Political Science Research and Methods (2023), page 1 of 9
doi:10.1017/psrm.2023.29

RESEARCH NOTE

How to improve the substantive interpretation of regression results when the dependent variable is logged

Oliver Rittmann[1]*, Marcel Neunhoeffer[2,3] and Thomas Gschwend[1]

[1]School of Social Sciences, University of Mannheim, Mannheim, Germany, [2]Rafik B. Hariri Institute for Computing and Computational Science & Engineering, Boston University, Boston, USA and [3]Department of Statistics, LMU Munich, München, Germany
*Corresponding author. Email: orittman@mail.uni-mannheim.de

(Received 7 July 2022; revised 3 March 2023; accepted 5 June 2023)

Abstract

Regression models with log-transformed dependent variables are widely used by social scientists to investigate nonlinear relationships between variables. Unfortunately, this transformation complicates the substantive interpretation of estimation results and often leads to incomplete and sometimes even misleading interpretations. We focus on one valuable but underused method, the presentation of quantities of interest such as expected values or first differences on the original scale of the dependent variable. The procedure to derive these quantities differs in seemingly minor but critical aspects from the well-known procedure based on standard linear models. To improve empirical practice, we explain the underlying problem and develop guidelines that help researchers to derive meaningful interpretations from regression results of models with log-transformed dependent variables.

Keywords: Linear regression; logged dependent variable; quantities of interest; simulation

Regression models with log-transformed dependent variables are widely used by social scientists to investigate nonlinear relationships between variables. Unfortunately, this transformation complicates the substantive interpretation of respective estimation results. In an effort to improve empirical practice, we clarify one popular strategy for the substantive interpretation of such regression results—the presentation of quantities of interest such as predicted values, expected values, or first differences on the original scale of the dependent variable (King *et al.*, 2000). We show that calculating such quantities together with their associated uncertainty is different from well-known procedures that work in the case of linear regression models without log-transformed dependent variables. Ignoring this difference can lead to erroneous communication of regression results when the dependent variable is log-transformed.

The key point of confusion is this: a regression with a logged dependent variable estimates $E[\ln(y|X)]$. For a substantive interpretation we want to calculate quantities of interest and their associated uncertainty on the original scale of the dependent variable rather than the logged scale. However, scholars cannot simply exponentiate expected values, standard errors, or upper and lower bounds of the estimated confidence intervals on the logged scale in order to transform them to the original scale. While $\ln(y|X)$ is normally distributed, its transformation $y|X$ back to the original scale is skewed. The consequence is that $e^{E[\ln(y|X)]} \neq ; e^{\ln(E[y|X])} = E[y|X]$. This is well known among methodologists (e.g., Manning, 1998), but often neglected by substantive scholars. To derive the desired quantities on the original scale together with the associated uncertainty, scholars need to carefully apply appropriate transformation formulas and simulate their respective confidence intervals correctly.

Popular methods textbooks also acknowledge this "retransformation problem" (e.g., Cameron and Trivedi, 2022: Section 4.2.3; see also Cameron and Trivedi, 2005: Section 20.5.2). In this research note, we add two important aspects. Cameron and Trivedi (2022) declare that the prediction $e^{E[ln(y|X)]}$ is a "very poor" and incorrect prediction for $E[y|X]$. We show that $e^{E[ln(y|X)]}$ is indeed not a valid prediction for the mean on the original scale. However, it is still an interesting quantity as it represents the conditional median of the log-normal distribution of $y|X$. Furthermore, the solutions described in Cameron and Trivedi (2022) focus on the point estimate of the conditional mean, but offer little guidance on the uncertainty that comes with those estimates. Here, we integrate solutions to the retransformation problem of point estimates with the simulation algorithm by King *et al.* (2000). In doing so, our solution also provides accurate confidence intervals for $E[y|X]$ and other quantities of interest on the original scale of the dependent variable.

Even though the presentation of meaningful quantities of interest became best practice for the interpretation of a wide range of statistical models, our review of current practice shows that substantive scholars make little use of this approach when interpreting results from regression models with logged dependent variables. We base this conclusion on a content analysis of all research articles published in the *American Political Science Review* and *American Journal of Political Science* between 2015 and 2020. In total, we identify 39 articles in which scholars report at least one statistical model with a log-transformed dependent variable.

