Regression with Archaeological Count Data

Brian F. Codding and Simon C. Brewer

ABSTRACT

Archaeological data often come in the form of counts. Understanding why counts of artifacts, subsistence remains, or features vary across time and space is central to archaeological inquiry. A central statistical method to model such variation is through regression, yet despite sophisticated advances in computational approaches to archaeology, practitioners do not have a standard approach for building, validating, or interpreting the results of count regression. Drawing on advances in ecology, we outline a framework for evaluating regressions with archaeological count data that includes suggestions for model fitting, diagnostics, and interpreting results. We hope these suggestions provide a foundation for advancing regression with archaeological count data to further our understanding of the past.

Keywords: archaeological statistics, computational archaeology, digital archaeology

Los datos arqueológicos a menudo vienen en forma de conteos. Comprender por qué los recuentos de artefactos, restos de subsistencia o características varían a lo largo del tiempo y el espacio es fundamental para la investigación arqueológica. Un método estadístico central para modelar dicha variación es a través de la regresión, pero a pesar de los avances sofisticados en los enfoques computacionales de la arqueología, los profesionales no tienen un enfoque estándar para construir, validar o interpretar los resultados de la regresión de conteo. Basándonos en los avances en ecología, aquí describimos un marco para evaluar regresiones con datos de conteo arqueológico que incluye sugerencias para el ajuste de modelos, diagnósticos e interpretación de resultados. Esperamos que estas sugerencias proporcionen una base para avanzar en la regresión con datos de conteos arqueológicos para mejorar nuestra comprensión del pasado.

Palabras clave: estadísticas arqueológicas, arqueología computacional, arqueología digital

Archaeological data often comes in the form of counts: the number of ceramic sherds, rabbit bones, projectile points, tin cans, or residential structures. Understanding why these counts vary across time or space is central to interpreting past human behavior from archaeological data. One common method is to examine this variation formally through regression analysis.

For example, an investigator may want to know how the number of obsidian artifacts at archaeological sites varies with distance from the volcanic source. To evaluate this relationship using regression, counts of obsidian artifacts are referred to as the "response" or "dependent variable" (also known as "*y*"). When plotted, the response is shown on the vertical axis (see Figure 1). The distance from the volcanic source is referred to as the predictor or independent ("*x*") variable. When plotted, it is shown on the horizontal axis (see Figure 1). Arranged in this way, we are proposing that *y* (the number of obsidian artifacts) varies as a function of *x* (the distance from the volcanic source). In other words, we suspect we can learn something about the number of obsidian artifacts if we know something about how far that site is from the source. Yet another way to phrase this is that we expect the number of obsidian artifacts is *dependent* on the distance from the source.

This is an example of statistical inference: we can understand a population through a sample of observations. Such inference is critical in archaeology given that we only have a sample of past human behavior represented by material remains, but we want to leverage that material to broadly understand patterning in our history. Of course, regression alone does not provide a causal inference. As Pearl and McKenzie (2018) note, "Data do not understand causes and effects; humans do." Inferring causality must come from carefully thinking through both theoretical links between variables and the factors that may complicate causal relationships (for more, see Pearl and McKenzie 2018). In our hypothetical example, our inference is that variation in counts of obsidian artifacts observed in our sample of archaeological sites represents a general relationship of how past individuals acquired, transported, used, and discarded obsidian depending on how far they were from the source, which we have solid theoretical foundations to expect (e.g., Beck et al. 2002; James et al. 2022; Shott et al. 2015). Regression analysis makes this inference possible by estimating the underlying "unknown parameters" that structure the observed relationships between *x* and *y*.

The most common way to estimate these unknown parameters is through ordinary least squares (OLS) regression, which does so by determining the slope and *y*-intercept that minimizes the sum of squared error (see Table 1 for definitions). However, count data are intrinsically unsuited to OLS given that counts are discrete and cannot go below zero, and their distributions generally have fewer cases as counts increase (i.e., are right skewed), leading to

Advances in Archaeological Practice 12(2), 2024, pp. 163–172

DOI:10.1017/aap.2024.7

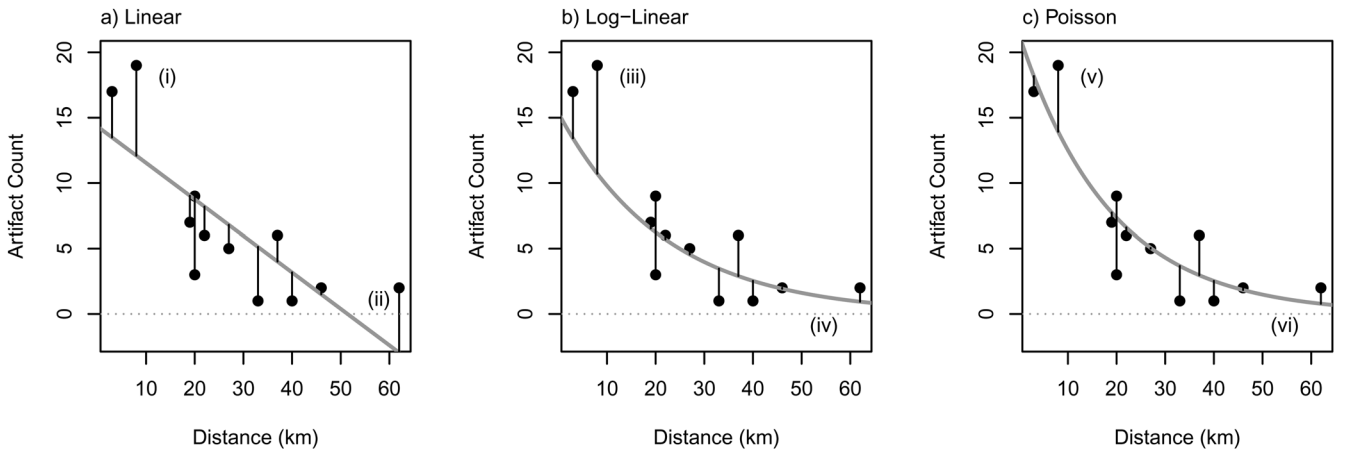

FIGURE 1. Results of (a) linear (ordinary least squares [OLS]), (b) log-linear (OLS with a logged response variable), and (c) Poisson regression predicting counts of obsidian artifacts across hypothetical archaeological sites as a function of the distance from the volcanic source. Black dots show the observed values at each site. Gray solid lines show the predicted model fit. Black vertical lines show the distance between the predicted and observed value for each site (the residuals), which the model is trying to minimize. Gray horizontal dashed lines indicate zero. See Supplemental Text 1 for a more formal comparison of model fits.

violations in OLS regression assumptions and poor model fits. These issues are illustrated in Figure 1a, where we estimate the relationship between our hypothetical obsidian artifact counts and distance from the source using OLS: the model underpredicts both high values (see i in Figure 1a) and low values (see ii in Figure 1a)—including predicting counts below zero, which is impossible.

