

---

# Relative importance analysis for count regression models

Journal Title  
XX(X):1–28  
©The Author(s) 0000  
Reprints and permission:  
sagepub.co.uk/journalsPermissions.nav  
DOI: 10.1177/ToBeAssigned  
www.sagepub.com/

SAGE

Joseph N. Luchman<sup>1</sup>

## 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 science often uses how many times a behavior is observed as an outcome to answer research questions. Such behavior counts arise from many different sources at the organization- and individual-level of aggregation. Organization-level behavior counts used in the literature include the number of organizations adopting a

---

<sup>1</sup>Fors Marsh

specific practice in a week (Naumovska, Zajac & Lee, 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 by an academic (Rotolo & Messeni Petruzzelli, 2013) or number of errors that resulted in an accident in the last three months among medical doctors (Naveh, Katz-Navon & Stern, 2015). Behavior counts such as the above examples are valuable outcomes given that organizational science concepts are often defined 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).

The conceptual value of behavior counts as outcomes aside, data analysis with behavior counts as a dependent variable/DV 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, Tsang & Spain, 2015). These count regression models/CRMs include the Poisson regression/PR and negative Binomial regression/NBR models. Both PR and NBR are commonly used in the organizational science literature and implemented in many data analytic software environments.

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

The log-linear predictive equations and coefficients CRMs estimate are challenging to interpret directly. Log-linear CRM coefficients are challenging to interpret because they describe how the natural logarithm of the DV changes given a 1 unit change to an independent variable/IV. When back-translated through an exponential function, CRM coefficients are known as incidence rate ratios and describe the percentage change in the DV per unit change to the IV. Importantly, the percentage change described by CRM coefficients makes the count values produced by those changes relative to where they began. For example, a CRM 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 as each describe a 100% increase in the count DV value.

Model post-estimation methods such as graphing estimated marginal means are useful interpretive tools for log-linear CRMs to better contextualize the logarithmic predicted values they produce. Another increasingly common model post-estimation tool used to contextualize model predictions are relative importance analysis methods (Tonidandel & LeBreton, 2011). Relative importance analysis methods focus on how different IVs in the model contribute to the  $R^2$  as computed by methods such as dominance analysis/DA (Azen & Budescu, 2003).

To date, published methodological work has extended DA from the linear regression model/LRM on which it was originally developed, to generalized linear models including binary (Azen & Traxel, 2009), ordered, and multinomial logit models

(Luchman, 2014). This work extends DA to CRMs and makes three contributions to the literature. This work first reviews DA as a relative importance methodology and recommends using a specific pseudo- $R^2$  statistic for CRM-based DA and implements multiple data analytic examples of DA with that statistic. Despite the applicability of DA to any model fit statistic that could be computed for a CRM, pseudo- $R^2$  values are valuable for DA due to their interpretability. The recommended pseudo- $R^2$  statistic is a direct generalization of the explained variance  $R^2$  in the LRM to generalized linear models based on model deviance. The data analytic examples then use this deviance-based pseudo- $R^2$  as the fit statistic in a series of DA computations with simulated Poisson and negative Binomial distributed data.

Second, this work discusses data generation mechanisms with count DVs and outlines how research questions with count DVs can, under certain circumstances, be affected by a concept known as differential exposure. Differential exposure is a concept whereby two observations have different numbers of opportunities to produce behavior counts. Oftentimes, such differential exposure is a nuisance factor related to data collection context and not of substantive interest and is correctable in the model using an offset term with a variable measuring differential exposure. As I will show in this work, differential exposure is of particular import for DA statistic computation as it can affect estimates of DA statistics even when model coefficients are relatively unaffected.

Finally, this work extends on that of Blevins, Tsang, and Spain (2015)'s review and recommendations for the use of CRMs in organizational science. Blevins et al. discuss the extent of CRM use in the organizational science literature, describe model and analytic details about the PR and NBR models, and provide a flowchart that researchers can use to identify which CRM might be best to choose for their data analysis. In this work, I extend on their review to add an in-depth discussion of the DA methodology 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, produce improvements to predicting the count DV. In addition, the discussion of differential exposure extends on Blevins et al.'s conceptual review to cover more of the considerations important for CRM-based data analysis.

In this work, I begin with a detailed discussion of the conceptual background of DA—the relative importance analysis method most applicable to CRMs. Here I discuss the different levels of dominance between two IVs as defined by DA, how these levels are determined in the data, and what this level of dominance means in terms of the importance of any IV in the model. The conceptual discussion of DA is followed by a more focused discussion where I apply DA to CRMs. The goal of the focused CRM-based DA discussion is to recommend a model fit statistic for use in dominance designations. In the third section, I apply the topics discussed in the previous two sections in a detailed PR- and NBR-based DA using simulated data designed be useful in illustrating key CRM-based DA concepts. In a fourth and final section, I turn to discussing the concept of differential exposure in count DVs, show how it can affect inferences from CRM analysis, and provide the method used to correct for it in analysis. My goal in discussing differential exposure is to outline the effect of this data generation mechanism on DA statistics and designations. I compute and compare

CRM-based DA statistics with and without corrections based, again, on simulated data in this section to illustrate the impact of ignoring this issue during analysis.

## Dominance Analysis: A Unique Relative Importance Analysis Method

Research methodologists and statisticians have used many methods over the years to determine how important an IV is in a LRM (see reviews in Grömping, 2007; Johnson & LeBreton, 2004). Importance methods have included the correlation between the IV and DV, an IV's standardized regression coefficient, as well as the increment to the  $R^2$  makes, an IV makes when predicting the DV. These methods, however informative in specific circumstances, are not considered to be useful as importance metrics as they fail to simultaneously account for the IV's bivariate and multivariate (i.e., controlling for other IV) relationships with the DV (Johnson & LeBreton, 2004). Accommodating both bi- and multivariate relationships is important in determining importance as, for most statistical models, the order in which IVs are included is arbitrary yet the order of inclusion of IVs fundamentally affects the inferences made about IV importance. Hence, it is essential to use importance methods that are independent of inclusion order.

One importance method that is independent of ordering and has been favorably received in the methodology community, is DA. DA was originally developed as a method for determining IV importance for the LRM by Budescu (1993) that extended on past methods by defining importance in terms of pairwise IV comparisons as inferred from the direct comparison of sub-model  $R^2$  values. Sub-model  $R^2$  are those computed by re-fitting the model to the data with some subset of all the IVs. In a model with  $p$  IVs there are a total of  $2^p$  sub-models including the full model with all  $p$  IVs and the null model with none of the IVs (i.e., an intercept-only model)—which always produces a value of 0 and is often omitted. DA achieves order independence by comparing the two focal IVs'  $R^2$  values across sub-models that contain all combinations of the non-focal IVs including the null set of no other non-focal IVs. In this way, the dominance comparisons produced by DA do not depend on the order of IV inclusion and, indeed, require that an IV obtain a higher value than another IV irrespective of inclusion of other non-focal IVs.

As an example of how the dominance comparison is implemented, consider a model with 4 IVs:  $IV_x$ ,  $IV_z$ ,  $IV_w$  and  $IV_v$  predicting  $Y$ . If the dominance comparison was focusing on  $IV_x$  versus  $IV_z$ , there would be a total of 3 possible  $R^2$  comparisons each of which is outlined below in Table 1. In each comparison, the subscripted model indicates the form of the LRM equation in symbolic form (i.e., similar to an R formula interface).  $IV_x$  dominates  $IV_z$  only when all the sub-model  $R^2$  values that include  $IV_x$  are greater than the sub-model  $R^2$  values that include  $IV_z$  across all three of these sub-models. The dominance comparison described in this section has, more recently, been named complete dominance and has been recognized as the most stringent, or hardest to achieve, dominance designation in a series of three possible designations (Azen & Budescu, 2003).

**Table 1.** Example Dominance Comparisons

	Sub-model with $IV_x$	Sub-model with $IV_z$
Across null set/no other IVs	$R_{Y \sim IV_x}^2$	$R_{Y \sim IV_z}^2$
Comparing across $IV_w$	$R_{Y \sim IV_x + IV_w}^2$	$R_{Y \sim IV_z + IV_w}^2$
Comparing across $IV_v$	$R_{Y \sim IV_x + IV_v}^2$	$R_{Y \sim IV_z + IV_v}^2$
Comparing across both $IV_w$ and $IV_v$	$R_{Y \sim IV_x + IV_w + IV_v}^2$	$R_{Y \sim IV_z + IV_w + IV_v}^2$

### Complete Dominance

Complete dominance is the most stringent of the dominance designations as it is a difficult designation for an IV to achieve over another. Complete dominance is difficult to achieve as it involves direct comparisons between IVs at the level of individual sub-models'  $R^2$  values. The reason this designation is so difficult to achieve is that it is non-compensatory; all sub-model  $R^2$  value comparisons must favor one IV over another or the designation will fail to be achieved.