We identify three main styles of interpretation.[1] First, an "old school" strategy that uses the mere direction (positive/negative) and statistical significance of regression coefficients for interpretation (in 6 out of 39 articles). Those articles provide no substantive interpretation of the respective results and their uncertainty. Second, the most popular practice is the interpretation of regression coefficients as "percent increase" of the dependent variable (used in 31 out of 39 articles). While this practice is not incorrect, we see an important shortcoming. Any concrete interpretation of a "percent increase" does not provide us with an adequate sense of the effect's absolute magnitude. Furthermore, scholars rarely present the uncertainty associated with a "percent increase." Finally, there is an interpretation of regression results through the presentation of quantities of interest such as predicted values, expected values, or first differences (in 13 out of 39 articles). This is the one we recommend if done correctly and effectively. Surprisingly, few authors provide uncertainty assessments for their quantities of interest (only 3 of 13 quantities are presented with uncertainty estimates). Because there is no reason why scholars should not be interested in communicating uncertainty on the original scale of the dependent variable, we interpret this result as stemming from a lack of guidance in how to correctly derive quantities of interest together with appropriate confidence intervals on the original scale when the dependent variable is log-transformed.

In this research note, we provide guidance on how substantive scholars can improve their interpretation of regression results when the dependent variable is logged. We show how to calculate quantities of interest on the original scale even when the dependent variable is log-transformed, and how to derive respective confidence intervals using simulations. Furthermore, we highlight how the nonlinear nature of the log-transformation has important consequences for the calculation of first differences and how the presentation of first differences is especially useful for the interpretation of estimation results from models that include interaction terms. We illustrate the utility of our approach with a reanalysis of a recent study on executive appointment processes in the USA.

1. Calculating quantities of interest when the dependent variable is logged

To calculate quantities of interest on the original scale for models with logged dependent variables, the workflow is to log-transform the dependent variable y, estimate the regression model

[1]The categories are not mutually exclusive. A list of all articles can be found in the Supporting information (SI.1).

Table 1. Transformation formulas for point estimates of common quantities of interest, and their approximated distributions to construct correct confidence intervals by using $\alpha/2$ and $1-\alpha/2$ percentiles

Quantity of interest		Point estimate	Simulated confidence interval: use $Q_{\frac{\alpha}{2},1-\frac{\alpha}{2}}(\cdot)$
Variance	σ^2	$\hat{\sigma}^2$	$\tilde{\sigma}^2 \sim \text{Inv}-\Gamma(\frac{N-k}{2},\ \frac{\hat{\sigma}^2(N-k)}{2})$
Regression coefficients	β	$\hat{\beta}$	$\tilde{\beta} \sim MVN(\hat{\beta},\ \tilde{\sigma}^2(X'X)^{-1})$
Expected value	$E(Y_c)$	$e^{X_c\hat{\beta}+\frac{1}{2}\hat{\sigma}^2}$	$e^{X_c\tilde{\beta}+\frac{1}{2}\tilde{\sigma}^2}$
Median value	$\text{Med}(Y_c)$	$e^{X_c\hat{\beta}}$	$e^{X_c\tilde{\beta}}$
First difference	$E(Y_{c_1}) - E(Y_{c_2})$	$e^{X_{c_1}\hat{\beta}+\frac{1}{2}\hat{\sigma}^2} - e^{X_{c_2}\hat{\beta}+\frac{1}{2}\hat{\sigma}^2}$	$e^{X_{c_1}\tilde{\beta}+\frac{1}{2}\tilde{\sigma}^2} - e^{X_{c_2}\tilde{\beta}+\frac{1}{2}\tilde{\sigma}^2}$
First difference	$\text{Med}(Y_{c_1}) - \text{Med}(Y_{c_2})$	$e^{X_{c_1}\hat{\beta}} - e^{X_{c_2}\hat{\beta}}$	$e^{X_{c_1}\tilde{\beta}} - e^{X_{c_2}\tilde{\beta}}$

with $\ln(y)$ as dependent variable, and then use estimation results to calculate quantities of interest, such as $E[\ln(y)]$, although substantively rarely meaningful. To get substantively meaningful quantities, one needs to transform those quantities back to the variable's original scale to get $E[y]$. While the transformation in the first step is fairly simple—we take the natural log of each value of y—the back transformation requires careful thinking.