To accommodate the disconnect between OLS and count data, investigators often log-transform their counts before running regression (e.g., log obsidian artifact counts) so that the response more closely approximates a normal distribution, leading to a better model fit. As shown in Figure 1b, fitting OLS to the logged counts of obsidian artifacts in our hypothetical example does avoid predicting values below zero (see iv in Figure 1b), but the model fit still underestimates high values at sites close to the volcanic source (see iii in Figure 1b). More generally, logging counts can generate additional problems that complicate modeling decisions (e.g., what to do with counts of zero that cannot be logged) and moves analysis farther away from the empirical observation (i.e., counts of things), which can complicate the interpretation of results.

An alternative method for estimating the unknown parameters that structure the underlying relationships between *x* and *y* is through generalized linear models (GLM). These are extensions of linear models that can accommodate a variety of data types and that use an iterative process to estimate the unknown parameters through a maximum likelihood function (see definitions in Table 1), which gives the probability of observing the empirical data. As noted by the name, these models can be thought of as a generalization of OLS to accommodate more diverse data types and relationships by specifying an *error distribution* and *link function*, respectively. GLMs with count data typically specify a Poisson error distribution to account for the characteristics described above and a log link function to linearize the relationship between the counts and covariates (more below). As seen in Figure 1c, modeling our hypothetical obsidian artifact counts as a function of distance from the source using this technique produces the best model fit and does not violate any model assumptions; the model does a good job of predicting high counts at sites close to the source (see v in Figure 1c) and does not predict counts below zero (see vi in Figure 1c). In other words, count regression does a better job of inferring the underlying relationship between *x* and *y*, which is the goal of applying regression to answer archaeological questions in the first place.

There is also a second issue that arises when applying standard OLS to archaeological count data: such counts are often observed over varying periods of time or areas in space. Because larger sampling windows should have more counts—all else equal—this needs to be accounted for in any modeling exercise or else the investigator risks coming to spurious conclusions (e.g., the smallest site may have the lowest count but the highest density). To accommodate this issue, investigators may divide their counts by the sample window to transform them to rates or densities (e.g., ceramic sherd density by excavation volume), or they may divide them by the sum of that object class to create a sample-size normalized ratio (e.g., proportion of pottery type; e.g., Fulford and Hodder 1974; Hodder 1974). In doing so, the response variable is no longer a count; it is either a rate, density, or proportion. This generates further problems for using OLS regression because rates and densities are equally likely to have skewed distributions that lead to poor model fits (e.g., Figure 1a), and proportions cannot go above one or below zero, yet OLS will predict both. Moreover, such abstraction can obfuscate interpretation (i.e., What does a change in density actually mean about human behavior? Do changes in proportions reflect variation in the numerator or denominator?).

These problems are not unique to archaeology. Ecology also often deals with counts that may be sampled across varying windows of time or space. For example, these could be counts of endangered animals or invasive plants, the number of times bees visit a flower, or the duration of toxic algae blooms. Over the past several decades, ecological informaticians have debated

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

Table 1. Definitions.

Term	Definition
Deviance	How well the proposed model accounts for variation in the response relative to a saturated model that fits the data perfectly. Measured as twice the log-likelihood (2LL) of the saturated model minus the 2LL of the proposed model. Analogous to residual variance in OLS.
Deviance residuals	The contribution that each individual observation has to the overall residual deviance. Calculated as the square root of the 2LL of an observation in a saturated model minus the 2LL in the proposed model. This is analogous to residuals in OLS regression.
Generalized linear model (GLM)	A family of models that estimate the unknown parameters that best fit to a set of observations using maximum likelihood based on a specified distribution and link function established determined by the nature and underlying distribution of the response variable.
Imputation of missing data	To replace missing cases with values estimated by other information. This could simply be the median or mean value of the variable, or it could leverage information from other variables.
Likelihood	How likely an observation or set of observations is, given parameters specified in the model.
Likelihood ratio test	An analysis of goodness of fit by examining how well the proposed model performs relative to a simpler, baseline model (usually a null model).
Likelihood r-squared (r_l^2)	A measure of goodness of fit. Also referred to as McFadden's R-squared, r_l^2 is calculated as one minus the proportion of the residual deviance (not accounted for by the model) over the null deviance. This is equivalent to the proportion of the proposed model deviance over the null model deviance.
Maximum likelihood	The value of the likelihood corresponding to the best, unbiased estimate of model parameters.
Negative binomial distribution	A probability distribution describing nonnegative integers or counts, with each observation being independent from one another. Unlike the Poisson distribution, this includes a separate parameter that describes the scale of the variance.
Null model	A model that only accounts for variation in the response variable with the *y*-intercept, estimated as the sample mean.
Ordinary least squares (OLS) regression	A common regression model that estimates unknown parameters by minimizing the sum of squared error or residuals.
Overdispersion	When variance in the data (i.e., counts) is greater than the mean.
Poisson distribution	A probability distribution describing nonnegative integers or counts, with each observation being independent from one another. This is defined by a single parameter (λ) that represents both the mean and variance of the distribution.
Predictor variables	The independent variable(s) or *x* variable(s) that are hypothesized to lead to variation in the response or dependent (*y*) variable.
Residuals	Deviations from observed values and those predicted by a model. Two common forms are Pearson's residuals and deviance residuals.
Residual degrees of freedom	The number of observations less the number of model parameters. Higher values indicate that the model is less dependent on this exact dataset and has greater generalizability.
Response variable	The dependent or *y* variable being approximated by the predictor or independent or *x* variables.
Saturated model	A theoretical model with a parameter for each data point (i.e., perfect fit). Used to estimate the total deviance in a set of observations.
Sum of squared residuals	The sum of the squares of residuals. Also referred to as the error or residual sum of squares.
Variance inflation	The degree to which model coefficients are higher than they should be due to correlation between multiple predictor variables.
Zero inflation	Occurs when the model systematically under- or overfits zeros relative to the observed number of zeros.

strategies and methods to deal with these issues common to count data.

