PRICELESS OR PRICEY? "ARBITRARY" CHOICES IN LOG-LINEAR MODELS AND THE "ARBITRARY" COST OF HAVING CHILDREN

Briggs Depew[1]

Abstract

A popular fix when dealing with zeros in the dependent variable, y, is to add a scalar value, a, within the log transformation, i.e. $\log(y + a)$. However, the choice of the scalar value is often seemingly arbitrary. Using data from the Current Population Survey, I step-by-step walk through an empirical investigation of how an additional child in the household affects childcare cost, and I show that the choice of the arbitrary scalar value significantly affects the estimates of a log-linear regression model. For those "special couples" who are mining through data from the Current Population Survey to inform them on life decisions, they can estimate a model to justify any decision by their choice of a. We demonstrate that the best practice may be to forgo the log-linear regression model when dealing with zeros and turn to a Poisson regression.

Keywords: regression analysis, zeros, missing data, censored data, Poisson, Childcare

JEL Classification: A22, J13, C18

Introduction

Undergraduate courses in regression analysis, whether in the fields of statistics, economics, or business analytics, among others, are designed to equip students with the necessary tools to tackle real-world data. However, it is often the case that students encounter issues that they are unprepared to handle. Online resources, like Stack Exchange or ChatGPT, are often helpful, but these resources can lead a novice practitioner down an endless path of misunderstanding and confusion. A common problem in applied work is encountering a dependent variable that includes zeros, but the practitioner would like to estimate a log-linear regression. The purpose of this article is to demonstrate how the choice of the "arbitrary" scalar in the popular fix of $\log(y + a)$ can lead to results that are not arbitrary. As students are gearing towards the completion of college, some may be wondering what is next in life and some may even be pondering marriage and children. By walking through an example that regresses log childcare expenses on the number of young children in the household, we demonstrate that the "cost" of an additional child will greatly depend on the choice of the arbitrary parameter used to deal with zeros. In this setting, one can estimate a model to justify any outcome by the choice of the value of a. This provides a clear and intuitive example of how results can greatly be affected by choices that may seem arbitrary.

A quick survey of degree requirements for majors in economics, statistics, and business analytics across institutions suggests that students often take one or two courses in linear regression models but rarely, at least at the undergraduate level, do they take courses covering more advanced non-linear regression models such as Probit and Logit, Poisson, Tobit, and other methods. As a

[1] Associate Professor, Department of Economics and Finance, Utah State University Old Main Hill, Logan, UT 84322; IZA, Germany

result, it is challenging for novice practitioners to want to deviate from the known linear-regression framework[2].

Log transformations are a common practice in empirical work. Justification for choosing log-linear models often include: 1) ease of interpretation (the estimated parameter can be interpreted as an elasticity or semi-elasticity), 2) logs can linearize a non-linear model such as a Cobb-Douglas production function, 3) when skewed data is logged it becomes more normally distributed, and 4) homoskedasticity is less likely to be violated when variables are logged. When novice practitioners encounter the issue of having zeros in the dependent variable, they may turn to ChatGPT for advice.

When we prompted ChatGPT (OpenAI, 2024) with the following, "How to estimate a log-linear regression when there are zeros in the dependent variable?", we received a reply that provided five solutions. Four of the solutions suggested more complicated non-linear models (Tobit Model, Zero-inflated Model, Poisson or Negative Binomial Model, and a two-step model using a Logistic model in part with a log-linear regression). The first suggestion by ChatGPT was to add a small constant to the log transformation: "One common approach is to add a small constant (usually denoted as a "pseudo-count" or a "small shift") to the dependent variable before taking the logarithm. This ensures that no value in the dependent variable is zero or negative, avoiding the problem of undefined logarithms." Adding a scalar to deal with the log of zero dates back at least to Williams (1937) and despite the common practice of doing so, it has been shown that the transformation will bias the Ordinary Least Squares (OLS) estimates (Flowerdew and Aitkin, 1982; King, 1988).

Recently, academic researchers have turned their attention to dealing with zeros in logged data. Bellego, Benantia and Pape (2022) reviewed all articles published in the *American Economic Review* between 2016 and 2020 to survey the extent of academic researchers deal with the log of zero. They found that 40% of empirical papers used a log-specification and 36% of these articles faced the problem of the log of zero. Bellego, Benantia and Pape (2022) found that it was most common for authors to keep the zero observations but to also add a positive arbitrary value to the dependent variable (48% of articles). In 35% of the articles, they found that authors used a Poisson-type estimator, and in 15% of the articles the authors used the inverse hyperbolic sine transformation. Finally, 31% of articles discarded the observations with zeros. Note, the choice of modeling was not mutually exclusive since in 20% of articles, the authors compared more than one method to assess the robustness of the modeling choice. The article by Chen and Roth (2023) studies the popular fix of $\log(y+1)$, and their main finding shows that the average treatment effect for such transformations should not be interpreted as percentages, since they depend arbitrarily on the units of the outcome when there is an extensive margin.