The back transformation is not straightforward because it maps $\ln(y)$ back to y, which is skewed log-normally distributed conditional on the model. For $y = e^{\ln(y)} = e^{X\beta+\epsilon}$ one can show that $E[y] = E[e^{\ln(y)}] = e^{E[\ln(y)]} \cdot E[e^{\epsilon}] = e^{E[\ln(y)]+(1/2)\hat{\sigma}^2} > e^{E[\ln(y)]}$ (e.g., Manning, 1998). Therefore, we cannot simply exponentiate $E[\ln(y)]$ to obtain $E[y]$. Table 1 provides an overview of the correct transformation formulas: if we are interested in $E[y]$, we need to transform the estimates on the log-scale with $E[y] = e^{E[\ln(y)]+(1/2)\hat{\sigma}^2}$. But $e^{E[\ln(y)]}$ is an interesting quantity as well, as this represents the median of the resulting log-normal distribution. Both the mean and the median of y can be interesting and reasonable quantities to present, but researchers must be aware of the difference and should not confuse one with the other when interpreting results. For unimodal, continuous skewed distributions, such as the log-normal distributed y, the median is often considered to be a more typical value than the mean (von Hippel, 2005).

To develop a deeper intuition and to illustrate the consequences of choosing the mean or the median as a typical value in the context of linear models, we walk through a motivating example introduced by Rainey (2017). Consider the following data generating process (DGP):

$$\ln(\text{Income}) = \beta_{\text{cons}} + \beta_{\text{edu}}\ \text{Education} + \epsilon, \text{ and } \epsilon \sim N(0, \sigma^2) \tag{1}$$

The true values of the coefficients are given by $\beta_{\text{cons}} = 2.5$, $\beta_{\text{edu}} = 0.1$, and $\sigma^2 = 1$. The challenge with this DGP is that the dependent variable is log-transformed. Scholars, however, are usually interested in interpreting the results on the original scale of the dependent variable because we are interested in the results in dollars rather than ln(dollar). Suppose that we are interested in a typical income for a person with 20 years of education given our model. How can we calculate such a typical value of income in dollars, even if the dependent variable is income in ln(dollar)?

First, we need to choose whether we want to present the median or mean as our typical value. Both can be interesting. To calculate a point estimate of the median income conditional on our scenario of 20 years of education, we get $\text{Med}(Y_c) = e^{2.5+0.1\times20} = e^{4.5} \approx 90.06$. If we are interested in a point estimate of the mean income conditional on 20 years of education, we get $E(Y_c) = e^{2.5+0.1\times20+1/2\times1} = e^{5} \approx 148.41$. This shows that the mean and median are two very distinct quantities.[2]

[2]The sampling distributions of these estimators for $\widehat{\text{Med}(Y_c)} = e^{X_c\hat{\beta}}$ and $\widehat{E(Y_c)} = e^{X_c\hat{\beta}+(1/2)\hat{\sigma}^2}$ are right skewed as well and thus the estimators are biased. Rainey (2017) describes this bias as transformation induced bias.

2. Simulating confidence intervals when the dependent variable is logged

Every estimation entails uncertainty. Transparent communication of this uncertainty is fundamental to scientific practice. In this section, we show how the simulation approach by King *et al.* (2000) can be used to get confidence intervals for both, the median and mean of y on the original scale. Figure 1 provides the algorithm to simulate confidence intervals for the mean when the dependent variable is not transformed (left column), the median when the dependent variable is log-transformed (center column), and the mean when the dependent variable is log-transformed (right column). Consider the workflow for models with untransformed dependent variables first.