The process for determining complete dominance between two IVs,  $IV_x$  and  $IV_z$  with an arbitrary number of IVs in the model ( $p$ ) can also be defined as:

$$IV_x D IV_z \quad \text{if} \quad k = \sum_{j=1}^k \begin{cases} \text{if } R_{Y \sim IV_x + \{u_j\}}^2 > R_{Y \sim IV_z + \{u_j\}}^2 & \text{then } 1 \\ \text{else} & 0 \end{cases} \quad (1)$$

Where the total number of combinations of IVs not including  $IV_x$  or  $IV_z$  is  $k = 2^{p-2}$  and  $u_j$  is a distinct subset of the other  $p - 2$  IVs. The braces surrounding  $u_j$  indicate that it is a subset of IVs such as those in Table 1. The  $D$  in this case is a designation indicating complete dominance of the left hand IV over the right hand IV.

Ultimately, if  $IV_x$  completely dominates  $IV_z$  as is outlined in Equation 1,  $IV_x$  has bested  $IV_z$  in all possible directly comparable ways in the model. As such, when  $IV_x$  completely dominates  $IV_z$ , it clearly explains more of  $Y$ , and is thus more important for explaining differences observed in  $Y$ , in the model

Because complete dominance is a difficult criterion to achieve in comparing two IVs, alternative, less-stringent and more compensatory, dominance designations have been proposed to provide more ways to compare the predictive usefulness of IVs against one another. The criteria used in the sections to come involve averaging the  $R^2$  values associated with the IVs in the comparison and comparing those resultant averages.

### Conditional Dominance

A less stringent dominance designation between IV pairs than complete dominance is called conditional dominance. Conditional dominance relaxes the stringency of the comparisons across pairs of IVs by changing the focus from comparing  $R^2$  values of individual sub-models to averages of  $R^2$  increments across sub-models. By comparing averages of  $R^2$  increments, as opposed to individual sub-model  $R^2$  values, conditional dominance allows for some sub-models with higher values of  $R^2$  increments to compensate for sub-models with lower values of  $R^2$  increments in the averaging. Thus, it is more likely that an IV will be determined to be conditionally dominant over another given a set of sub-model  $R^2$  values.

The specific comparisons made to determine conditional dominance of one IV over another focus on the average increment an IV makes to the  $R^2$  with a specific number of IVs in the sub-model. The averages are aggregated by the number of IVs in the sub-model as it is necessarily the case that, as more IVs are added to a model, the incremental contribution any one IV can make to the  $R^2$  shrinks. Hence, conditional dominance designations acknowledge that increments to the  $R^2$  are best compared in the context of the number of IVs in the model.

The average increments to the  $R^2$  computed for determining conditional dominance are known as conditional dominance statistics. Each IV has a conditional dominance statistic for each number of IVs in the sub-model. Thus, with  $p$  IVs, each IV will have  $p$  conditional dominance statistics to compare to another IV. Extending on Table 1, determining conditional dominance between  $IV_x$  and  $IV_z$  would involve all four different conditional dominance statistics the computation of which are outlined below in Table 2.

**Table 2.** Example Conditional Dominance

Comparing at	Average with $IV_x$	Average with $IV_z$
One IV	$\Delta R_{Y \sim IV_x + \{\emptyset\}}^2$	$\Delta R_{Y \sim IV_z + \{\emptyset\}}^2$
Two IVs	$(\Delta R_{Y \sim IV_x + \{IV_w\}}^2 + \Delta R_{Y \sim IV_x + \{IV_v\}}^2 + \Delta R_{Y \sim IV_x + \{IV_z\}}^2)^{\frac{1}{3}}$	$(\Delta R_{Y \sim IV_z + \{IV_w\}}^2 + \Delta R_{Y \sim IV_z + \{IV_v\}}^2 + \Delta R_{Y \sim IV_z + \{IV_x\}}^2)^{\frac{1}{3}}$
Three IVs	$(\Delta R_{Y \sim IV_x + \{IV_w + IV_v\}}^2 + \Delta R_{Y \sim IV_x + \{IV_w + IV_z\}}^2 + \Delta R_{Y \sim IV_x + \{IV_v + IV_z\}}^2)^{\frac{1}{3}}$	$(\Delta R_{Y \sim IV_z + \{IV_w + IV_v\}}^2 + \Delta R_{Y \sim IV_z + \{IV_w + IV_x\}}^2 + \Delta R_{Y \sim IV_z + \{IV_v + IV_x\}}^2)^{\frac{1}{3}}$
Four IVs	$\Delta R_{Y \sim IV_x + \{IV_w + IV_v + IV_z\}}^2$	$\Delta R_{Y \sim IV_z + \{IV_w + IV_v + IV_x\}}^2$

The  $\Delta$  indicates that the values in Table 2 reflect that the  $R^2$  value is an increment made by the first IV beyond the subset of IVs in braces. For the first set of conditional dominance statistics, at one IV, the  $\emptyset$  indicates the null set with no IVs—thus, is the  $R^2$  with the IV by itself in a subset. Note that the conditional dominance statistics are not only averaged but also include all IVs in their computation. Hence, conditional dominance comparisons between two IVs' conditional dominance statistics will include increments from the IV against which they are being compared in their each average. This is a noteworthy difference from complete dominance designations that ignore sub-model  $R^2$  values that include the IV against which the focal IV is being compared.

The process of computing conditional dominance statistics for  $IV_x$  for a number of IVs in the model  $i$  can be formalized as in Equation 2 below.

$$C_{IV_x}^i = \frac{\sum_{g=1}^{l_i} \Delta R_{IV_x + \{o_g\}}^2}{l_g} \quad (2)$$

Where  $l_i$  is the number of combinations of size  $i$  given  $p$  IVs and  $o_g$  is a distinct subset of the  $p$  IVs of size  $i - 1$ , not including  $IV_x$ , that are included in the sub-model.

The process for determining conditional dominance between two IVs,  $IV_x$  and  $IV_z$ , can also be defined as in Equation 3.

$$IV_x D_c IV_z \quad \text{if} \quad p = \sum_{i=1}^p \begin{cases} \text{if } C_{IV_x}^i > C_{IV_z}^i & \text{then } 1 \\ \text{else } & 0 \end{cases} \quad (3)$$

The  $D_c$  in this case is a designation indicating conditional dominance of the left hand IV over the right hand IV.

When complete dominance is not achieved between two  $IV_x$  and  $IV_z$ , but conditional dominance is,  $IV_x$  tends to explain more of  $Y$  than does  $IV_z$  irrespective of inclusion order. Thus, when achieved, conditional dominance suggests that  $IV_x$ 's effects are not generally sensitive to the order in which it enters the model compared to  $IV_z$  and that it is more important conditional on the number of IVs in the sub-model.

Conditional dominance between two IVs is much less stringent than complete dominance but can still be a difficult designation to meet for IV pairs in models with a great deal of between-IV overlap. The final dominance designation involves yet another averaging step to allow reduce the stringency of the comparison and allow for compensation of values over inclusion orders.

### General Dominance

The least stringent dominance designation between IV pairs is called general dominance. General dominance further relaxes the stringency of the comparisons between IV pairs by changing the focus from comparing average increments grouped by the number of IVs in a sub-model to the arithmetic average of these averages. By averaging over conditional dominance statistics, general dominance allows higher contributions at specific numbers of IVs in the sub-model to compensate for lower contributions at other numbers of IVs in the sub-model. The values generated by general dominance almost guarantee that IVs will produce a dominance designation and permits an explicit rank ordering of IV predictive usefulness of each IV in the model.

The averaged conditional dominance statistics computed for determining general dominance are known as general dominance statistics. The single general dominance statistic computed for each IV Extending on Table 2, determining general dominance between  $IV_x$  and  $IV_z$  would involve both general dominance statistics. Both general dominance statistics are included below in Table 3 in their full computational form.

As is implied by the computations in Table 3, each general dominance statistics are a weighted average of the individual increments to the  $R^2$ s.

The process of computing the general dominance statistic for  $IV_x$  is formalized below in Equation 4.

$$C_{IV_x} = \frac{\sum_{i=1}^p C_{IV_x}^i}{p} \quad (4)$$

Using the general dominance statistics computed in Equation 4, the process of determining general dominance of  $IV_x$  over  $IV_z$  can also be defined as in Equation 5.