Here, we draw on this literature to recommend analytical guidelines for archaeologists following two related principles: (1) keep the data as close to the unit of observation as possible, and (2) let the model do the work to accommodate peculiar data structures (i.e., model reformation, not data transformation; St-Pierre et al. 2018). This is because models designed for count data tend to do a better job of describing observations than standard OLS regression operating on log-transformed counts, and they are less likely to violate model assumptions (O'Hara and Kotze 2010), especially when counts are small (Warton et al. 2016).

For the remainder of this article, we offer a step-by-step approach to model fitting, diagnostics, and interpretation with archaeological count data, drawing on examples from the literature. Reproducible examples of all cases are available in Supplemental Text 1. The overall workflow is summarized in Figure 2. An extended version of this flowchart for models with multiple predictor variables is in Supplemental Text 1 (Section 5).

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

STAGE 1. EXPLORATORY DATA ANALYSIS

The research process should begin with asking a clear question, identifying the response and predictor variables needed to answer the question, and collecting or compiling data on those variables (see, for example, Zuur and Leno 2016). Researchers should have a clear rationale as to why they expect the response to vary as a function of the predictor, and they should identify any potentially confounding variables. This stage should also include a consideration of the size of the sampling windows for each count. If these vary, then data to represent that variation should also be collected (e.g., length of time window, area of spatial unit, size of population). Once the data are organized (e.g., in simple form, a table with cases in rows and variables in columns, or a more complex relational database), begin with some exploratory data analysis (see Chambers et al. 1983; Drennan 2009; Tukey 1977; Zuur et al. 2010; Figure 3). Figure 3 provides an example where we plot a simple histogram of the response variable to visualize its distribution, and we plot the bivariate relationships between the response and predictor. These exercises may reveal valuable information, such as outliers that are worth checking for data entry error or patterning that suggests nonlinear responses (e.g., Shennan 1997:161). Once the investigator has a good sense of what the data look like, move on to regression analysis.

STAGE 2. REGRESSION WITH COUNT DATA

The most common form of regression with count data is through GLMs (McCullagh and Nelder 1989). As noted above, these extensions of linear regression estimate the unknown parameters through maximum likelihood (see Table 1). GLMs are