The procedure consists of four steps. In *step 1* we approximate the distributions of the estimated coefficients to account for *estimation uncertainty*.[3] We start by drawing $\tilde{\sigma}^2$ from an inverse-gamma distribution, $\text{Inv}{-}\Gamma((N-k)/2,\ (\hat{\sigma}^2(N-k))/2)$, where $\hat{\sigma}^2$ is the estimated variance, N is the number of observations, and k is the number of coefficients ($\hat{\beta}$). Next, we approximate the distribution of $\hat{\beta}$ by drawing simulations $\tilde{\beta}$ from the multivariate normal distribution

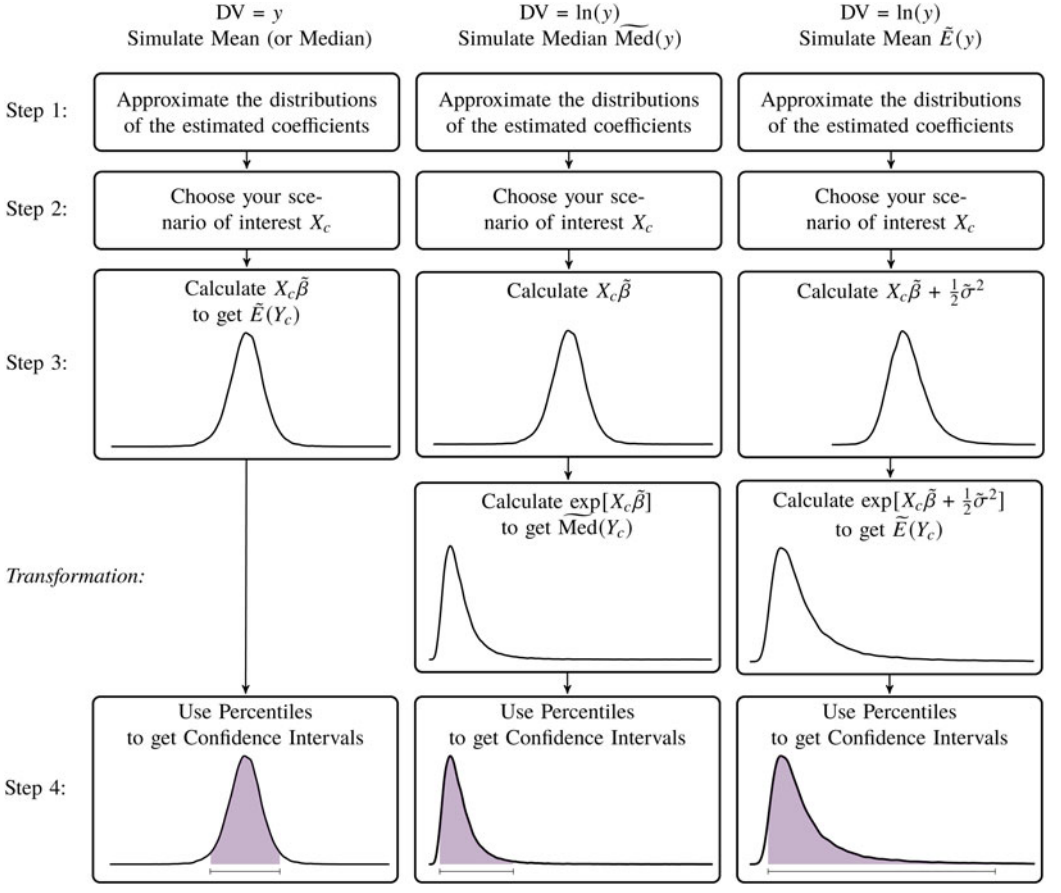

Figure 1. Algorithm to simulate confidence intervals (King *et al.*, 2000) when the dependent variable is untransformed (left column), for the median when the dependent variable is logged (center column), and for the mean when the dependent variable is logged (right column).

[3]The simulation approach follows an informal Bayesian logic. When we write that we simulate the distribution of a quantity, then this means that we draw from an informal posterior distribution of that quantity.