**Table 3.** Example General Dominance

Average with $IV_x$	Average with $IV_z$
$\Delta R_{Y \sim IV_x + \{\emptyset\}}^2 \frac{1}{4} +$	$\Delta R_{Y \sim IV_z + \{\emptyset\}}^2 \frac{1}{4} +$
$(\Delta R_{Y \sim IV_x + \{IV_w\}}^2 +$	$(\Delta R_{Y \sim IV_z + \{IV_w\}}^2 +$
$\Delta R_{Y \sim IV_x + \{IV_v\}}^2 +$	$\Delta R_{Y \sim IV_z + \{IV_v\}}^2 +$
$\Delta R_{Y \sim IV_x + \{IV_z\}}^2) \frac{1}{12} +$	$\Delta R_{Y \sim IV_z + \{IV_x\}}^2) \frac{1}{12} +$
$(\Delta R_{Y \sim IV_x + \{IV_w + IV_v\}}^2 +$	$(\Delta R_{Y \sim IV_z + \{IV_w + IV_v\}}^2 +$
$\Delta R_{Y \sim IV_x + \{IV_w + IV_z\}}^2 +$	$\Delta R_{Y \sim IV_z + \{IV_w + IV_x\}}^2 +$
$\Delta R_{Y \sim IV_x - \{IV_v + IV_z\}}^2) \frac{1}{12} +$	$\Delta R_{Y \sim IV_v + \{IV_w + IV_x\}}^2) \frac{1}{12} +$
$\Delta R_{Y \sim IV_x + \{IV_w + IV_v + IV_z\}}^2 \frac{1}{4}$	$\Delta R_{Y \sim IV_v + \{IV_w + IV_w + IV_x\}}^2 \frac{1}{4}$

$$IV_x D_g IV_z \text{ if } C_{IV_x} > C_{IV_z} \quad (5)$$

The  $D_c$  in this case is a designation indicating conditional dominance of the left hand IV over the right hand IV.

When complete and conditional dominance is not achieved between two  $IV_x$  and  $IV_z$ , but general dominance is,  $IV_x$  tends to explain more of  $Y$  than does  $IV_z$  but is sensitive to inclusion order. Hence, when achieved, general dominance suggests that  $IV_x$ 's predictive usefulness is on average, better than  $IV_z$  and that it is more important when ignoring the effect of inclusion order.

It is important to note that the general dominance statistic values, when summed across the  $p$  IVs, equals the sub-model  $R^2$  when all  $p$  predictors are included. This is a useful feature of these statistics that, as has been discussed in other reviews (Grömping, 2007; Johnson & LeBreton, 2004), ties this method to earlier work that has focused on the decomposition of the  $R^2$  statistic.

In the sections above, I have provided a discussion of DA's development and reviewed how DA statistics are computed and DA designations are determined. In the section below, I transition from a broad outline of DA to a more targeted discussion of the application of DA to CRMs. The focus of the section discussion of CRM-based DA focuses on considering which fit metric should be used when applying DA to CRMs.

## Applying Dominance Analysis to Count Regression Models

A complication of applying DA to CRMs is that most of the literature on DA has focused on its application to LRM with the variance explained  $R^2$ . Given the semi-continuous nature of count DVs and that the output from CRMs is typically in the form of a predicted count/mean value, the  $R^2$  could be applied to CRM-based DA.

Although the variance explained  $R^2$  could be applied, there are good conceptual reasons to choose another fit metric. In the section below, I provide a detailed rationale for the choice of a different metric, the deviance  $R^2$  (i.e.,  $R_{DEV}^2$ ), that better reflects how CRMs fit to data.



### Count Regression Fit Metric: Deviance $R^2$

CRMs such as PR and NBR are not only log-linear models but follow the discrete Poisson and negative Binomial probability distributions. That both of these CRMs follow a distribution that is not the Normal distribution is important to consider when evaluating model fit. This is because statistical models are fit using information about the data as applied to a probability distribution to find their most likely parameter values. Thus, choosing a fit metric that matches the probability distribution's underlying fitting criterion will best reflect how the model fits to data.

In considering how fit metrics reflect underlying fitting criteria, I find it instructive to first consider the LRM. LRM is based on a Normal or Gaussian probability distribution that has been shown to have least-squares as its fitting criterion which seeks to minimize  $SS_{residual} = \sum (Y - \hat{Y})^2$  or the residual sums of squares between the predicted values from the LRM and the observed DV. 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  $SS_{residual}$  can then be cast as  $DEV_{model}$  or a model deviance for the LRM.

Deviance computations apply to CRMs but can differ substantially from the LRM's least squares criterion. For example, the deviance for PR and a special case of the NBR\* is  $\sum Y \ln \frac{Y}{\hat{Y}} - (Y - \hat{Y})$ . Note that this PR-focused deviance value differs notably 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 DVs 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 deviance has no such constraint and penalizes discrepancies from observed values equally in either direction.

The difference between the deviance computations across LRM and CRMs in terms of their implications for model fit was first discussed by Cameron and Windmeijer (1996) who devise the  $R^2_{DEV}$  or deviance  $R^2$  outlined in Equation 6 below.

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

Where  $DEV_{null}$  is a deviance for an intercept- or mean-only model.

As Cameron and Windmeijer note, the explained variance  $R^2$  used in LRM is, in fact, a  $R^2_{DEV}$  for the LRM as it can be cast as  $1 - \frac{SS_{residual}}{SS_{total}}$  where  $SS_{total} = \sum (Y - \bar{Y})^2$  which is equivalent to  $DEV_{null}$  for the LRM. Consequently, applying the explained variance  $R^2$  to a CRM applies a fit metric meant for LRM's fitting criterion to CRMs. Because the  $R^2_{DEV}$  using each CRM's unique deviance computation necessarily fit better to each CRM's fitting criterion, it is then the fit metric I recommend for CRM-based DA.

In this section I have provided a rationale for the choice of a specific fit metric to use when applying DA to CRMs. In the sections to follow, I transition to the description

---

\*This special case is the NBR estimated using a quasi-likelihood method. Maximum likelihood methods require a more complex form given the estimation of the  $\alpha$  parameter.

of the methodology for applying DA to CRMs. This discussion briefly outlines the data generated for this purpose, describes the models estimated from the data, and also describes how DA statistics and designations were determined with the results.

### *Count Regression Model-based Dominance Analysis: An Analytic Example*

The goal of this section is to provide a walk through of an analysis of CRM-based DA using the  $R^2_{DEV}$  recommended in the previous section. This walk through is intended to be useful as a guide for researchers and analysts interested in applying this method to CRMs. In the section below, I begin by describing how I generated data for the walk through based on a PR model<sup>†</sup>. The methods used to generate these data are not directly relevant to the goals of this manuscript and, as such, have been put into an online supplement. Please note that the online supplement outlines the literal generation and development perspective behind the generation of these data in great detail. The section below is focused primarily on describing the conceptual nature of the data so that the reader can follow along prior to transitioning to analysis.

*Data Generation* The study described below intends to evaluate the effect of four different tailoring shop characteristics on the output of each shop. The data for this analysis describe a sample of 6,780 such tailoring shops over a two-week period of time. One count DV output of each of these shops is the number of sport jackets (called *SJ* below) finished by the shop. Sport jackets are garments developed from a template and are of standard sizes and measurements. As such, the jackets are expected to show little extra variation and were expected to be Poisson distributed.

There were four measured tailoring shop characteristics used to predict sport jacket output. The first characteristic was an expert's assessment of the shop's equipment reliability (called *ER* in results below). Equipment reliability reflects the ease of the tailoring shop's equipment use for staff and the extent to which it is likely to fail to operate when needed. The second characteristic was the sufficiency of tailoring assistant staffing levels (called *AS* in results below). Understaffed tailoring shops are likely to suffer from reduced productivity given the shortage of labor. A third characteristic is the managing tailor's level of skill (called *SL* in results below). More skilled managing tailors are more likely than less skilled managers to have designed processes that facilitate jacket completion. Finally, the last characteristic is the managing tailor's work experience (called *WE* in results below). Similar to skill-level, more experienced managing tailors are likely to have designed processes that improve garment production.

The means, standard deviations, and correlations between all four IVs and two DVs are reported below in Table 4.

One point of note about the 6,780 tailoring shops revealed in Table 4 is that, on average, each of them produced a single jacket in the two-week period under study. In addition, the variance of sport jacket production was consistent with its expected their

---

<sup>†</sup>An example using an NBR is available in the online supplement.