This article does not contribute to understanding the bias or the proper fix. Instead, we provide data and a step-by-step example of how the choice of the scaling parameter, a, can affect the results in a setting that is easily accessible to students. We also digress into the alternative estimation practice of using a Poisson regression model. In the next section, we provide a brief background on the log function of the log-linear regression equation with the addition of the scalar value a. We demonstrate that adding the scalar value will bias the results. The following sections present data from the Current Population Survey (CPS) 2010-2018 (Ruggles et al., 2024) and an empirical investigation into household childcare expenditures using the log-linear model. We

[2] The growth of data analytics there has been more emphasis placed on applying data tools, rather than a depth of understanding and the causal pathways of black box models (Zhao et al., 2021).

conclude by introducing the Poisson regression model, and we discuss the benefits and interpretation of the model parameters.

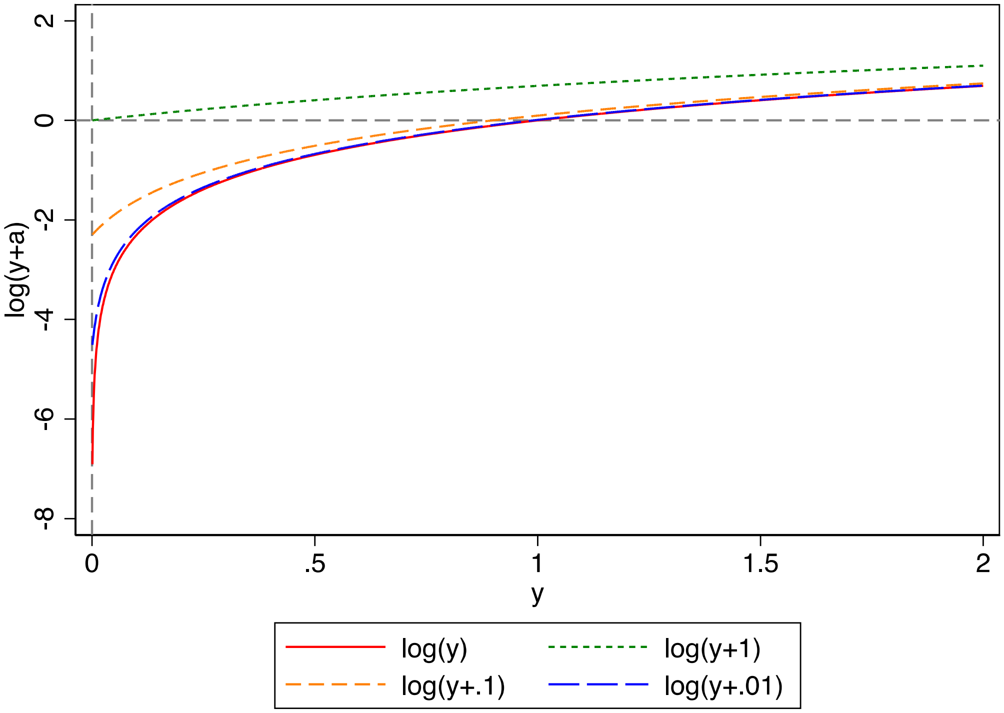
Log Function and Zeros

The log function is the inverse function to the exponential function. If one considers the equation $b^y = x$, then the log of x to base b is equal to y. In other words, what power does one need to raise b to obtain the value x. The log function compresses large values and expands small values. As the input gets larger, the increase in the log value becomes smaller. For example, the log of 10 is 2.30, the log of 100 is 4.61, and the log of 1,000 is 6.91, even though the numbers (10, 100, 1,000) are growing by factors of 10, the corresponding log values only grow by a constant amount (2.30). This behavior compresses large values into a smaller range, and it is why logarithms are often used to compare data that spans wide arrays, like income or population. For inputs close to 1, the log function becomes very sensitive to small changes. For example, the difference between the log of 1.01 and the log of 1.00 is larger than the difference between the log of 101 and 100. In other words, logarithms expand small values, and in some sense magnify the relative differences in smaller numbers.