$MVN(\hat{\beta},\ \tilde{\sigma}^2(X'X)^{-1})$. In *step 2* we choose our scenario of interest, i.e., we specify covariate values X_c that are held constant during simulation. In *step 3* we calculate $X_c\tilde{\beta}$, the linear combination of the simulations $\tilde{\beta}$ and the chosen values of the covariates (X_c) that define the scenario of interest. This results in a simulated distribution $\tilde{E}(Y_c)$ of expected values of Y conditional on the specified scenario X_c. In *step 4* we get a $(1-\alpha)\times 100\%$-confidence interval by summarizing the distribution with the $\alpha/2$ and $1-\alpha/2$ percentiles.

Now consider the center column in Figure 1 where we outline the simulation procedure for a confidence interval for the conditional median of a model with a logged dependent variable. Using the example in Equation 1, suppose that we are interested in the median income in dollars of a person with 20 years of education. Steps 1–3 do not differ from the standard procedure: we draw simulations of the model coefficients ($\tilde{\beta}$ and $\tilde{\sigma}^2$), we set Education = 20, and compute the linear combination of the simulated coefficients and the scenario of interest. This yields $\tilde{E}(\ln(Y_c))$, a simulated distribution of ln (Income) conditional on 20 years of education. To get confidence intervals for the median in dollars, we exponentiate the distribution of $\tilde{E}(\ln(Y_c))$ as shown in the transformation step in Figure 1. Step 4 then is the same as before. Note that the resulting distribution is skewed and the confidence intervals will not be symmetric around the point estimate from the previous section.[4]

In the right column of Figure 1, we show how to get confidence intervals for the mean, e.g., the mean income in dollars of a person with 20 years of education. The procedure mostly remains the same with the exception of step 3 where we need to add $(1/2)\tilde{\sigma}^2$ to $X_c\tilde{\beta}$. The derivation of confidence intervals for first differences follows equivalently (see Table 1 for an overview).[5]

3. First-differences and interaction terms with log-transformed dependent variables

We point out two additional issues that arise when presenting quantities of interest on the original scale from regression models with log-transformed dependent variables. First, the magnitude of a first difference based on regression models with log-transformed dependent variables depends on *all* covariates in the scenario, even those that are held constant. This is different for regular linear regression models where the point estimate of the first difference does not depend on the values of covariates that are held constant across scenarios.

Second, if an interaction term is present in the regression model, e.g., $\ln(y) = \beta_0 + \beta_1 D + \beta_2 X + \beta_3(D\times X) + \epsilon$, then the first difference $E(y|D=1, X) - E(y|D=0, X)$ can *increase* at higher levels of X, even if β_3 is *negative* (and vice versa). This is different for regular linear models where the first difference is given by the marginal effect of D, i.e., $\beta_1 + \beta_3 X$, and changes as a linear function of X at a rate of β_3. Figure 2 demonstrates this notion.

For applied work, these peculiarities show the benefit of presenting quantities of interest of regression models with log-transformed dependent variables on the original scale. At the same time they alert researchers to choose and justify all values in their scenarios carefully. An insensitive selection of any of a model's covariate values may artificially inflate or deflate the size of a first difference.

4. Application—a reanalysis of Hollibaugh and Rothenberg (2018)

To demonstrate the utility of our approach, we reanalyze the results of a recently published study by Hollibaugh and Rothenberg (2018). The study investigates factors that influence executive appointment processes in the US context. The authors, among other things, study the relation between agency dependence and appointee ideology. We are specifically interested in one of

[4]This is also the reason why taking the average of the simulations is not a good strategy to get a point estimate. As the distribution is right skewed, continuous, and unimodal, the mean of this distribution will be biased upward.