**Table 4.** Descriptive Statistics

Variable	Mean	Standard Deviation	Correlations				
			<i>ER</i>	<i>AS</i>	<i>SL</i>	<i>WE</i>	<i>SJ</i>
<i>ER</i>	−0.0355	1.1853	1.0000	0.4168	0.1446	0.2180	0.4419
<i>AS</i>	−0.0261	1.4494	0.4168	1.0000	0.1958	0.2850	0.4025
<i>SL</i>	−0.0056	1.5480	0.1446	0.1958	1.0000	0.3295	0.3393
<i>WE</i>	0.0144	1.9211	0.2180	0.2850	0.3295	1.0000	0.3973
<i>SJ</i>	0.9920	1.0040	0.4419	0.4025	0.3393	0.3973	1.0000

underlying distribution. Specifically, sport jacket production had a variance of 1 which closely matches the mean as is assumed of the Poisson distribution.

Given that the sport jacket DV fits the requirements of a Poisson distribution, in the next section I estimate a PR predicting it using the four shop characteristic IVs.

**Regression Results** The PR results using the four shop characteristic IVs to predict sport jacket counts are reported in Table 5. Table 5 includes coefficients ( $\beta$ ), standard errors (SE), the 95% confidence interval for the unstandardized coefficients, and the exponentiated coefficient or incidence rate ratio/IRR ( $e^\beta$ ).

**Table 5.** Poisson Regression Predicting Sport Jackets Produced

	$\beta$	SE	95% Confidence Interval		$e^\beta$
			Low	High	
<i>ER</i>	0.2463	0.0114	0.2239	0.2687	1.2793
<i>AS</i>	0.1320	0.0095	0.1133	0.1506	1.1411
<i>SL</i>	0.1264	0.0083	0.1100	0.1427	1.1347
<i>WE</i>	0.1167	0.0069	0.1032	0.1302	1.1238
<i>Intercept</i>	−0.1805	0.0142	−0.2084	−0.1529	0.8349

Table 5 shows that each of the IVs has a positive effect on the number of sport jackets and appear to be statistically significant (at the  $p < .05$  level) as is implied by the confidence intervals.

In terms of effect size magnitude, equipment reliability had the largest effect on sport jacket counts. The IRR value shows that each one point increase in equipment reliability led to a 27.9% increase in the number of sport jackets finished. Assisting staffing levels obtained the second largest effect on sport jacket counts with its IRR value showing that each one point increase in staffing levels led to a 14.1% increase in the number of sport jackets finished. Skill level and work experience obtained the smallest effect sizes as reflected by their IRRs. Skill level resulted in a 13.5% increase and work experience a 12.4% increase in the number of sport jackets produced in two weeks.

The IRR value for the model intercept reports on the average number of sport jackets when all IVs are 0; here at their means. The .8349 value obtained is similar to the overall mean of near 1 reported in Table 4. This .8349 value is also used as

a baseline prediction for the PR. For instance, the expected number of sport jackets with equipment reliability at 1 and all other IVs at 0 is  $.8349 * 1.2793 = 1.0681$  or just above 1 jacket. Contextualizing model predictions by reporting out specific predicted values is useful for non-linear models such as PR as they more clearly depict the multiplicative effects produced inherently by CRMs. Hence, reporting on and visualizing predicted marginal means adds value to the model interpretation process (see Rönkkö, Aalto, Tenhunen & Aguirre-Urreta, 2022, for a similar perspective). Although I do not report marginal means plots in the body of this manuscript for brevity, such plots are reported in the online supplement.

The PR modeling result reported in Table 5 shows that all four IVs have predictive effects on sport jacket production and, in addition, show different effect size magnitudes that indicate likely differences in the importance of each IV for explaining sport jacket production.

In the section below, I transition to the focal analysis of this work where I determine the importance of each IV for predicting sport jacket production using DA. In this way, the section below provides an empirical example that applies the designation and computational formulas in Equations 1, 2, 3, 4, and 5 to the PR model in Table 5.

**Dominance Analysis Results** DA designations and statistics are built from the collection of model fit statistics representing all sub-models. The four tailoring shop IVs used in this manuscript result in  $2^4 = 16$  sub-models. The results from all model sub-sets are reported below in Table 6 omitting the sub-model with no predictors as it will produce a 0 value.

**Table 6.**  $R^2_{DEV}$  by Sub-model

$SJ \sim ER$	0.1704
$SJ \sim AS$	0.1418
$SJ \sim SL$	0.1008
$SJ \sim WE$	0.1384
$SJ \sim ER + AS$	0.2209
$SJ \sim ER + SL$	0.2388
$SJ \sim ER + WE$	0.2530
$SJ \sim AS + SL$	0.2047
$SJ \sim AS + WE$	0.2200
$SJ \sim SL + WE$	0.1816
$SJ \sim ER + AS + SL$	0.2748
$SJ \sim ER + AS + WE$	0.2821
$SJ \sim ER + SL + WE$	0.2868
$SJ \sim AS + SL + WE$	0.2524
$SJ \sim ER + AS + SL + WE$	0.3113

A few points can be gleaned from the results in Table 6. For example, Table 6. shows that equipment reliability appears to be associated with larger  $R^2_{DEV}$  values and is likely to be an important variable consistent with its coefficient size. The  $R^2_{DEV}$  values associated with the remaining variables is harder to discern from the table but

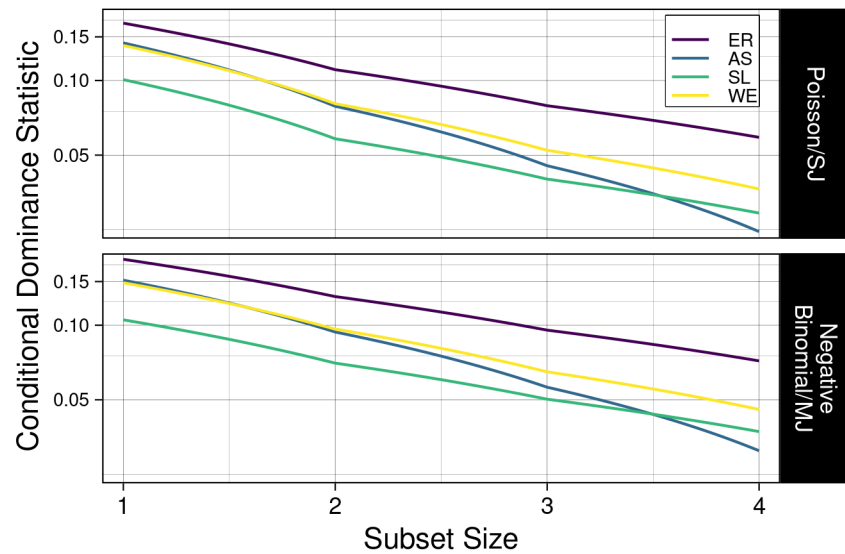
do not as clearly follow the coefficient sizes reported in Table 5. As such, it is useful to instead compare these values using the dominance analysis designations.

The first dominance designation to evaluate using the results in Table 6 is whether IVs show evidence of completely dominating one another. Equipment reliability was noted as having larger  $R^2_{DEV}$  values than other variables above and, when using Equation 1 to evaluate equipment reliability relative to the other three IVs, is shown to completely dominate all three. Thus, equipment reliability is the undisputed top predictor of sport jacket production irrespective of IV inclusion ordering in the model.

The complete dominance relationships between the remaining variables is far less clearly defined. For instance, work experience completely dominates skill level but does not completely dominate assistant staffing. Moreover, assistant staffing and skill level have no complete dominance relationship either. This makes clearly ranking the final three variables more nuanced as work experience is undisputedly better than skill level irrespective of inclusion order, but no such clear statement can be made of the rest of the comparisons. Thus, work experience is likely the 'second best' predictor but its will require the one of the expanded dominance designations to confirm this is the case.

In order to rank the remaining variables, the conditional dominance statistics for all the tailoring shop IVs were computed using Equation 2. The results of the conditional dominance statistic computations are depicted graphically in Figure 1 below. For this Figure 1, the y-axis is on a log scale to improve legibility nearer subsets sizes of 4.

**Figure 1.** Conditional Dominance Statistics



The graphic format is useful for the reporting of conditional dominance statistics with only a few IVs as the relative orientation of each of the lines reflects the nature of the conditional dominance designation implemented in Equation 3. Specifically,

if an IV's conditional dominance trendline is always above, and thus never crosses over, another IV's conditional dominance trendline, the IV that is above conditionally dominates the one below. Figure 1 confirms the complete dominance results in that equipment reliability's line is above the lines for the other three IVs indicating that it dominates each of them. Similarly, work experience's line is above skill level's line consistent with the complete dominance results.