The problem with taking the log of zero is that it is undefined. The solid line of Figure 1 displays the relationship between y and log (y).[3] A vertical asymptote exists at $y = 0$. As y approaches zero from the right, $\log(y)$ approaches negative infinity. In Figure 1, we present three transformations: log $(y + 1)$, log $(y + .1)$, and log $(y + .01)$. Not surprisingly, the smaller the value of the parameter of a, the closer the transformation matches log (y). However, in the applied setting we show that it is not always ideal to let the parameter a equal a small positive value.

[3] Throughout the paper, we use the natural log function in our empirical demonstrations.

Figure 1: Log Function



Teaching Points of Emphasis #1:

- *The log function compresses large values and expands small values.*
- *The log of zero is undefined.*
- *The closer that y is to zero (limit as y approaches zero from the right), the smaller the* $log(y)$ *becomes to the point where* $\lim_{y \to 0^+} log\ (y) = -\infty$.
- *The smaller the value of the parameter of a, the closer the transformation of* $log(y + a)$ *matches* $log\ (y)$.
    - *As we will later show, this does not justify using the smallest parameter possible.*

Data

I now turn our attention to the data from the 2010-2018 Annual Social and Economic Supplement (ASEC) of the CPS[4]. The CPS is conducted annually by the U.S. Census Bureau and the Bureau of Labor Statistics and is designed to provide comprehensive data on income, poverty, health insurance coverage, and a variety of demographic and economic characteristics of households in the United States. The ASEC is typically conducted in March as a follow-up to the regular monthly CPS. We limit the sample to years 2010-2018 for households with young children. Particularly, an observation is a household with at least one child in the home and with the oldest child being five years of age or younger. Column 1 of Table 1 presents the summary statistics for the data used in the analysis. Average childcare expenses are $2,535 with a standard deviation of $6,791. Childcare expenses are calculated at the Supplemental Poverty Measure (SPM) family unit level. The SPM family unit includes people who live in the same housing unit and are related by

[4] In 2019, the Poverty Supplement of the ASEC implemented survey changes that affected income and relationship variables, making poverty measurements before and after 2019 incomparable, including childcare expenses.

birth, marriage, or adoption. It also includes cohabiting couples and their children and foster children. The average number of children under 6 in the household is 1.45. The head of the household has an average of 14.16 years of education, 72% of the households are married, the average family income is $81,000, and the average age of the youngest child in the home is 1.84. In total, we have 49,278 observations.

