no easy to discern pattern and, moreover, did not as clearly follow the coefficient sizes reported in Table 5. As such, the information that could be obtained from the dominance analysis designations was likely would be useful, not only for confirming *ability*'s importance, but also for clarifying the more complex patterns of interrelations between the other three independent variables.

The first dominance designation evaluated using the results in Table 6 was complete dominance. *ability* generally had larger R^2_{DEV} values than other variables and, when using Equation 1 to determine complete dominance over to the other three independent variables, was shown to completely dominate all three. Thus, *ability* was the undisputed top independent variable predicting solution production irrespective of independent variable inclusion ordering in the model.

The complete dominance relationships between the remaining variables was far less clearly defined. For instance, *skill* completely dominated *tact* but did not completely dominate *motivation*. *motivation* and *tact* also had no complete dominance designation. The failure of these independent variable comparisons to have had complete dominance designations prevented developing a clear hierarchy between the last three variables as *skill* was a better predictor than *tact* irrespective of inclusion order, but no such determination could be made of the rest of the comparisons. Thus, *skill* was likely the second-best predictor but it required use of one of the expanded dominance designations to confirm that was the case.

In order to rank the remaining independent variables, the conditional dominance statistics for all the survey scale independent variables were computed using Equation 2. The results of the conditional dominance statistic computations were depicted graphically in Figure 1 below. For this Figure, the y-axis was depicted on a logarithmic scale to improve legibility nearer values where the independent variable was included last (i.e., subset sizes of 4).

I used a graphical format to depict the conditional dominance statistics' values as this format more clearly conveys the dominance designations between independent variables than does a table of values. In the graphic, the orientation of each independent variable's conditional dominance statistic trendline relative to other independent variables trendlines represented the information conveyed by Equation 3. Specifically, when an independent variables conditional dominance trendline was always above another independent variable's conditional dominance trendline, the independent variable that was above conditionally dominated the one below.

The trendlines in Figure 1 confirmed the complete dominance results in that *ability*'s trendline was above the trendlines for the other three independent variables indicating that it conditionally dominated each of them. Similarly, *skill*'s trendline was above *tact*'s trendline consistent with the complete dominance results. In addition, *skill*'s trendline was above *motivation*'s trendline indicating that it conditionally dominated *motivation*. Although *skill* did not explain more information than *motivation* in all comparable models (see $solutions \sim motivation + tact$ versus $solutions \sim tact + skill$'s in Table 6), when considering the average information explained given inclusion order, *skill* produced bigger increments to the R^2_{DEV} than did *motivation* levels. As such, *skill*'s dominance of *motivation* was model dependent, but not generally dependent on independent variable inclusion order.

The last two independent variables did not result in conditional dominance designation as *motivation* failed to conditionally dominate *tact* given *tact*'s conditional dominance statistic when included last in the model (i.e., at subset size 4) was larger than *motivation*'s statistic when included last in the model. These conditional dominance results further reinforce the idea that attempting to rank these independent variables is not straightforward and their contributions to prediction depend on the order of their inclusion in the model.

Given that no complete or conditional dominance designations were possible for comparing *motivation* with *tact*, I proceeded to evaluate the general dominance designations between these variables. The general dominance statistics for each survey scale independent variable was computed using Equation 4 as reported in Table 7.

Evaluating the general dominance designations determined by the general dominance statistics in Table 7 using Equation 5 again showed that *ability* dominated each other variable and that *skill* dominated *tact* and *motivation*. A useful addition that general dominance designations added was in determining general dominance between *motivation* and *tact*. The general dominance designations then added to the previous dominance results in that they were the final component needed to construct an importance hierarchy among the independent variables. In combination, the results across all three dominance designations resulted in a clear rank ordering of the survey scale independent variables in predicting solutions produced: *ability* was most important followed by *skill*, then *motivation*, and *tact* was least important.

In conclusion, the DA results have built on and extended the coefficient reporting in Table 5 by adding additional information about each of the survey scale independent variables' predictions. In particular, the dominance results offered conclusive evidence for differences between independent variables in their ability to explain variation or information in the solution production variable. The DA results supported the inference about the predictive usefulness of *ability* given its large IRR value. The DA results showed that, regardless of the order in which *ability* might be included in the model, it always produced the biggest increment to the R^2_{DEV}. In addition, *ability* was associated with nearly one-third (i.e., $\frac{.0965}{.2624} \approx \frac{1}{3}$) of the explained information in solutions produced. The DA results also provided useful contextualization of *skill*, *motivation*, and *tact*. These three independent variables obtained similar IRR values, each within .01 of one another, which made their dominance hierarchy less easy to guess at the outset. The dominance designations obtained showed that the hierarchy between these variables was indeed more nuanced as *skill* completely dominated only *tact* and conditionally dominated *motivation* levels. Moreover, *motivation* only generally dominated *tact*. The DA results then provided useful additional information to a researcher about the differences in strength of prediction between these independent variables that would not be possible to have obtained only with their IRR values.