The conditional dominance relationships among the final three variable comparisons do not offer a different designation than did the complete dominance results. Work experience fails to conditionally dominate assistant staffing given assistant staffing's conditional dominance statistic at subset size 1. Similarly, assistant staffing fails to conditionally dominate skill level given skill level's conditional dominance statistic at subset size 4. These conditional dominance results further reinforce the idea that attempting to rank these IVs 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 assistant staffing with skill level and work experience, I proceed to evaluating the general dominance designations between these variables. The general dominance statistics for each tailoring shop IV was computed using Equation 4 and are reported in Table 7.

**Table 7.** General Dominance Statistics

<i>ER</i>	0.1048
<i>AS</i>	0.0726
<i>SL</i>	0.0570
<i>WE</i>	0.0769

Evaluating the general dominance designations implied by the general dominance statistics in Table 7 using Equation 5 again shows that equipment reliability dominates each other variable and that work experience dominates skill level. The general dominance statistics add to the prior dominance designations by determining that work experience also generally dominates assistant staffing and that assistant staffing generally dominates skill level. The general dominance designations then allow for a clear rank ordering of the tailoring shop IVs in predicting sport jackets with equipment reliability being the most important followed by work experience, then assistant staffing, and ending with skill level.

In conclusion, the DA results have built on reports of the model in Table 5 by adding additional information about how each of the tailoring shop IVs explains variation or information in the sport jacket production variable. The DA results work to support the inference about the large IRR value obtained by equipment reliability, confirming that this variable, regardless of the order in which it might be included in the model, is most important. In addition, equipment reliability is associated with nearly one-third (i.e.,  $\frac{.1048}{.3113} \approx \frac{1}{3}$ ) of the explained information in sport jackets alone. The DA results also provide useful contextualization of the other three IV's IRR values from Table 5. An important point of clarification is on the importance of the work experience variable which, despite having the lowest IRR value, was generally recognized as the

second most important IV in the model. Although work experience's effect changed depending on the order in which it was included when comparing against assistant staffing, its large variance reported on in Table 4 likely resulted in it superseding the other two IVs in terms of explained information in sport jacket production. Finally, the DA results also provided useful information about the stability of the assistant staffing effect relative to skill level. Assistant staffing obtained the second largest IRR value but, as was most clearly revealed in Figure 1, its usefulness in terms of variance explained was the most dependent on the order of inclusion of all of the IVs. As such, it was only able to generally dominate the skill level IV.

*Section Summary* This section provided an extensive analytic walk through of a data analysis in a PR model using simulated data. The PR model described in this section interpreted model coefficients using IRRs and applied the recommended DA methodology using the  $R^2_{DEV}$  to dominance statistics and designations.

The data analytic conditions under which the PR model above was generated and estimated assumed that the time window under which each tailoring shop was observed was two weeks. In many real world situations involving count DVs, the time window might vary from observation to observation. Under such conditions, the observations are "exposed" to the count generating mechanism differently. In the section below, I shift the focus to a conceptual discussion of what exposure means for count DVs, how differential exposure is accommodated in CRMs, and offer an example of differential exposure affects DA statistics and designations with another data analytic example.

## Additional Consideration for CRMs

CRMs are defined by their accommodation of the semi-continuous and non-negative values on which count DVs take. Count DVs, in addition to these two facets, are also marked by an implied level of aggregation. Note that all the count DVs discussed in the introduction, as well as the simulated DV in the example above, all aggregate a number of events over a specific period of time.

That count DVs as aggregated events is a common interpretation of a Poisson distributed variable. Indeed, Poisson variables can be defined as accumulated repeated sampling from a Binomial distribution with a set probability of occurrence and a set number of trials (e.g., Hodges & Le Cam, 1960). In this interpretation of the Poisson, the Poisson mean parameter  $\mu$  is the product of the number of trials  $n$  for the Binomial distribution times the probability  $p$  of the event occurring in the Binomial distribution. Given the Binomial interpretation of the Poisson, it is clear that CRMs assume that the counts of events realized in the data are derived from Binomial processes that share the same number of trials and probabilities of occurrence. The linkage between aggregated Binomial processes and the negative Binomial distribution is even clearer as it is, in fact, defined as aggregated Binomial events or trials. In the end, both PR and NBR models can be seen as a method for modeling aggregated event counts that stem from a specific set of Binomial processes.

That count DVs can differ in their exposure to, or the number of trials in producing, the observed count values is can be accommodated in the analysis assuming that the differences in exposure to the count generating processes are encoded into a variable

that can be used to adjust the results of the CRM. In the section below, I elaborate more on what differential exposure is, the practical effects of differential exposure on answering research questions with count DVs, and how to adjust CRM results to ensure that count DVs that might be affected by differential exposure produce accurate DA statistics and designations.

### *Differential Exposure and Offset Terms*

Analysis of data with CRMs is, fundamentally, a method to answer research questions. Consider, for example, evaluating whether a customer service training increases the number of positive reviews for a business. From the perspective of the Binomial interpretation of count events like positive reviews, the researcher must consider whether the training will affect the probability of receiving a positive review from any given customer, the number of customers who can provide a review, or both factors. For this research question, the customer service training is likely to be most relevant to the probability of receiving a positive review from any given customer. Thus, with a fixed set of customers, the customer service training should increase the count of positive reviews due to increasing the probability that any one customer is happy and thus provides a positive review.

A complication that can easily arise in evaluating a research question like the one above is that two business locations can have different numbers of customers. If location A has 100 customers and location B has 1,000 customers, I would expect that location B would have more positive reviews than location A given the same underlying probability of observing the positive review event. The count DV, however, does not naturally contain this information and, in fact, the default CRM assumption is that the two locations have the same number of customers—a fact known not to be true.

That both locations have different numbers of customers means that the locations have different levels of exposure or 'differential exposure' to the count generating probability. The differential exposure in locations A and B is a problem when interpreting the count DV for answering the research question. Consider what would happen in a CRM if locations A and B both receive 10 positive reviews. I know that both values of 10 imply starkly different probabilities of observing an event (i.e.,  $\frac{10}{1000} = .01$  and  $\frac{10}{100} = .1$ ). The CRM, however, does not have any built-in way to distinguish between the location's values given just the count DV values and they would be erroneously treated the same in the model. What the CRM needs is to be instructed how to adjust the results for the known differential exposure that obscures the differences in the underlying probability of event occurrence.

CRMs can control for differential exposure using an offset term. The offset term, for a log-linear CRM, is natural log transformed version of a variable that reflects the known differential exposure between observations set to a coefficient value of 1. Thus, to control for differential exposure, the researcher must have access to a variable that reflects, for each observation, the exposure that the observation had to the count generating process. When this exposure variable is natural logarithm transformed and included in the CRM with a coefficient of 1, the DV predicted by the model becomes the ratio of counts out of the value of the exposure or the rate of events given exposure.



The transformation of the count DV to a rate follows is easier to see when making a few algebraic manipulations to the model. Consider the most simple case of a CRM, an intercept only model that is fit with an offset term. An intercept-only model with an offset would, as a formula, look like:  $Y = e^{\beta_0 \ln X_{exposure} + \beta_1}$ , where  $X_{exposure}$  is the exposure variable,  $\beta_1$  is the offset term, and  $\beta_0$  is the model intercept. Taking the natural logarithm of both sides of the equation and given the offset term is 1, the equation above simplifies to  $\ln Y = \ln X_{exposure} + \beta_1$ . Subtracting  $\ln X_{exposure}$  from both sides results in  $\ln Y - \ln X_{exposure} = \beta_1$ . Finally, the difference between two logarithms is the logarithm of their quotient, thus producing  $\ln \frac{Y}{X_{exposure}} = \beta_1$  or  $\frac{Y}{X_{exposure}} = e^{\beta_1}$ . The natural logarithm transformed exposure variable, when included with an offset term, necessarily results in the dependent variable being a rate.

The practical effect of the offset term is that the DV better reflects the underlying probability of the event and is no longer contaminated with different levels of exposure to the count generating process. An additional side effect of the offset term is that it reduces the extent of unexplainable variation, or deviance, in the count DV. The reduction in unexplainable variation arises as observations with high counts and high exposure values are pulled nearer those with low counts and low exposure values. As a result of the reduced unexplainable variation, inclusion of an offset term affects the values obtained  $R^2_{DEV}$  values and DA statistics on which they are based.

...ended here - need a good way to work the below in ... feels abrupt at current ...

It is also important to note that the relationships IVs have with the exposure variable can have a substantial impact on the estimated coefficients and DA statistics that can bias results and affect inference for probability-oriented count research questions.

The sections below extend the modeling discussed above by describing the generation of two new variables that have different patterns of exposure to the count generating process. The modeling process above is then repeated on these new variables to illustrate the effect that exposure variables, and modeling them as offset terms, have on both modeling and DA results.