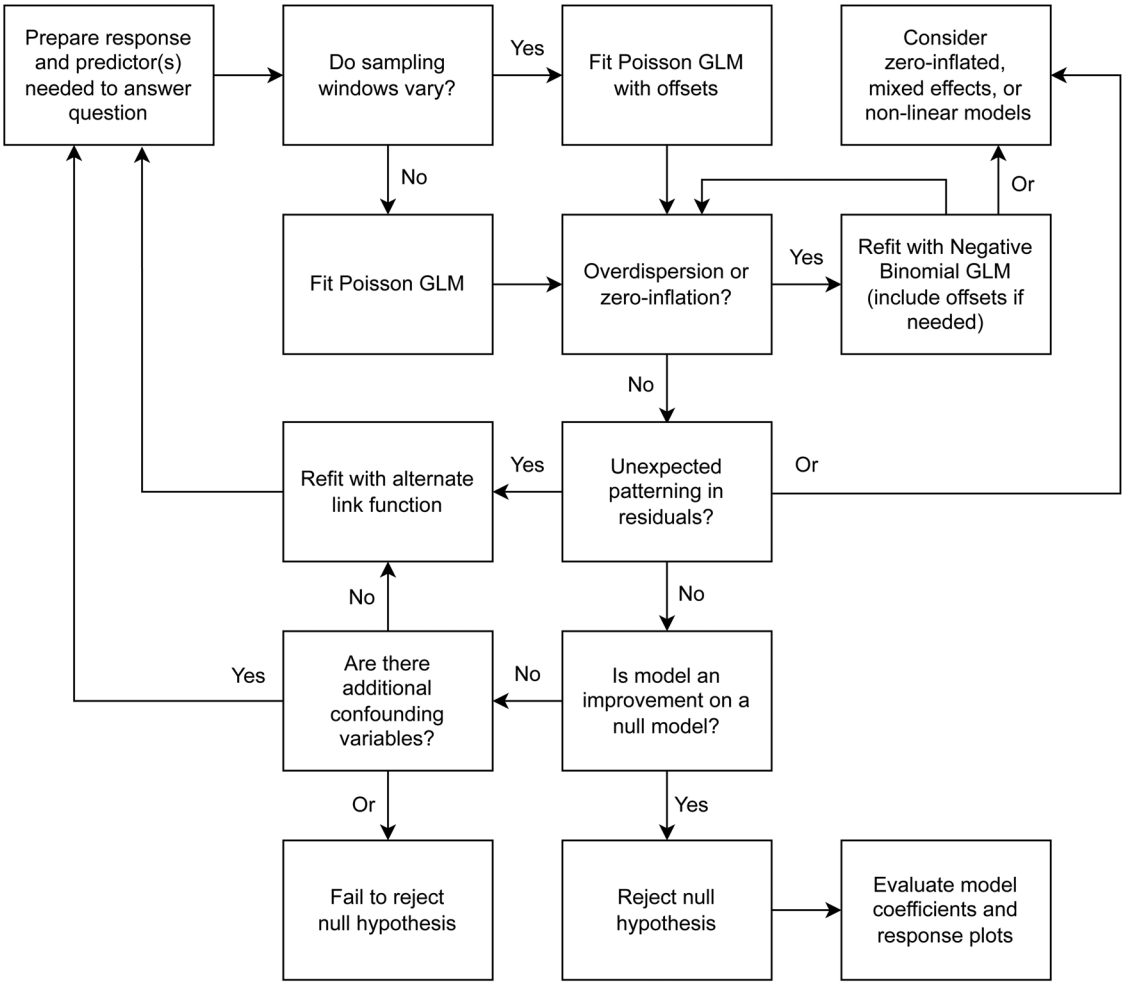

FIGURE 2. Flowchart outlining recommended procedures for fitting and evaluating count regression models. At each box or node, practitioners are directed to complete a specific step. Where a decision is needed based on the results of that step, practitioners are directed to proceed whether the answer is a "yes" or "no" to the question in the box. If practitioners follow a loop and return to that node, they may be presented with an "or" option for further model revision. This is not exhaustive, but it provides guidance on model fitting and diagnostics that archaeologists are likely to encounter. An example flowchart for models with multiple predictors is available in Supplemental Text 1.

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

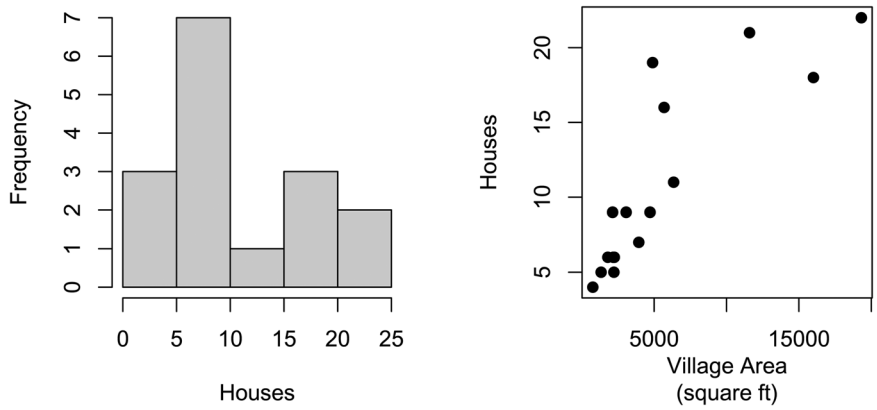

FIGURE 3. Examples of exploratory graphical data analysis examining the number of houses in Yurok villages and how they vary by village size (data from Cook and Treganza 1950; Waterman 1920).

generalizations in that they allow the specification of an error distribution that best fits the nature of the response variable, whereas OLS requires the error to be normal. For count data, the most commonly specified distribution is a Poisson or negative binomial. GLMs also require specification of a link function, which indicates the expected relationship between the response and predictors. Count regression typically begins with a log link function to linearize the relationship between the counts and covariates. Other types of distributions and link functions for other data structures are described in standard introductory texts (e.g., Venables and Ripley 2002:184). Supplemental Text 1 provides several examples in the R environment for statistical computing (R Core Team 2024), which is a common platform for archaeological data analysis (e.g., Carlson 2017; Marwick 2018). Given their wide use, these models should be available in all standard statistical software packages. As a first example, we illustrate how the number of houses (and by proxy, people) in Yurok villages increases with total village area (Cook and Treganza 1950; Figure 3; see Section 1 in Supplemental Text 1). For a published example, Codding and colleagues' (2024a) model site counts as a function of effective precipitation to assess how foraging populations responded to drought in the Basin-Plateau region of North America.

Modeling Count Data with Varying Sampling Windows

As noted above, archaeological research often needs to account for varying sampling windows. Counts may be summed by survey blocks that vary in size or by stratigraphic units of varying volumes that were deposited over varying periods of time. In order to account for this, models can *offset* the count for each case based on the size of the sampling window. Because the link function that relates x to y is logged, the offset too must be logged.

It is also possible to normalize archaeological counts with offsets that use the total counts of material in that class (i.e., all ceramic sherds or all bones from animals in the same patch), which can potentially account for variation in spatial and temporal sampling windows simultaneously, assuming similar processes of deposition and fragmentation. This is akin to modeling proportions of a specific artifact over the total artifact class, but it focuses analysis on the outcome variable of interest as opposed to a ratio that can vary as a result of the numerator and denominator. An examination of how counts of aurochs and cattle (*Bos* spp.) bones vary across late Mesolithic to early Bronze Age sites in Europe (data from Manning et al. 2015) provides an example and can be found in Supplemental Text 1 (Section 4). For a published example, Vernon and colleagues (2024) examine counts of farming sites across watersheds of varying sizes to evaluate the environmental factors that limit maize agriculture across the greater US Southwest (also see Codding et al. 2024b).

Models That Include More Than One Predictor

Oftentimes, an investigator may want to examine how counts vary based on multiple predictor variables. Multiple regression requires additional checks on data before model fitting (see page 45 in Supplemental Text 1), especially checking for multicollinearity or covariance between predictors that may bias model results (see, for example, Dormann et al. 2013; Zuur et al. 2010). Informally, multicollinearity can be thought of as trying to use the same information to explain the outcome twice. Specifically, multicollinearity can inflate the variance of model parameters resulting in the misidentification of significant predictors (Dormann et al. 2013). If multicollinearity is present, consider dropping correlated variables in favor of the most explanatory predictor (e.g., precipitation over elevation to predict settlement decisions; Vernon et al. 2024) or conducting a dimensional reduction technique (e.g., principal component analysis) to generate uncorrelated composite predictors (although this may reduce interpretability because the predictors become abstractions of empirical values).

STAGE 3. MODEL DIAGNOSTICS

After fitting the model, there are several diagnostic checks that should be performed prior to assessing model results. We discuss three common diagnostics for all models, and one additional diagnostic for models with multiple predictors.

Overdispersion

First, consider checking the model for overdispersion. One assumption of Poisson regression is that the mean and variance of

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

the counts are equal, or at least roughly of the same order of magnitude. To check if the variance is greater than the mean, divide the sum of squared residuals over the residual degrees of freedom (see Table 1 and Section 3 in Supplemental Text 1). If this dispersion parameter is greater than one, then the model may be overdispersed. This can be formally tested using a χ^2 test (see Supplemental Text 1). If present, consider refitting the model with a negative binomial distribution instead of a Poisson (or a Poisson model with quasi-likelihood estimation), which adds a dispersion parameter to account for overdispersion and removes potential bias in the model results.