[5]We show that the confidence intervals calculated with this procedure have the correct coverage through a Monte Carlo simulation we present in the Supporting information (SI.2).

https://doi.org/10.1017/psrm.2023.29 Published online by Cambridge University Press

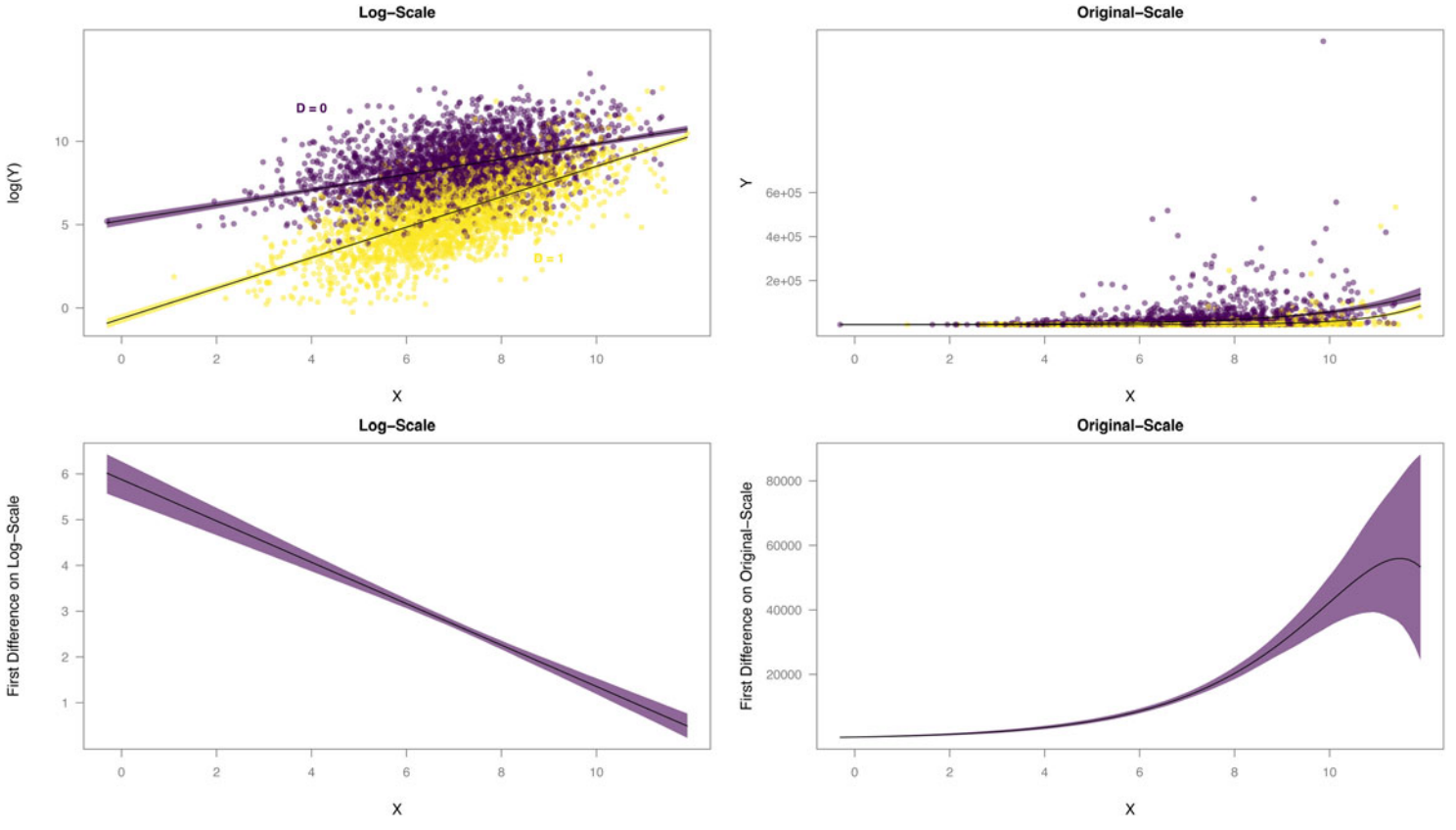

Figure 2. Quantities of interest with an interaction effect. The DGP follows $\ln(Y) = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + \epsilon$ with $\beta_0 = 5$, $\beta_1 = -5.5$, $\beta_2 = 0.5$, $\beta_3 = 0.4$, and $\epsilon \sim N(0,\ 1.5^2)$. As X increases, the first difference on the log-scale of Y *decreases*, but it *increases* on the original scale of Y.