### *Dominance Analysis with Offset Terms: Analytic Examples*

This section is an extension of the data analysis section above that discussed the PR and NBR modeling. Here, an additional two Poisson distributed variables are created that have different levels of exposure to the count generating process and different relationships with the IVs.

The next section provides a description of the variables and how they were generated. Full details on data generation are available in the online supplement.

**Data Generation** The data described in this section extended on the generated data produced for the analytic examples discussed previously. These data were then generated with the intention to balance their "realism" with built-in properties that illustrate the nature of the exposure effects.

The data simulated in this section again focused on the 6,780 tailors generating sport jackets in a fortnight. One key difference from the data above is that the period of time in which these data simulated was for two different local holidays. The first sport jacket variable occurred for near a religious holiday ( $SJ_{RH}$ ) that is not celebrated in the

same way by all tailors. The observation of this religious holiday produced inconsistent leaves of absence, and thus less exposure, across tailors that were uncorrelated with their IV values.

The second sport jacket variable occurred near a common time when tailors take voluntary holidays ( $SJ_{VH}$ ) that depends on tailor workload. Because tailors who produce more jackets are more likely to have their workload sufficiently light to take the holiday, the IVs are inversely related to the likelihood of taking holiday time and thus exposure. The sport jacket variable was otherwise identical to the previous version discussed above. The exposure days for religious holidays (RH) and voluntary holidays (VH) ranged from a minimum of 1 working day to 10 (i.e., all working days in a fortnight) and were uniformly distributed.

The means, standard deviations, and correlations with the IVs and one another are reported in Table 8.

**Table 8.** Descriptive Statistics with Differential Exposure

Variable	Mean	Standard Deviation	Correlations			
			$ER$	$AS$	$SL$	$WE$
$SJ_{RH}$	0.5447	0.7575	0.3455	0.3247	0.2773	0.3174
$SJ_{VH}$	0.4709	0.6549	0.2332	0.2325	0.1631	0.2405

Table 6 shows that the means and standard deviations for  $SJ_{RH}$  and  $SJ_{VH}$  were effectively identical but had shrunk from the value of SJ in Table 1. This is because SJ implicitly assumes an exposure value of 10 for all observations. Because the exposures for  $SJ_{RH}$  and  $SJ_{VH}$  were both around 5.5 days, their means shrunk proportional to the number of days of exposure. If 10 days produces a mean of 1, then 5.5 average days should produce an average of .55.

The pattern of correlations between the SJ-exposed variables and their exposures with the IVs is also of note. In the case of both  $SJ_{RH}$  and  $SJ_{VH}$ , their correlations with the IVs are smaller than SJ as reported in Table 1. This is due to a restriction in variance from reduced exposure time to the count generating process for these variables. Note also the differences between RH, the exposure variable for  $SJ_{RH}$ , and VH, the exposure variable for  $SJ_{VH}$ . RH is, as expected, effectively uncorrelated with the IVs. By contrast, VH shows strong, negative correlations with the IVs. It is these negative correlations that produce an additional suppressing effect on the correlations between SH\_VH and the IVs as the restriction in variance due to exposure is inversely related to counts. These negative correlations are also the reason  $SJ_{VH}$ 's mean is less than the expected value of .55.

In sections to come, the new  $SJ_{RH}$  and  $SJ_{VH}$  variables are analyzed in separate PR models. Both models are evaluated with and with out the use of their exposure variable modeled as an offset. The results below are subdivided by the type of exposure variable and begins with  $SJ_{RH}$ , the uncorrelated exposure variable.

**Uncorrelated Exposure Variable** The two models using  $SJ_{RH}$ , one with and one without the exposure variable modeled with an offset, are reported in Table 7. Only

coefficient and IRR values are reported in Table 7 as their values are the focus of this discussion.

**Table 9.** Uncorrelated Exposure Results

	Without Offset			With Offset		
	$\beta$	$e^\beta$	$C$	$\beta$	$e^\beta$	$C$
<i>ER</i>	0.2575	1.2937	0.0708	0.2657	1.3043	0.0863
<i>AS</i>	0.1521	1.1643	0.0541	0.1502	1.1621	0.0641
<i>SL</i>	0.1455	1.1566	0.0437	0.1515	1.1636	0.0545
<i>WE</i>	0.1271	1.1355	0.0551	0.1347	1.1442	0.0698
<i>Intercept</i>	-0.8170	0.4418		-2.5276	0.0799	

Table 7 shows that the coefficients and IRR values for the models with and without an offset term are very similar to one another with the exception of their intercept value. That the exposure variable does not noticeably affect the magnitude of the coefficients, despite the clear impact the exposure variable has on the *SJ* distribution's mean and variance relative to when there is no exposure effect, may come as a surprise. Indeed, the exposure variable works in a similar way to random/uncorrelated error variance which tends to shrink the magnitude of coefficients in LRM. The key difference between error variance in LRM and 'random' exposure variables in CRMs is the multiplicative nature of the CRM's coefficients. For an LRM, the additional error variance dilutes what a 1 unit change in an IV does in terms of units of change to the DV. For a CRM, the interpretation of the coefficient is that a 1 unit change in an IV does in terms of a percent change to the DV. Here the random variation from an exposure variable would have affect change disproportionately across levels of the DV to shrink the magnitude of the coefficients in the same way as in an LRM.

Compared to the coefficients, the intercept across both models does show substantial change with the addition of the offset term. The difference across the offset and no offset models is to re-scale the mean value in a way that reflects the new interpretation of the DV when an offset is estimated. The effect of the re-scaling is clearest when comparing IRR values across models. The IRR value for the no offset model is .4418, which is an IV-adjusted mean based on the .5447 mean value for *SJ<sub>RH</sub>*. The IRR value for the offsets model is .0799, which is an IV-adjusted rate based on the theoretical underlying rate of .1000. This theoretical rate is obtained from for *SJ* which is what *SJ<sub>RH</sub>* would be equivalent to given no exposure differences. In this case, *SJ*'s mean is 1 and implied exposure is 10 working days. The rate of counts for *SJ* is then  $\frac{1}{10}$  or .1000. Thus, the offset term has, as expected, re-scaled the DV from a count to a rate.

It is also important to note that the coefficient values obtained for the IVs for both models in Table 7 are, again with the exception of the intercept, similar to the values obtained for the non-exposure affected *SJ* variable in Table 2. Whereas the effects of differential exposure that are uncorrelated with the IVs has had a minimal effect on the estimated coefficients, the effect of uncorrelated differential exposure is more noticeable among the next set of results; the general dominance statistics for both models reported in Table 8.

Table 8: Uncorrelated Exposure Poisson-based General Dominance Statistics

	No Offset	Offset
<i>ER</i>	0.0708	0.0863
<i>AS</i>	0.0541	0.0641
<i>SL</i>	0.0437	0.0545
<i>WE</i>	0.0551	0.0698
Overall	0.2238	0.2748

Table 8 shows that the general dominance statistics for models with the offset term are all larger than their corresponding values without the offset term. However ostensibly different, the differences in magnitudes between these results is due to the differences across models in the magnitude of the  $R_{DEV}^2$  from which each IV has been partitioned. In fact, the relative proportions of the  $R_{DEV}^2$  explained by each of IVs within each model is nearly identical. For example, *ER* is responsible for approximately 31% of the  $R_{DEV}^2$  for both models (i.e.,  $\frac{.0708}{.2238} = .3164$ ;  $\frac{.0863}{.2748} = .3140$ ). The reason for the notable difference in explained deviance across models is because of the offset term's role in reducing unexplained variation in the DV for probability-based research questions. As was discussed previously, the exposure effect produces variation in the count DV that is unrelated to the focal probability of the occurrence of an event. The inclusion of the offset term reduces this unexplained variation directly and obtains results that are closer to those that would be obtained if exposure was equal. To this last point, note that the general dominance statistics and  $R_{DEV}^2$  for the offset model are much closer to the results of *SJ* in Table 5 than are the no offset model's results.

The section below moves to discussing the results of the use of exposure terms for the  $SJ_{VH}$  variable in which the IVs are related to exposure. As with the results for  $SJ_{RH}$ , the models with and without offset terms are reported.

*Correlated Exposure Variable* The no offset and offset models using  $SJ_{VH}$  are reported in Table 9. As before, only the coefficient and IRR values from each model are reported.

Table 9: Correlated Exposure Poisson Regression

	No Offset		Offset	
	Coefficient	IRR	Coefficient	IRR
Inter	-0.8512	0.4269	-2.5182	0.0806
<i>ER</i>	0.1638	1.1780	0.3054	1.3572
<i>AS</i>	0.1145	1.1214	0.1735	1.1895
<i>SL</i>	0.0642	1.0663	0.1668	1.1815
<i>WE</i>	0.1118	1.1182	0.1568	1.1698

As compared to the uncorrelated exposure results in Table 7, the correlated exposure results show different patterns of results across no offset and offset models for all IVs.