Zero Inflation

Second, consider checking if the model is over- or underfitting zeros, referred to as "zero inflation." This can result when there are a large number of zeros relative to values above one (e.g., when modeling the distribution of a very rare artifact or species that only occurs at a few sites). If the ratio of observed and predicted zeros is near one, then this should not be a problem for prediction. If the value is much higher or lower than one, consider refitting the model with a negative binomial distribution, which can better handle zero inflation, or consider other zero-inflated models (e.g., hurdle models; Zuur et al. 2009).

Residual Patterning

Third, check if there is unexpected patterning in the model residuals. Plotting the residuals by the fitted (predicted) values can give a sense of this. The points will likely show some patterning given the nature of count data (e.g., points in a diagonal line across the bivariate space), but should overall be roughly centered on zero (on the *y*-axis) across the full range of predicted values (the *x*-axis). Figure 4a shows what "acceptable" patterning may look like with Poisson regression on small sample sizes typical of archaeological data. When there is clear patterning in the residuals, it may indicate a misspecified link function or unaccounted for relationships (e.g., interactions) between predictor variables. This can be checked by specifying alternative links (i.e., identity instead of log) or by adding interaction terms between predictors (if there are multiple predictors). Patterning may also be driven by underlying structure in the observations not accounted for in the model, such as may occur with temporal or spatial autocorrelation (i.e., when neighboring points in time or space are more like one another than expected by chance). This may also result from group-level patterning, as shown in Figure 4b; for example, when modeling projectile point counts as a function of time across hunting and residential sites, where the former always have higher counts than the latter, regardless of time period. When present, consider including a group identifier as a factor (categorical) interaction term, or turn to mixed effects models that can account for underlying structure by treating groups as random effects (see Bolker 2015:Box 13.1; for step-by-step procedures, see Bolker et al. 2009:Box 4). Additionally, patterning in residuals may indicate that the underlying trend is nonlinear (Figure 4c). This could occur, for example, when modeling site counts as a function of precipitation in a region where farming was most productive at some intermediate level of rainfall. If this is the case, consider nonlinear extensions of GLMs, such as generalized additive models (Wood 2017). Mixed effects and nonlinear models can also account for temporal and spatial structure in the data (Simpson 2018; Wood 2017).

Variance Inflation

Finally, models with multiple predictors require additional diagnostics to assess variance inflation, or the degree to which model coefficients are higher than they should be due to correlated predictor variables (see Section 5 in Supplemental Text 1). This can be accomplished by calculating a variance inflation factor (VIF), which estimates how well each independent term can be predicted from the remaining set of independent variables. High VIF values indicate multicollinearity, often the result of having several variables that relate to the same underlying driver (e.g., precipitation and soil moisture).

STAGE 4. MODEL RESULTS

After running diagnostics, consider evaluating how well the model performs by comparing it to a null model, by assessing the goodness

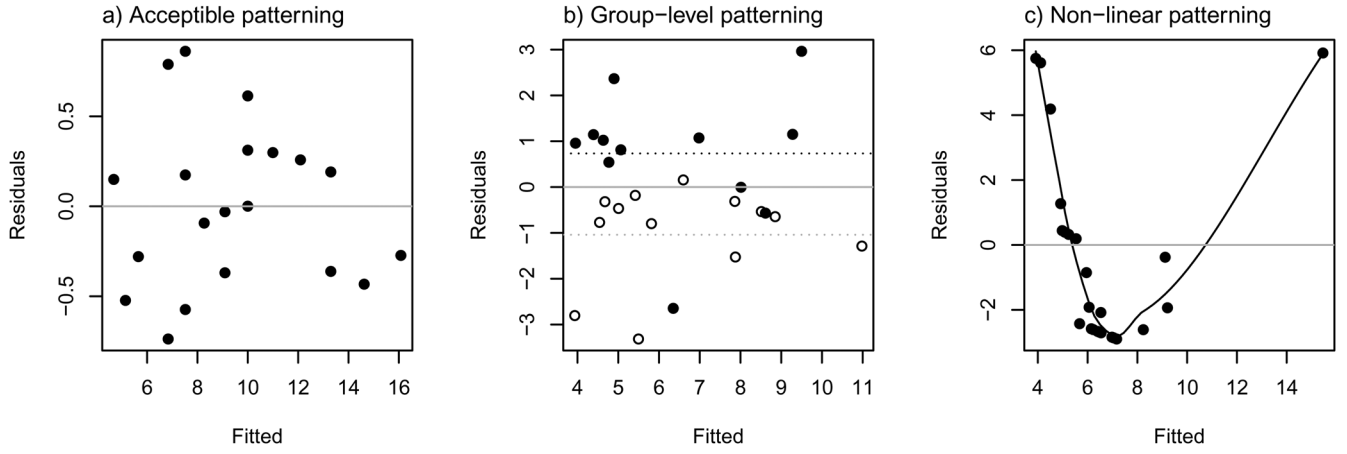

FIGURE 4. Examples of residual by fitted plots to examine (a) acceptable pattering in Poisson residuals, (b) patterning structured by between-group variation (dashed lines show group-level mean residuals), and (c) patterning structured by a nonlinear relationship between *y* and *x* not accounted for in the model.

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

of fit, and by evaluating the model coefficients. We also discuss how to examine variable importance with multiple regression.

Null Model Comparison

A first step is to compare the fitted model to a null model that includes only a parameter representing the *y*-intercept (compare Figure 5a and Figure 5b). This can be done by fitting a null model without any predictors and comparing it to the proposed model using a likelihood ratio test, which determines whether the proposed model is a significant improvement on the null. If not, then there is limited information gained about *y* from *x* in the proposed model; or the inclusion of the covariates does not improve our understanding of variation in the outcome.