Discussion

In this manuscript I have recommended a methodology for determining the relative importance of independent variables in count regression models. I recommend the DA methodology as an approach that is comprehensive in the information it provides about independent variable's prediction and, when using an appropriate fit statistic such as the R^2_{DEV}, can provide information analogous to the explained variance R^2 using the linear regression model.

I have also walked the reader through an example data analysis applying Poisson regression to simulated data. In walking the reader through this example, I use the recommended DA with R^2_{DEV} fit statistic approach to evaluate the relative importance of four independent variables. This walk through of the DA has focused on the utilization of different levels of DA designation stringency and how these different levels offer different weights of evidence for the importance of the independent variables over one another in predicting the count dependent variable.

By combining these two topics, this manuscript picks up where Blevins et al. (2015) had left off by recommending the use of DA as a postestimation method to better understand count regression model predictions. Specifically, I recommend consulting Blevins et al's work when choosing to implement a count regression model as their decision flowchart can help you to choose the most appropriate count regression model given the nature of your data and follow the estimation of the count regression model with a DA to better contextualize the predictions made by each independent variable.

In this manuscript I have discussed many key considerations for researchers contemplating implementing DA in using count regression models, but I acknowledge that several relevant topics have not been included. Before closing, I discuss some noteworthy limitations and additional extensions of this work.

Limitations and Future Directions

In this work, I use only simulated data in the empirical examples. The use of simulated data was an intentional choice to avoid the need to work through many of the decision points outlined by Blevins et al. (2015) related to the selection of an appropriate count regression model as I know the dependent variables are distributed as Poisson or negative Binomial. I acknowledge that the use of the simulated data adds additional complexity to following along with the methodology in the online supplement. That said, all the procedures used to simulate the various count dependent variables are fully replicable, well-documented, and available as a markdown that can be used by interested readers with a working knowledge of R (R Core Team, 2022). In addition, I acknowledge that only reporting the results from the Poisson regression models is a limitation. The Poisson regression and negative Binomial regression results were very similar and, in my view, including the negative Binomial regression results would not meaningfully add to the manucscript's narrative. Although the negative Binomial regression model's results are not in the manuscript, I have included the negative Binomial regression results in the supplement for interested readers.

Prior work on DA has recommended that researchers use bootstrapping to estimate the reproduce-ability of dominance designations (Azen & Budescu, 2003). Bootstrapping the count regression model-based dominance statistics and designations is also possible but was not examined in the present work. Although I have not provided an example of bootstrap reproduce-ability in this work, evaluating the bootstrap reproduce-ability of dominance designations is a useful and important practice. Evaluating reproduce-ability allows researchers assess a level of confidence that specific designations between independent variables will hold under resampling. Thus, like standard hypothesis testing, evaluating bootstrap reproduce-ability can allow a researcher to better determine whether a set of dominance designations between independent variables in a count regression model are likely to generalize beyond the sample at hand.

In addition, zero inflation is commonly observed of count dependent variables. Zero inflation is a condition where the distribution of the count dependent variable has more 0s than would be expected given a standard Poisson or negative Binomial distribution and requires the use of specialized models (**bhaskar2023regression**; e.g., Blevins et al., 2015). Cameron and Windmeijer (1996) discuss the application of the R^2_{DEV} to zero-inflated count regression models and, thus, a DA methodology based on the same general approach as discussed above could be applied to zero-inflated count regression models. One additional complication that arises with considering how to determine importance with zero-inflated count regression models is that these models encompass two predictive processes. The first process is the standard count generating process whereby independent variables increase or decrease dependent variable counts. The second is an "opt out" process whereby independent variables increase or decrease the likelihood of the count being 0. These two processes add complexity in that they can be modeled differently. Any one independent variable can predict the count generating process, the opting out process, or both. When using a DA with zero-inflated count regression models, I recommend the researcher consider whether they are truly interested in determining the importance of independent variables or are actually interested in determining the importance of parameter estimates (Luchman et al., 2020). The key difference between the two perspectives is that, if one independent variable is included in both the count and opt out process, the independent variable approach would ascribe the independent variable a single set of dominance statistic designations whereas the parameter estimate perspective would break the designations into one focused on the independent variable's effect in the count process and, separately, the independent variable's effect in the opt-out process.