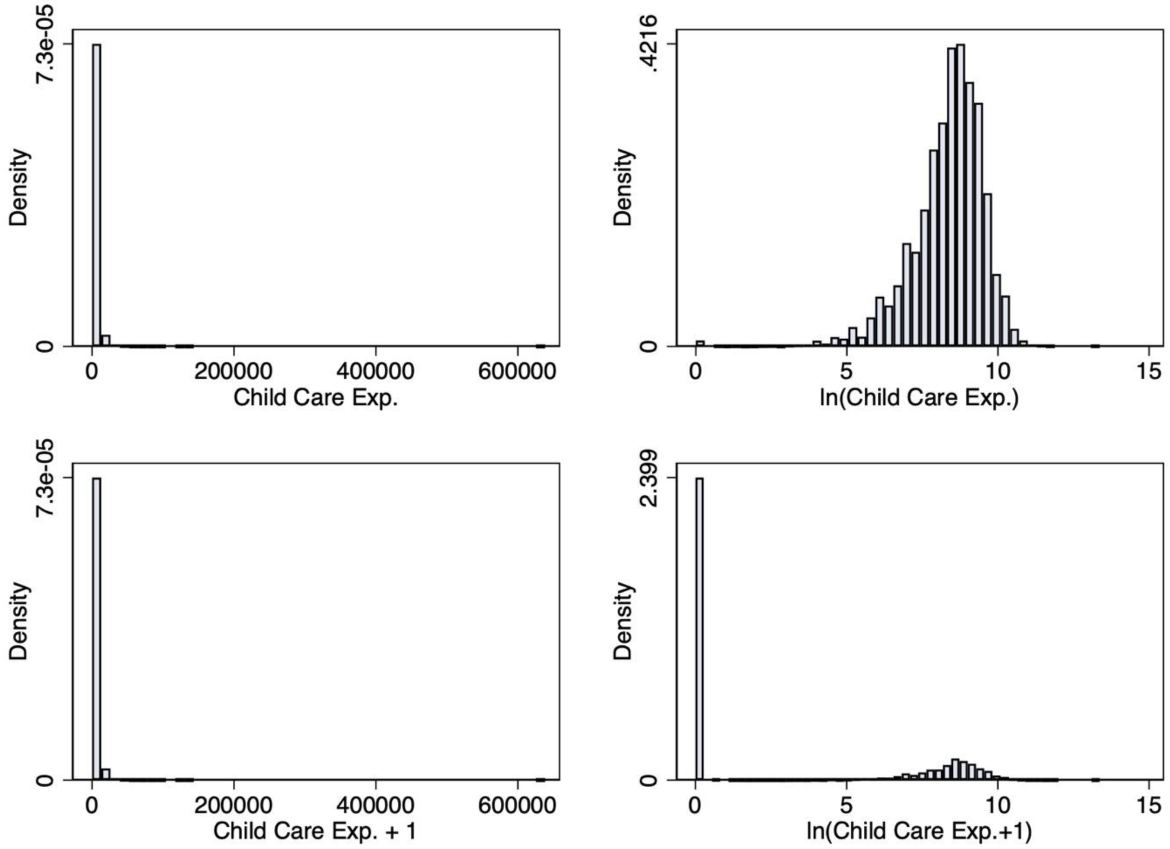
Table 1: Summary Statistics

		Childcare Exp.		Difference
	All	$>0	$0	In Means
	(1)	(2)	(3)	(4)
Childcare Exp.	2,534.57	6,999.46		
	(6,790.89)	(9,803.34)		
Children in Household	1.45	1.45	1.45	-0.007
	(0.62)	(0.59)	(0.63)	(0.006)
Years of Education	14.16	14.86	13.76	-1.096***
	(2.79)	(2.65)	(2.78)	(0.026)
Married	0.72	0.75	0.70	-0.041***
	(0.45)	(0.44)	(0.46)	(0.004)
Family Income ($10,000)	8.11	10.36	6.83	-3.526***
	(9.11)	(10.43)	(7.99)	(0.084)
Age of Youngest Child	1.84	2.02	1.73	-0.284***
	(1.54)	(1.52)	(1.54)	(0.014)
Number of Observations	49,278	17,844	31,434	

Data are from the CPS ASEC supplement. Means and standard deviations in parentheses are presented in columns 1-3. Column 4 presents the difference in means across columns 2 and 4 with the standard error in parenthesis.
* 0.10 ** 0.05 and *** 0.01 denote significance levels.

The distribution of the variable of interest, Childcare Expenses, is displayed in Figure 2. The top left panel presents the raw data. The top right panel displays the log transformation. As displayed, this causes all the zeros in the data to be dropped. In the bottom row of the figure, I display the distributions of the level and the log transformation when the value of one is added to each observation. The two figures on the right show very different distributions of the dependent variable. If the zeros are random, i.e. in our context, not correlated with number of children, then it can be shown that discarding those observations will not affect the regression estimates. However, in most cases the zeros are not random, and one needs to seriously consider how to move forward. Furthermore, adding an arbitrary scalar will impact the distribution. For example, if one was to instead transform the data by log(y+.000001), the spike in the distribution of the log transformation would be much further to the left than the distribution displayed in the bottom right figure.

Figure 2: Distribution of Childcare Expenses



Columns 2 and 3 of Table 1 present the summary statistics for households that have positive childcare expenses and households that have zero childcare expenses. Column 4 reports the difference in means between the two groups. Aside from the number of children in the household, each of the means are statistically different from the other at the .01 significance level, suggesting that the zeros in the data are not random.

Teaching Points of Emphasis #2:

- *The non-linear nature of the log function helps to normalize skewed data. It reduces positive skewness, meaning that if your data is heavily right-skewed (a long tail on the right), the log transformation can help make the distribution more symmetrical.*
- *The log transformation only works for positive values since the logarithm is undefined for zero or negative values.*
- *Adding an arbitrary scalar inside the transformation makes the zeros relevant but creates a spike in the distribution. Since the log transformation expands for values close to zero, adding a small constant, like 10^{-8}, within the transformation would increase the spacing between the spike and the bell-shaped data displayed in the bottom right panel of Figure 1.*
  - *The minimum, mean, and max of the log of childcare expenses plus 1 ($ln\ (child\ care + 1)$) is 0, 3.05, and 13.38, respectively.*
  - *The minimum, mean, and max of the log of childcare expenses plus .00000001 ($ln\ (child\ care + 10^{-8},)$) is -18.42, -8.70, and 13.38, respectively.*