Whereas the offset model looks relatively similar to the pattern of results from the no offset model in Table 7 as well as the results in Table 2, the no offset model shows a notable shrinkage in the magnitude of the coefficient and IRR for SL and *AS* relative to other results. The magnitude of SL's coefficient across all models has been between *AS* and WE whereas for no offset model using  $SJ_{VH}$ , SL's value is notably lower than all other coefficients. *AS*' coefficient in the no offset model is also relatively smaller than in all other models to this point; nearly equal in value to WE's coefficient value. ... note sim of intercept...

The no offset model results in Table 9 show the biasing effect of the correlated exposure variable on coefficient estimates when it is not modeled as an offset. The notable change to coefficients due to the correlated exposure are likely to impact DA results as well. Table 10 below reports on the general dominance statistics derived from both models in Table 9.

**Table 10: Correlated Exposure Poisson-based General Dominance Statistics**

	No Offset	Offset
<i>ER</i>	0.0307	0.0889
<i>AS</i>	0.0280	0.0656
SL	0.0127	0.0512
WE	0.0342	0.0737
Overall	0.1055	0.2795

Table 10's results show, similar to those of Table 9, that the model including an offset is very similar to the results obtained from the offset model in Table 7 as well as  $SJ$ 's results from Table 5. Thus, the effect of a correlated exposure variable is relatively minimal on DA statistics given that the appropriate offset term is included in the model and reproduces results similar to a model not affected by differential exposure. By contrast, the no offset model shows both a large reduction in the magnitude of the  $R^2_{DEV}$  as well as a change in the rank orders of the IVs. The reduction in the magnitude of the  $R^2_{DEV}$  was produced, in part, for reasons similar to the uncorrelated exposure models as additional unexplainable variation was added to the count DV. In the case of the correlated exposure model, the extra variation in counts due to differential exposure also affected coefficient estimates as can be seen in Table 9. In all cases, the estimated magnitude of the IV's coefficients, when affected by differential exposure that was correlated with the IVs, was smaller than expected relative to the models without such differential exposure. The effect of correlated, differential exposure also changed the rank order of top two IVs. In the no offset model, WE obtains the highest general dominance statistic value followed by *ER*. The correlations in Table 6 reveal the probable reason for this effect in that WE both has the highest relationship with  $SJ_{VH}$  as well as the lowest correlation with the exposure variable VH. Hence, without the offset term adjustment, WE appears to have the strongest overall impact on  $SJ_{VH}$ .

**Section Summary** This section has discussed the effects of unequal exposure to a count generating process on both estimated coefficients as well as DA statistics. When

unequal exposure to the count generating process is uncorrelated with the IVs in the model, the effect of unequal exposure is minimal, mostly changing the total  $R^2_{DEV}$  value obtained as well as the model intercept. When unequal exposure to the count generating process is, at least in part, determined by the IVs in the model, the effect of unequal exposure is far more influential. The results in Tables 9 and 10 show that correlated, unequal exposure can bias coefficient estimates as well as explained deviance metrics, as well as change the rank order of IV importance.

Although unequal exposure can have notable biasing effects, such biases can be corrected for using the exposure variable as an offset term in the model. Modeled offset terms do not necessarily recover the effect of each IV that would have been obtained had there been no unequal exposure, but do result in notable improvements to the model estimates and DA statistics over the alternative where an offset term is not modeled. These results then show the importance of considering whether it is possible that observed counts in a CRM could result from different probabilities of the event occurring, different exposure to the count generating probability, or both. If it is the case that both factors might be producing counts and the focus of the research question is on the probability of events, it is important to attempt to model the exposure variable as an offset term to avoid potential bias in the results.

## Discussion

This manuscript provided a rationale for application of a specific methodology when using relative importance analysis with CRMs. I argue that DA, as an extension of Shapley value decomposition, is an applicable method to CRMs and in addition provided a fit metric for use in the relative importance analysis of any CRM. Specifically, I recommend the use of the deviance  $R^2$  or  $R^2_{DEV}$  as it is directly analogous to the explained variance  $R^2$  or  $R^2_{EXP}$ . Applying the  $R^2_{DEV}$  to CRMs then follows directly from the application of  $R^2_{EXP}$  to LRM.

In addition, this manuscript provides a discussion of one specific data analytic consideration of particular relevance to CRMs; differential exposure to a count generating process. In the discussion of this issue, I outline its relevance to the evaluation of count DV-based research questions and showed examples of how the concept affects both model coefficient estimates as well as DA statistics. The section focusing on differential exposure also showed the extent to which modeling such differential exposure variables as offset terms could correct the biasing effect and produces more accurate estimates of parameters and DA statistics.

In combination, the discussion of these issues supplies organizational scientists with the tools necessary to accurately evaluate IV relative importance as applied to a series of widely used CRMs. It is important when modeling count DVs that the recommendations provided by Blevins et al. (2015) are considered first as their guidance can point a researcher to the appropriate model to apply to the data. This article seeks only to pick up where they left off and recommend practices for the evaluation of the selected CRM after estimation.

This article touched on many of the key aspects necessary for extending DA to CRM. Below I discuss additional potential extensions of this work and future directions for the evaluation of IV importance in organizational science.

### *Extensions and Future Directions*

PR in particular, but also NBR, are instances of the broader class of models known as generalized linear models (McCullagh & Nelder, 2019). Specifically, PR is a generalized linear model in the Poisson "family" with a natural logarithm "link" (See implementation of generalized linear models in R; R Core Team, 2022). In extending DA to CRMs, I also effectively make an argument for the applicability of DA using the  $R^2_{DEV}$  to any generalized linear model using the same approach. That is, given that DA is applicable to CRMs, it is also applicable to any generalized linear model with a different distribution family or link function. This is an important point to note as it is not necessary for researchers to hesitate in applying DA to generalized linear models that have not been explicitly discussed in the literature. Indeed, because this article links Shapley value decomposition to DA and then provides a method for implementing DA with a specific generalized linear model (i.e., CRMs), it is trivial to extend the same approach to a generalized linear model with a different family and link function. Therefore, DA with the  $R^2_{DEV}$  approach is applicable to any generalized linear model.

It is worth noting before moving on that the  $R^2_{McFadden}$ , recommended by Azen and Traxel (2009) for use in DA with logistic regression generalized linear model, is equivalent in computation to the  $R^2_{DEV}$ . This is because the form of the deviance computation for a logistic regression produces a value that is a scalar multiple of its log-likelihood. As such, the perspective taken in the paper agrees with that of the past research in that recommends the use of  $R^2_{McFadden}$ . Additionally, Cameron and Windmeijer (1996) note that the  $R^2_{McFadden}$  differs from the  $R^2_{DEV}$  by a scalar multiple and is valid in application to CRMs and, by implication, the computation of DA statistics using CRMs.

Shapley value decomposition and, by extension, DA are well-known to be computationally expensive methodologies that require averaging of contributions by IVs to a model over a potentially extremely large pool of values. For example, a DA with 20 IVs would require  $2^{20} = 1,048,576$  different sub-models in order to collect fit statistics associated with all possible combinations of IVs and compute general dominance statistics. Whereas modern computing power could feasibly accommodate around one million models in a reasonable amount of time, adding additional IVs beyond 20 continues to grow the number of required models geometrically. An additional 10 IVs results in  $2^{30} = 1,073,741,824$  or over one billion model required to obtain dominance statistics; even if each CRM took  $\frac{1}{100}$  of a second to estimate, this many models would take nearly 4 months to complete on a single machine.

Many approaches have been developed to approximate Shapley values/DA that are far less computationally expensive. Relative weights analysis (Johnson, 2000) is a commonly applied alternative to DA that approximates the averaging process between all possible models using singular value decomposition. Relative weights analysis has shown some promise in past research (LeBreton, Hargis, Griepentrog, Oswald & Ployhart, 2007), and has been extended to one generalized linear model in published research (Tonidandel & LeBreton, 2011, i.e., logit regression) but has been subject to sharp criticism (Thomas, Zumbo, Kwan & Schweitzer, 2014) and is not among commonly recommended methods for relative importance of linear models (Grömping, 2007).