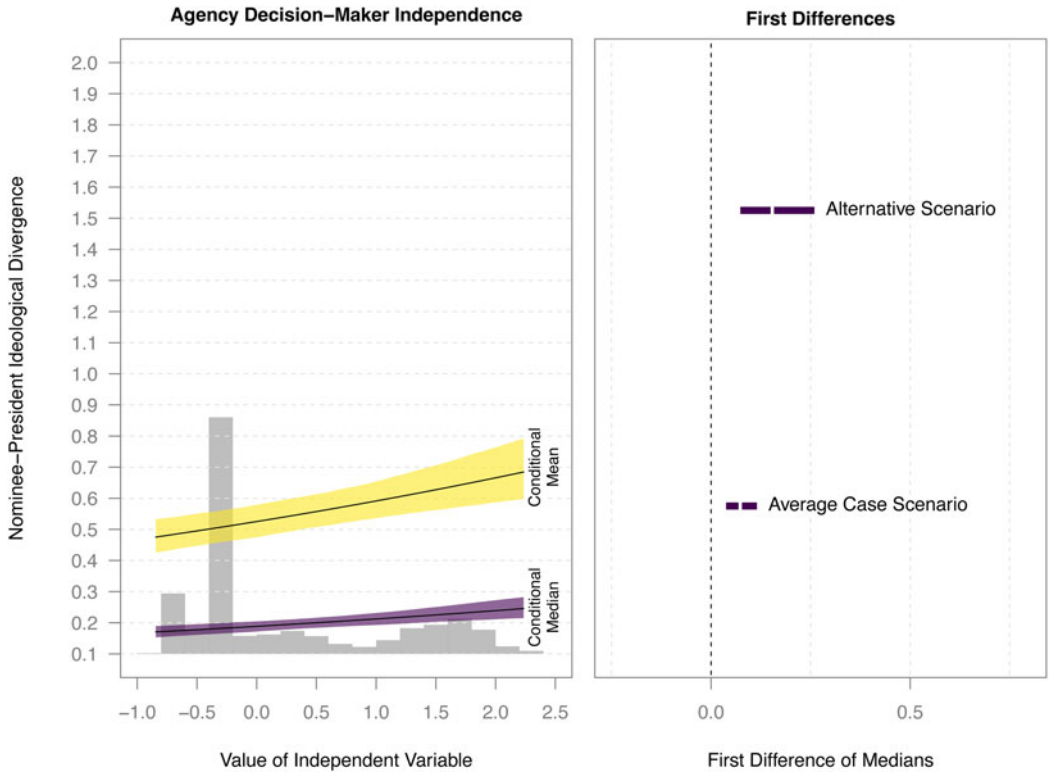

Figure 3. Conditional mean and median values of nominee-president ideological divergence and first difference between minimum and maximum values of the independent variable, conditional on two different sets of covariate values.

their hypotheses: the higher the independence of the decision-maker in the targeted agency, the higher the ideological divergence between the president and a nominee.[6]

To test this hypothesis, Hollibaugh and Rothenberg (2018) estimate linear models. Their dependent variable is the natural log of the ideological divergence between a nominee and the president (*Nominee-President Divergence*). The key independent variable is *Agency Decision Maker Independence*. In support of the hypothesis, Hollibaugh and Rothenberg (2018) find that *Agency Decision Maker Independence* is positively associated with the divergence between the president and the nominee.

To facilitate the interpretation of this effect, Hollibaugh and Rothenberg (2018) report "expected values" of ideological divergence between the president and the nominee from low to high agency decision-maker independence. All binary variables are set to zero, and following an average case approach, all covariate values of continuous variables to their means. We replicate the analysis following our guidelines and present both conditional mean and median values of the dependent variable in Figure 3.[7] Our reanalysis reveals two things: first, researchers are not always aware of the difference between conditional mean and conditional median values of y. What Hollibaugh and Rothenberg (2018) present as expected values are actually conditional median values. Second, this difference is not trivial. Figure 3 demonstrates that the conditional mean

[6]This relates to hypotheses 6 in the original article. We refer interested readers to the original article for more details on the theoretical arguments behind this hypothesis.