Conclusion

DA is a useful post-estimation methodology for determining the importance of independent variables in statistical models such as count regression models. This manuscript has provided a recommended methodology for extending DA to count regression models and offered an extensive data analytic example focusing on the interpretation of DA statistics and designations with simulated data. In combination, the conceptual discussion of DA and count regression models when paired with the empirical example in this paper, will provide scientists with useful tools they can use to better understand the results of count regression models they estimate in support of research questions with count data.

References

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*(2), 129–148.

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.

Bettinazzi, E. L., & Feldman, E. R. (2021). Stakeholder orientation and divestiture activity. *Academy of Management Journal*, *64*(4), 1078–1096.

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.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*(3), 542–551.

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.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, *61*(2), 139–147.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, *7*(3), 238–257.

Lindeman, R. H., Merenda, P. F., Gold, R. Z., et al. (1980). *Introduction to bivariate and multivariate analysis* (Vol. 4). Scott, Foresman Glenview, IL.

Luchman, J. N. (2014). Relative importance analysis with multicategory dependent variables: An extension and review of best practices. *Organizational Research Methods*, *17*(4), 452–471.

Luchman, J. N., Lei, X., & Kaplan, S. A. (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.

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Motowidlo, S. J. (2003). Job performance. *Handbook of Psychology: Industrial and Organizational Psychology*, *12*(4), 39–53.

Naumovska, I., Zajac, E. J., & Lee, P. M. (2021). Strength and weakness in numbers? unpacking the role of prevalence in the diffusion of reverse mergers. *Academy of Management Journal*, *64*(2), 409–434.

Naveh, E., Katz-Navon, T., & Stern, Z. (2015). Active learning climate and employee errors: The moderating effects of personality traits. *Journal of Organizational Behavior*, *36*(3), 441–459.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rönkkö, M., Aalto, E., Tenhunen, H., & Aguirre-Urreta, M. I. (2022). Eight simple guidelines for improved understanding of transformations and nonlinear effects. *Organizational Research Methods*, *25*(1), 48–87.

Rotolo, D., & Messeni Petruzzelli, A. (2013). When does centrality matter? scientific productivity and the moderating role of research specialization and cross-community ties. *Journal of Organizational Behavior*, *34*(5), 648–670.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*, *26*(1), 1–9.

Table 1

Example Dominance Comparisons

	Sub-model with X	Sub-model with Z
Across null set/no other independent variables	$R^2_{Y\sim X+\{\emptyset\}}$	$R^2_{Y\sim Z+\{\emptyset\}}$
Comparing across W	$R^2_{Y\sim X+\{W\}}$	$R^2_{Y\sim Z+\{W\}}$
Comparing across V	$R^2_{Y\sim X+\{V\}}$	$R^2_{Y\sim Z+\{W\}}$
Comparing across both W and V	$R^2_{Y\sim X+\{W+V\}}$	$R^2_{Y\sim Z+\{W+V\}}$

Table 2

Example Conditional Dominance

Comparing at	Average with X	Average with Z
One independent variable	$\Delta R^2_{Y\sim X+\{\emptyset\}}$	$\Delta R^2_{Y\sim Z+\{\emptyset\}}$
Two independent variables	$\frac{1}{3}(\Delta R^2_{Y\sim X+\{W\}}+$ $\Delta R^2_{Y\sim X+\{V\}}+$ $\Delta R^2_{Y\sim X+\{Z\}})$	$\frac{1}{3}(\Delta R^2_{Y\sim Z+\{W\}}+$ $\Delta R^2_{Y\sim Z+\{V\}}+$ $\Delta R^2_{Y\sim Z+\{X\}})$
Three independent variables	$\frac{1}{3}(\Delta R^2_{Y\sim X+\{W+V\}}+$ $\Delta R^2_{Y\sim X+\{W+Z\}}+$ $\Delta R^2_{Y\sim X-\{V+Z\}})$	$\frac{1}{3}(\Delta R^2_{Y\sim Z+\{W+V\}}+$ $\Delta R^2_{Y\sim Z+\{W+X\}}+$ $\Delta R^2_{Y\sim Z+\{V+X\}})$
Four independent variables	$\Delta R^2_{Y\sim X+\{W+V+Z\}}$	$\Delta R^2_{Y\sim Z+\{W+V+X\}}$