Shapley values have also grown in use among methodologists seeking an approach that can provide "interpretable machine learning" results (Molnar, 2020). In many cases, such efforts apply Shapley values, or approximations to them, that decompose predicted values for each observation in the data (e.g., Lundberg et al., 2020). Such Shapley decompositions of predicted values can be used, when cross-referenced with a observations' IV values, that, in the aggregate, can be used to get a sense for the nature of the effect of that IV on the DV much like a coefficient in a CRM. Machine learning models often have a great number of IVs which makes the "exact" computation of Shapley values infeasible. Methods to facilitate Shapley value computations in the literature have included sampling permutations (Castro, Gómez & Tejada, 2009) and normal approximations under specific assumptions (Fatima, Wooldridge & Jennings, 2008), among others. In application to machine learning methods, Shapley value approximations have often use a weighted-least squares approximation with sampling and adjustments for dependence (Sellereite & Jullum, 2020; Aas, Jullum & Løland, 2021). Such approximation methods may be useful in application to CRM and other models as an alternative estimator of the DA statistics.

Finally, to the knowledge of the author, there has been relatively little work has considered the value in applying Owen decomposition (e.g., Huettnner & Sunder, 2012) to DA broadly. Owen decomposition differs from Shapley in that a grouping structure is applied to IVs in which groups of IVs "negotiate" values jointly between groups but subsequently also "negotiate" values within their group as well. The result is that each IV obtains a component of the fit metric, and the Owen values are a full decomposition of the fit metric, but a large number of sub-models can be eliminated. For example, by grouping 8 IVs into 4 groups of 2, the number of sub-models needing estimation goes from  $2^8 = 256$  to  $2^4 + 2^2 * 4 = 32$ —an eight-fold reduction in the number of sub-models needing estimation. An approach such as this can directly (i.e., without approximation) and substantially reduce the size of the pool of sub-models to estimate and retains feasibility for computing dominance designations such as complete dominance.

## Limitations

This work provides a justification for why DA is applicable to CRMs and offers multiple, extensive data analytic examples that model and dominance analyze simulated data. The data simulated in this manuscript were designed to have specific properties that make them applicable to the different CRMs and CRM-specific data analytic issues that were valuable to discuss as applied to DA. The use of simulated data is not, in and of itself, a limitation. The limitation in using simulated data is in avoiding many of the decision points noted by Blevins et al. (2015) related to the selection of the appropriate CRM. Careful consideration of the decision to choose a specific CRM is important, but were avoided in this work to focus on modeling and DA with data that were unequivocally known to apply to specific models. In addition to not covering the model selection process, I also acknowledge that using simulated data adds additional complexity to following along with the methodology in the online supplement. The procedures used to simulate the various count DVs are complex but could add value



to understanding count DVs for interested readers with a working knowledge of R (R Core Team, 2022).

In addition, this manuscript provides two examples of how omitting a relevant offset term can result in unexpected changes to DA statistics and designations. The examples provided show how the omission of an offset term affects DA statistics in ways different from, and arguably more serious than, estimated model coefficients. The purpose of this section was to raise awareness of this issue as one that has received less attention among organizational scientists but has the potential to strongly affect importance designations. This manuscript makes no attempt to quantify or characterize the extent of bias that can occur when omitting an offset term as I consider doing so out of scope. Although a thorough evaluation of the potential bias introduced by omitting an offset term was not undertaken in this work, I see future research focusing on this issue as applied to DA of potential value and help researchers identify situations when such bias is more or less likely to occur. An evaluation of omitted offset terms could be taken on in a way similar to Braun, Converse, and Oswald's (2019) evaluation of measurement reliability and sampling error on dominance statistics in meta-analytic LRM.

Related to sampling error, prior work on DA has also suggested the application of bootstrapping to estimate standard errors for general dominance statistics and reproduce-ability values for dominance designations (Azen & Traxel, 2009; Braun, Converse & Oswald, 2019). Bootstrapping the dominance statistics and designations is possible for CRM-based DA in the same way as was outlined for logistic regression by Azen and Traxel (2009) but is not examined in the present work. A severe complication with the bootstrapping method as applied to DA is, as discussed above, that as the number of IVs increases, the DA itself becomes computationally infeasible—let alone the entirety of the DA process being repeated for a fixed number of bootstrap iterations. A useful extension on this research would be to evaluate approximations to a full DA-with-bootstrapping process that could provide information on sampling variability but would remain scale-able and computationally feasible with CRMs (and other predictive models) of typical sizes encountered in the literature.

Finally, this work does not directly address how to conduct DA with a commonly observed complication with count DVs, zero inflation. Zero inflation is a condition where the count DV shows more zero counts than would be expected from a standard Poisson or negative Binomial distribution (e.g., Blevins, Tsang & Spain, 2015). A zero-inflated model is often thought of as having an "opt out" process that forces counts to 0 owing to making them not possible. For example, as applied to the generation of sport jackets, if a process in which tailors shops had to close and thus could not produce any jackets was modeled, such IVs could be modeled in a zero-inflation model to recover the effects of the other predictors on the "count" process. An additional complexity that results from zero-inflated models is that it is possible to allow IVs to predict the count generating process, the opting out process, or both. In such circumstances the researcher may have to consider whether an alternative approach that emphasizes parameter estimates as opposed to IVs is better suited to the importance determination with DA; indeed, past work has used zero-inflated Poisson regression as an example of the application of parameter estimate-based DA (Luchman, Lei & Kaplan, 2020). I expect that the recommendations offered in this paper could extend in a straightforward

way to zero-inflated models as Cameron and Windmeijer (1996) do consider zero-inflated models in terms of  $R^2_{DEV}$  computation. Such extensions would still need to consider whether IV- or parameter estimate-based DA fits the research question they are asking of the zero-inflated model.

## Conclusion

DA is a useful post-estimation methodology for determining the importance of IVs in statistical models such as CRMs. This manuscript has provided argumentation for how and why DA could extend to CRMs, offered extensive examples focusing on the computation of DA statistics with simulated data, and provided recommendations for how to adjust a CRM-based DA when observations in the count DV may have different levels of exposure to the count generating process.

In combination, the content of this manuscript will provide organizational and other behavioral scientists the necessary steps for determining the relative importance of IVs in CRMs they estimate. It is also important to note that several software implementations that can accommodate CRMs are available in different software. For instance, the `domin` package in Stata (Luchman, 2021) as well as the `parameters` (Lüdtke, Ben-Shachar, Patil & Makowski, 2020) and `domir` packages in R (Luchman, 2022) can accommodate CRMs. Thus, combined with this article's recommendations, all the necessary tools for the DA of CRMs are available to researchers to better understand their count DV-based models.

## References

- Aas, K., Jullum, M. & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298, 103502.
- Azen, R. & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, 8(2), 129.
- 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.
- Braun, M. T., Converse, P. D. & Oswald, F. L. (2019). The accuracy of dominance analysis as a metric to assess relative importance: The joint impact of sampling error variance and measurement unreliability. *Journal of Applied Psychology*, 104(4), 593.
- 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.

- 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.
- Castro, J., Gómez, D. & Tejada, J. (2009). Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5), 1726–1730.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281.
- Fatima, S. S., Wooldridge, M. & Jennings, N. R. (2008). A linear approximation method for the shapley value. *Artificial Intelligence*, 172(14), 1673–1699.
- Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2), 139–147.
- Hodges, J. L. & Le Cam, L. (1960). The poisson approximation to the poisson binomial distribution. *The Annals of Mathematical Statistics*, 31(3), 737–740.
- Huettner, F. & Sunder, M. (2012). Axiomatic arguments for decomposing goodness of fit according to shapley and owen values. *Electronic Journal of Statistics*, 6, 1239–1250.
- Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research*, 35(1), 1–19.
- Johnson, J. W. & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational research methods*, 7(3), 238–257.
- LeBreton, J. M., Hargis, M. B., Griepentrog, B., Oswald, F. L. & Ployhart, R. E. (2007). A multidimensional approach for evaluating variables in organizational research and practice. *Personnel Psychology*, 60(2), 475–498.
- Luchman, J. (2022). Package ‘domir’.
- 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.
- 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. (2021). Determining relative importance in stata using dominance analysis: domin and domme. *The Stata Journal*, 21(2), 510–538.
- Lüdecke, D., Ben-Shachar, M. S., Patil, I. & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using r. *Journal of Open Source Software*, 5(53), 2445.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1), 56–67.
- McCullagh, P. & Nelder, J. A. (2019). *Generalized linear models*. Routledge.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
- 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*. Vienna, Austria: R Foundation for Statistical Computing.
- 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.
- Sellereite, N. & Jullum, M. (2020). shapr: An r-package for explaining machine learning models with dependence-aware shapley values. *The Journal of Open Source Software*, 5(46), 2027.
- Thomas, D. R., Zumbo, B. D., Kwan, E. & Schweitzer, L. (2014). On johnson's (2000) relative weights method for assessing variable importance: A reanalysis. *Multivariate behavioral research*, 49(4), 329–338.
- Tonidandel, S. & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*, 26(1), 1–9.