Goodness of Fit

A second way to evaluate the model is to assess goodness of fit by estimating the proportion of variation in *y* that is accounted for in the model. In OLS, this is calculated as an *r*-squared (r^2) value. With generalized linear models this, is estimated as a likelihood r^2 or r_l^2 (see Faraway 2016). This is done by calculating the proportion of deviance accounted for in the proposed model over the deviance of the null model. This is equivalent to one minus the deviance of a saturated model that "perfectly" fits the data by theoretically including a term for each data point (Figure 5c) over the null deviance (Figure 5a). Michael Shott's (2018, 2022) ethnoarchaeological work provides an excellent case study to illustrate the difference between a null model (Figure 5a), a proposed or fitted model (Figure 5b), and a saturated model (Figure 5c). See Section 3 in Supplemental Text 1.

Model Coefficients

Next, evaluate the model coefficients. These include the estimated *y*-intercept—or the log value of *y* when *x* is zero—and how the response is predicted to change with each unit change in the corresponding predictor variable (e.g., the fitted line and where it crosses the *y*-intercept in Figure 5b). With default Poisson and negative binomial regression that use a log link, the coefficients are logged values. Taking the exponent of the coefficient returns it to the units of the response variable (i.e., a count). The interpretation of count regression coefficients differs from standard OLS regression, where a coefficient indicates a constant change in the response variable for a one unit increase in the independent variable. Here, the exponentiated coefficient is a multiplier, and it indicates the rate of change in the response with each unit increase in the predictor. A value of one indicates no change in the response; values above or below one indicate positive or negative relationships, respectively. To aid in interpretation, the exponentiated coefficient can be interpreted as a percentage change. For example, a value of 1.10 represents a 10% increase in the response variable for a one-unit increase in the independent variable. Returning to Shott's (2022) data, the fitted model response shown in Figure 5b illustrates the coefficient with predicted fit, which we will return to below (Section 3 in Supplemental Text 1).

Variable Importance

For models with multiple predictors, the relative importance of each variable can be assessed by examining standardized model coefficients. Because predictor variables may be on very different scales, it is often best to refit the model with scaled and centered predictor variables so that they vary on the same order of magnitude. Then, the investigator can compare the absolute values of scaled coefficients and their standard errors to assess which variables have the greatest influence on the response (see Section 5 in Supplemental Text 1). This can also aid in interpretation of the intercept parameter: with mean-centered covariates, the intercept represents the expected count under average conditions.

STAGE 5. MAKING PREDICTIONS

One of the main benefits of modeling data with regression is the ability to make informed predictions. This can be done by

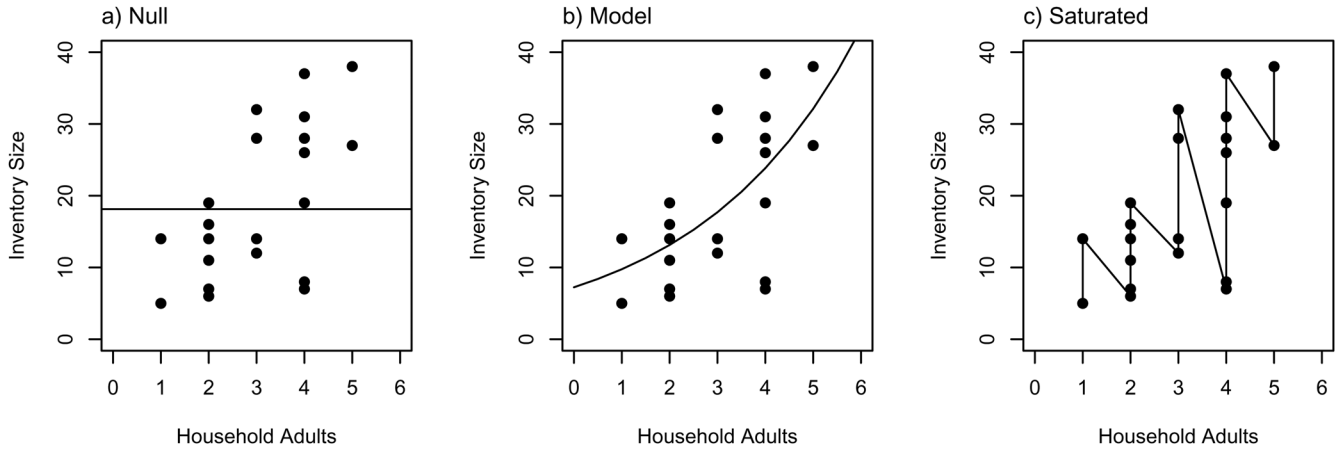

FIGURE 5. Data from Shott (2022) showing the relationship between pottery inventory size and the number of household adults in Michoacán overlaid with graphical representations of (a) null model, where only the mean of *y* is known; (b) the proposed fitted model; and (c) a saturated model with a parameter for every data point (i.e., a perfect fit). Generalized linear models compare how well the proposed model accounts for variation in *y* compared to the null (i.e., the log-likelihood of each value of *y* given *x* and the unknown parameters compared to the log-likelihood of each *y* value given only the *y*-intercept).

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

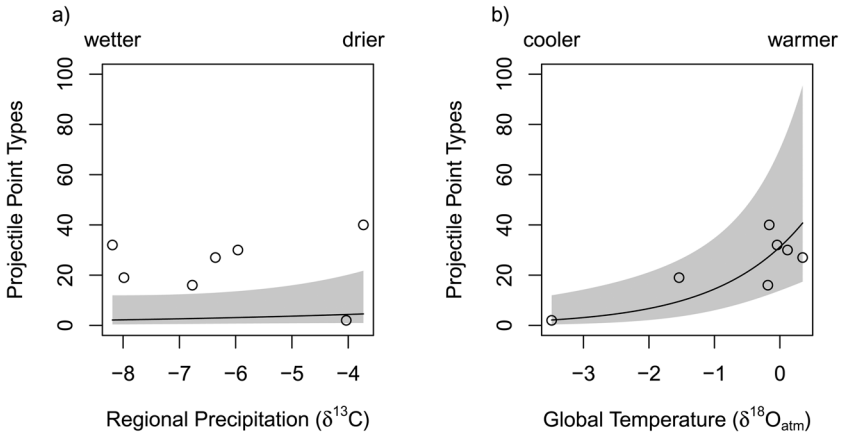

FIGURE 6. Partial response plots showing the predicted number of projectile point types per time period in Texas as a function of (a) regional precipitation inferred from sedimentary stable carbon isotope ratios and (b) global temperature inferred from atmospheric stable oxygen isotope ratios (data from Buchannan et al. 2016). Each panel shows the predicted response of projectile point type counts to the focal climate variable while holding the other climate variable constant at the minimum observed value. This is done to illustrate how regional precipitation (a) does not influence technological investment even under the coolest conditions, whereas (b) global temperatures promote increasing technological investment even under wetter conditions.