Table 3

Example General Dominance

Average with X	Average with Z
$\frac{1}{4}\Delta R^2_{Y\sim X+\{\emptyset\}}+$	$\frac{1}{4}\Delta R^2_{Y\sim Z+\{\emptyset\}}+$
$\frac{1}{12}(\Delta R^2_{Y\sim X+\{W\}}+$	$\frac{1}{12}(\Delta R^2_{Y\sim Z+\{W\}}+$
$\Delta R^2_{Y\sim X+\{V\}}+$	$\Delta R^2_{Y\sim Z+\{V\}}+$
$\Delta R^2_{Y\sim X+\{Z\}})+$	$\Delta R^2_{Y\sim Z+\{X\}})+$
$\frac{1}{12}(\Delta R^2_{Y\sim X+\{W+V\}}+$	$\frac{1}{12}(\Delta R^2_{Y\sim Z+\{W+V\}}+$
$\Delta R^2_{Y\sim X+\{W+Z\}}+$	$\Delta R^2_{Y\sim Z+\{W+X\}}+$
$\Delta R^2_{Y\sim X-\{V+Z\}})+$	$\Delta R^2_{Y\sim Z+\{V+X\}})+$
$\frac{1}{4}\Delta R^2_{Y\sim X+\{W+V+Z\}}$	$\frac{1}{4}\Delta R^2_{Y\sim Z+\{W+V+X\}}$

Table 4

Descriptive Statistics

		Standard	Correlations				
Variable	Mean	Deviation	*ability*	*motivation*	*tact*	*skill*	*solutions*
ability	−0.0355	1.1853	1.0000	0.4168	0.1446	0.2180	0.4175
motivation	−0.0261	1.4494	0.4168	1.0000	0.1958	0.2850	0.3592
tact	−0.0056	1.5480	0.1446	0.1958	1.0000	0.3295	0.2913
skill	0.0144	1.9211	0.2180	0.2850	0.3295	1.0000	0.3726
solutions	0.9940	1.0082	0.4175	0.3592	0.2913	0.3726	1.0000

Table 5

Poisson Regression Predicting Relationship Problem Solutions Produced

			95% Confidence Interval		
	β	σ_β	Low	High	e^β
ability	0.2431	0.0114	0.2208	0.2655	1.2753
motivation	0.1081	0.0095	0.0896	0.1267	1.1142
tact	0.0999	0.0083	0.0835	0.1162	1.1050
skill	0.1158	0.0069	0.1024	0.1293	1.1228
Intercept	−0.1507	0.0138	−0.1780	−0.1237	0.8601

Table 6

R^2_{DEV} by Sub-model

$solutions \sim ability$	0.1524
$solutions \sim motivation$	0.1131
$solutions \sim tact$	0.0744
$solutions \sim skill$	0.1219
$solutions \sim ability + motivation$	0.1887
$solutions \sim ability + tact$	0.2006
$solutions \sim ability + skill$	0.2248
$solutions \sim motivation + tact$	0.1584
$solutions \sim motivation + skill$	0.1844
$solutions \sim tact + skill$	0.1502
$solutions \sim ability + motivation + tact$	0.2266
$solutions \sim ability + motivation + skill$	0.2442
$solutions \sim ability + tact + skill$	0.2459
$solutions \sim motivation + tact + skill$	0.2050
$solutions \sim ability + motivation + tact + skill$	0.2624

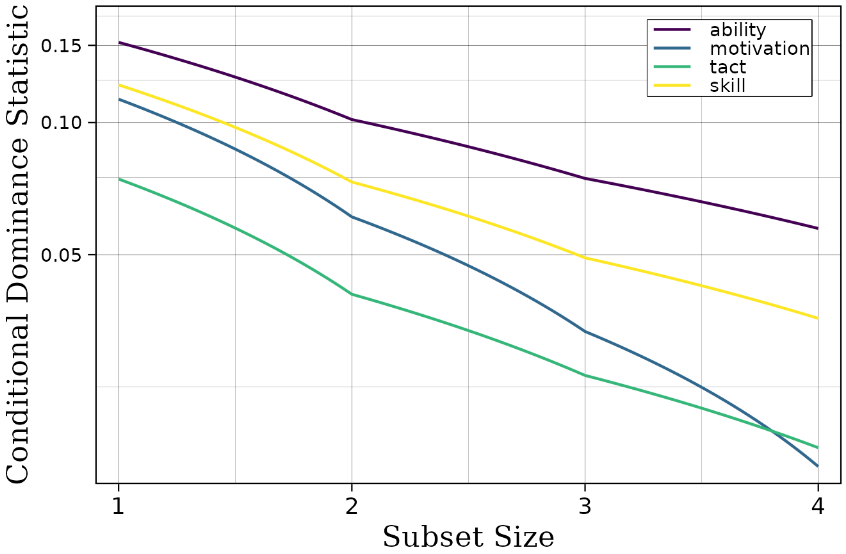
Table 7

General Dominance Statistics

ability	0.0965
motivation	0.0560
tact	0.0399
skill	0.0700

Figure 1

Conditional Dominance Statistics


ability
motivation
tact
skill
Conditional Dominance Statistic
Subset Size