[7]Hollibaugh and Rothenberg (2018) z-transform the independent variables and transform simulated values to the empirical percentile scale of the dependent variable.

and the conditional median of y are two very distinct quantities. The conditional mean values are considerably larger than conditional median values.

Next, we illustrate how the selection of covariate values matters for first differences. The right panel of Figure 3 presents two first differences. Both estimates show a first difference of the median between the minimum and maximum values of agency decision-maker independence, but we alternated the values of the covariates that are held constant. One first difference is based on an average case scenario (as in the left panel of Figure 3), for the other scenario we set all these covariates to either their minimum or maximum. This would not affect the magnitude of the first difference in regular linear models, but it clearly affects the magnitude in this case where the dependent variable is log-transformed. The first difference amounts to 0.075 for the average case setting, but it roughly doubles to 0.156 if we fix the control variables at more extreme values.

5. Conclusion

We have shown how to apply appropriate transformation formulas to estimated coefficients of linear regression models with logged dependent variables in order derive various quantities of interest on the original scale, and how to derive respective confidence intervals using simulations. We conclude with a set of four recommendations that researchers should keep in mind when improving the interpretation of such models.

First, it makes a difference whether conditional mean or median values are presented. Unless there is a special theoretical interest in only one of both quantities, our advice is to present both the conditional mean and the conditional median. Second, point estimates of conditional mean and median values should be calculated directly based on the point estimates of the regression model using appropriate transformation formulas (see Table 1). The simulation method in combination with the same formulas allows to derive respective confidence intervals. Third, even values that are held constant across simulations, typically values of control variables, are influential for quantities of interest on the original scale. These values have to be chosen and communicated transparently. Typical strategies are to set those variables to their means, their medians, or to observed values (Hanmer and Kalkan, 2013). Fourth, if the model includes one or more interaction terms, researchers should refrain from interpreting marginal effects on the logged scale. Larger marginal effects on the logged scale do not necessarily reflect larger marginal effects on the original scale. In fact, the opposite may be true as we have shown. To interpret respective results, our advice is to always calculate first differences on the original scale of the dependent variable.

6. Software

A software implementation of the proposed method is available as an open-source R package, simloglm, at https://github.com/mneunhoe/simloglm. We provide a code example for using the package, as well as calculating all mentioned quantities of interest and confidence intervals by hand in R in the Supporting information (SI.3).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2023.29 . To obtain replication material for this article, https://doi.org/10.7910/DVN/KZWKT6

Acknowledgments. We thank Marie-Lou Sohnius for excellent research assistance. We are grateful to Daniel Stegmueller, Lukas Stoetzer, Patrick Kraft, Andreas Murr, Marie-Lou Sohnius, and the anonymous reviewers' feedback on earlier versions of this manuscript. Moreover, we thank the students of our class "Advanced Quantitative Methods" at the University of Mannheim for inspiring us to write this paper. An earlier version of the manuscripts was presented at the first Meeting of the Society for Political Methodology in Europe 2021.

Competing interest. The authors declare none.

References

Cameron AC and Trivedi PK (2005) *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cameron A and Trivedi P (2022) *Microeconometrics Using Stata*. College Station, TX: Stata Press.

Hanmer MJ and Kalkan KO (2013) Behind the curve: clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science* **57**, 263–277.

Hollibaugh GE and Rothenberg LS (2018) The who, when, and where of executive nominations: integrating agency independence and appointee ideology. *American Journal of Political Science* **62**, 296–311.

King G, Tomz M and Wittenberg J (2000) Making the most of statistical analyses: improving interpretation and presentation. *American Journal of Political Science* **44**, 347–361.

Manning WG (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* **17**, 283–295.

Rainey C (2017) Transformation-induced bias: unbiased coefficients do not imply unbiased quantities of interest. *Political Analysis* **25**, 402–409.

von Hippel PT (2005) Mean, median, and skew: correcting a textbook rule. *Journal of Statistics Education* **13**, 965–971.

Cite this article: Rittmann O, Neunhoeffer M, Gschwend T (2023). How to improve the substantive interpretation of regression results when the dependent variable is logged. *Political Science Research and Methods* 1–9. https://doi.org/10.1017/psrm.2023.29