leveraging the model fit to estimate response variables across any of the predictor variables (e.g., Figure 5b). The investigator can also use the standard error of the prediction to provide confidence intervals around the model fit (see Figure 6 and Supplemental Text 1). However, when making predictions, it is important to recognize the limitations of the model, especially if considering making predictions outside the range of predictor variables included in the model fit.

With models that have multiple predictors, the investigator can estimate the partial response of the outcome variable to one of the predictors while holding the other predictor(s) constant. Such partial plots can also be used to evaluate different scenarios (e.g., how a response varies to population change under a constant environment, or vice versa). For example, Figure 6 shows how the number of projectile point types per time period in Texas (see Buchannan et al. 2016) varies across the range of regional precipitation during cold conditions, and across the range of global temperatures during wet conditions (Section 5 in Supplemental Text 1).

Other advanced methods that build on the basic principles discussed in this article use prediction for more complex problems. These include missing data imputation (Dakki et al. 2021) and generating probabilities of site occurrence across space or time in predictive models (see Bevan and Conolly 2009; Davis et al. 2020; Vernon et al. 2022; Yaworksy et al. 2020).

CAVEATS

The scenarios we discuss above are not exhaustive. Other issues may arise in the analysis of archaeological counts, such as differential breakage patterns. Additionally, there may be a priori reasons to decide to model ratios instead of counts, such as if evaluating theoretical predictions about resource trade-offs (e.g., Bayham 1979). Moreover, our message is not that OLS should never be used, but that it is more problematic than using models designed for count data (Warton et al. 2016). We encourage researchers to make informed decisions about which modeling approach is best, depending on their question and the nature of their data.

CONCLUSION

Archaeological data often occur as counts. Modeling these counts is therefore central to archaeological inference. Here, we draw on advances in ecology to make recommendations about potential best practices for analytical investigations. Our underlying message is that instead of contorting counts to meet model requirements, count data can be modeled using count regression, which is designed to handle such data and to deal with other issues that arise with using counts, such as varying sampling windows. We hope this tutorial provides a useful framework for advancing regression analysis of archaeological count data.

Acknowledgments

The authors thank Michael Shott for facilitating data access, and three reviewers for detailed and helpful comments on earlier versions of the manuscript. No permits were required for this work.

Funding Statement

The authors are funded in part by the National Science Foundation under grant BCS-2308299.

Data Availability Statement

All data and code are publicly available for others to copy and use in Supplemental Text 1, which is also posted on the Digital Archaeological Record (tDAR), an open-access repository (https://doi.org/10.48512/XCV8494446).

Competing Interests

The authors declare none.

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press

Supplemental Material

To view supplemental material for this article, please visit https://doi.org/10.1017/aap.2024.7.

Supplemental Text 1. Tutorial on count regression in R, including all data and code required to replicate analysis presented in the main manuscript.

REFERENCES CITED

Bayham, Frank E. 1979. Factors Influencing the Archaic Pattern of Animal Exploitation. *Kiva* 44(2–3):219–235.

Beck, Charlotte, Amanda K. Taylor, George T. Jones, Cynthia M. Fadem, Caitlyn R. Cook, and Sara A. Millward. 2002. Rocks Are Heavy: Transport Costs and Paleoarchaic Quarry Behavior in the Great Basin. *Journal of Anthropological Archaeology* 21(4):481–507. https://doi.org/10.1016/S0278-4165(02)00007-7.

Bevan, Andrew, and James Conolly. 2009. Modelling Spatial Heterogeneity and Nonstationarity in Artifact-Rich Landscapes. *Journal of Archaeological Science* 36(4):956–964. https://doi.org/10.1016/j.jas.2008.11.023.

Bolker, Benjamin M. 2015. Linear and Generalized Linear Mixed Models. In *Ecological Statistics: Contemporary Theory and Application*, edited by Gordon A. Fox, Simoneta Negrete-Yankelevich, and Vinicio J. Sosa, pp. 309–333. Oxford University Press, Oxford.

Bolker, Benjamin M., Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution. *Trends in Ecology and Evolution* 24(3):127–135.

Buchanan, Briggs, Michael J. O'Brien and Mark Collard. 2016. Drivers of Technological Richness in Prehistoric Texas: An Archaeological Test of the Population Size and Environmental Risk Hypotheses. *Archaeological and Anthropological Sciences* 8:625–634.

Carlson, David L. 2017. *Quantitative Methods in Archaeology Using R*. Cambridge University Press, Cambridge.

Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey. 1983. *Graphical Methods for Data Analysis*. CRC Press, Murray Hill, New Jersey.

Codding, Brian F., Heidi Roberts, William Eckerle, Simon C. Brewer, Ishmael D. Medina, Kenneth Blake Vernon, and Jerry D. Spangler. 2024a. Can We Reliably Detect Adaptive Responses of Hunter-Gatherers to Past Climate Change? Examining the Impact of Mid-Holocene Drought on Archaic Settlement in the Basin-Plateau Region of North America. *Quaternary International* 689–690:5–15. https://doi.org/10.1016/j.quaint.2023.06.014.

Codding, Brian F., Jack Meyer, Simon C. Brewer, Robert Kelly and Terry L. Jones. 2024b. Do Counts of Radiocarbon-Dated Archaeological Sites Reflect Human Population Density? A Preliminary Empirical Validation Examining Spatial Variation across Late Holocene California. *Radiocarbon*, in press.

Cook, Sherburne Friend, and Adan Eduardo Treganza. 1950. *The Quantitative Investigation of Indian Mounds*. University of California Press, Berkeley.

Dakki, Mohamed, Geneviève Robin, Marie Suet, Abdeljebbar Qninba, Mohammed A. El Agbani, Asmâa Ouassou, Rhimou El Hamoumi, et al. 2021. Imputation of Incomplete Large-Scale Monitoring Count Data via Penalized Estimation. *Methods in Ecology and Evolution* 12(6):1031–1039. https://doi.org/10.1111/2041-210X.13594.

Davis, Dylan S., Robert J. DiNapoli, and Kristina Douglass. 2020. Integrating Point Process Models, Evolutionary Ecology and Traditional Knowledge Improves Landscape Archaeology: A Case from Southwest Madagascar. *Geosciences* 10(8):287.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, et al. 2013. Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance. *Ecography* 36(1):27–46. https://doi.org/j.1600-0587.2012.07348.x.

Drennan, Robert D. 2009. *Statistics for Archaeologists*. Springer, New York.

Faraway, Julian J. 2016. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, New York.

Fulford, Michael G., and Ian Hodder. 1974. A Regression Analysis of Some Late Romano-British Pottery: A Case Study. *Oxoniensia* 39:26–33.

Hodder, Ian. 1974. Regression Analysis of Some Trade and Marketing Patterns. *World Archaeology* 6(2):172–189.

James, L. Brock, Kaley Joyce, Kate E. Magargal, and Brian F. Codding. 2022. A Stone in the Hand is Worth How Many in the Bush? Applying the Marginal Value Theorem to Understand Optimal Toolstone Transportation, Processing, and Discard Decisions. *Journal of Archaeological Science* 137:105518. https://doi.org/10.1016/j.jas.2021.105518.

Manning, Katie, Adrian Timpson, Sue Colledge, Enrico Crema, and Stephen Shennan. 2015. The Cultural Evolution of Neolithic Europe. *EUROEVOL Dataset*. https://discovery.ucl.ac.uk/id/eprint/1469811/, accessed May 14, 2023.

Marwick, Ben. 2018. R Coding and Modeling. In *The Encyclopedia of Archaeological Sciences*, edited by Sandra L. López Varela, pp. 1–5. John Wiley and Son, Hoboken, New Jersey. https://doi.org/10.1002/9781119188230.saseas0631.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall, London.

O'Hara, Robert, and Johan Kotze. 2010. Do Not Log-Transform Count Data. *Nature Precedings*. https://doi.org/10.1038/npre.2010.4136.1.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

R Core Team. 2024. R: A Language and Environment for Statistical Computing, version 4.3.1. R Foundation for Statistical Computing, Vienna, Austria. Electronic document, https://www.R-project.org/, accessed June 16, 2023.

Shennan, Stephen. 1997. *Quantifying Archaeology*. 2nd ed. University of Iowa Press, Iowa City.

Shott, Michael J. 2015. Glass Is Heavy, Too: Testing the Field-Processing Model at the Modena Obsidian Quarry, Lincoln County, Southeastern Nevada. *American Antiquity* 80(3):548–570.

Shott, Michael J. 2018. *Pottery Ethnoarchaeology in the Michoacán Sierra*. University of Utah Press, Salt Lake City.

Shott, Michael J. 2022. Inferring Use-Life Mean and Distribution: A Pottery Ethnoarchaeological Case Study from Michoacán. *American Antiquity* 87(4):794–815.

Simpson, Gavin L. 2018. Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution* 6:149. https://doi.org/10.3389/fevo.2018.00149.

St-Pierre, Anne P., Violaine Shikon, and David C. Schneider. 2018. Count Data in Biology: Data Transformation or Model Reformation? *Ecology and Evolution* 8(6):3077–3085. https://doi.org/10.1002/ece3.3807.

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S*. Springer, New York.

Vernon, Kenneth B., Peter M. Yaworsky, Weston C. McCool, Jerry D. Spangler, Simon C. Brewer, and Brian F. Codding. 2024. The Fremont Frontier: Living at the Margins of Maize Farming. *American Antiquity*, in press. https://doi.org/10.1017/aaq.2024.22.

Vernon, Kenneth B., Peter M. Yaworsky, Jerry Spangler, Simon C. Brewer, and Brian F. Codding. 2022. Decomposing Habitat Suitability across the Forager to Farmer Transition. *Environmental Archaeology* 27(4):420–433.

Warton, David I., Mitchell Lyons, Jakub Stoklosa, and Anthony R. Ives. 2016. Three Points to Consider When Choosing a LM or GLM Test for Count Data. *Methods in Ecology and Evolution* 7(8):882–890. https://doi.org/10.1111/2041-210X.12552.

Waterman, Thomas Talbot. 1920. *Yurok Geography*. University of California Press, Berkeley.

Wood, Simon. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. CRC Press, London.

Yaworsky, Peter M., Kenneth Blake Vernon, Jerry D. Spangler, Simon C. Brewer, and Brian F. Codding. 2020. Advancing Predictive Modeling in Archaeology: An Evaluation of Regression and Machine Learning Methods on the Grand Staircase-Escalante National Monument. *PLoS ONE* 15(10):e023942. https://doi.org/10.1371/journal.pone.0239424.

Zuur, Alain F., and Elena N. Leno. 2016. A Protocol for Conducting and Presenting Results of Regression-Type Analyses. *Methods in Ecology and Evolution* 7(6):636–645.

Zuur, Alain F., Elena N. Leno, and Chris S. Elphick. 2010. A Protocol for Data Exploration to Avoid Common Statistical Problems. *Methods in Ecology and Evolution* 1(1):3–14.

Zuur, Alain F., Elena N. Leno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.

AUTHOR INFORMATION

Brian F. Codding ■ Department of Anthropology, University of Utah, Salt Lake City, UT, USA, and Archaeological Center, University of Utah, Salt Lake City, UT, USA (brian.codding@anthro.utah.edu, corresponding author)

Simon C. Brewer ■ Department of Geography, University of Utah, Salt Lake City, UT, USA, and Archaeological Center, University of Utah, Salt Lake City, UT, USA

https://doi.org/10.1017/aap.2024.7 Published online by Cambridge University Press