- *Difference in means tests for observations that have zero childcare expenditures and positive childcare expenditures suggest that the zeros in the data are not random.*

Regression Model and Estimation

The regression model of interest is for household i is presented as,

$$\log(y_i) = \alpha + \beta nchild_i + \Gamma X_i + u_i,$$

where y_i represents household childcare expenditures, $nchild$ is number of children in the household, X is a vector of control variables, u is the unobserved term that accounts for all other variables that factor into household childcare expenses. The parameter of interest, β, is a semi-elasticity that approximates the proportionate change in childcare expenses for a unit increase in the number of children. A one unit increase in number of children is associated with a $(\exp(\beta) - 1) \times 100\%$ change in childcare expenses. This can be approximated by $\beta \times 100\%$ for relatively small values of β. I assume that the value of β is positive as each additional child in the home increases childcare expenses, though for this paper the more interesting question is the magnitude of the estimate.

Table 2: Regression Results

	$\ln(y + 1{,}000)$	$\ln(y + 1)$	$\ln(y + .01)$	$\ln(y + 10^{-4})$	$\ln(y + 10^{-8})$	Poisson
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Simple Linear Reg.						
Children	0.079***	0.126***	0.148***	0.179***	0.210**	0.295***
	(0.007)	(0.030)	(0.045)	(0.069)	(0.093)	(0.014)
N	49,278	49,278	49,278	49,278	49,278	49,278
Panel B: Mult. Linear Reg.						
Children	0.106***	0.271***	0.373***	0.527***	0.680***	0.327***
	(0.007)	(0.029)	(0.045)	(0.069)	(0.092)	(0.015)
Years of Educ.	0.059***	0.228***	0.342***	0.513***	0.684***	0.156***
	(0.002)	(0.007)	(0.011)	(0.017)	(0.023)	(0.004)
Married	-0.031***	-0.263***	-0.442***	-0.711***	-0.980***	0.175***
	(0.010)	(0.042)	(0.066)	(0.101)	(0.136)	(0.024)
Family Inc.	0.023***	0.074***	0.108***	0.158***	0.209***	0.017***
	(0.001)	(0.004)	(0.006)	(0.008)	(0.011)	(0.001)
Yng Child = 1	0.190***	0.712***	1.062***	1.587***	2.112***	0.405***
	(0.011)	(0.048)	(0.074)	(0.113)	(0.153)	(0.042)
Yng Child = 2	0.279***	1.066***	1.595***	2.387***	3.180***	0.586***
	(0.012)	(0.052)	(0.080)	(0.123)	(0.165)	(0.048)
Yng Child = 3	0.343***	1.401***	2.117***	3.190***	4.263***	0.623***
	(0.014)	(0.059)	(0.091)	(0.139)	(0.187)	(0.044)
Yng Child = 4	0.347***	1.473***	2.237***	3.381***	4.525***	0.627***
	(0.015)	(0.067)	(0.104)	(0.159)	(0.215)	(0.047)
Yng Child = 5	0.280***	1.219***	1.855***	2.807***	3.759***	0.511***
	(0.016)	(0.074)	(0.115)	(0.177)	(0.239)	(0.051)
State FE	Yes	Yes	Yes	Yes	Yes	Yes
N	49,278	49,278	49,278	49,278	49,278	49,278

Data are from the 2010-2018 CPS ASEC supplement. The table displays results from 12 regressions. Column 1-5 are from the log-linear OLS regressions. Column 6 is from the Poisson regression. The results in panel B include controls for years of educations for the head of household, marital status, total family income, fixed effects for the age of the youngest child, and state of resident fixed effects. Robust standard errors are in parentheses.
* 0.10 ** 0.05 and *** 0.01 denote significance levels.

Panel A and B of Table 2 report results from the simple and multivariate regression models. Columns 1-5 display results from the log-linear regression models and column 6 presents results for the Poisson model (discussed later). For the log-linear models, the dependent variable is listed in the heading of the table. For example, column 1 presents the results for the transformation of the dependent variable as $\ln(y + 1{,}000)$ where y is household expenditures on childcare. Column 2 reports results where the dependent variable is $\ln(y + 1)$, and so on. Robust standard errors are presented in parenthesis. As can be seen, each of the estimates is precisely estimated.

Teaching Points of Emphasis #3:

- *β can be interpreted as an approximate semi-elasticity: a one unit increase in number of children is associated with a $\beta \times 100\%$ change in childcare expenses. The exact semi-elasticity is $(exp(\beta) - 1) \times 100\%$.*
- *The results from the simple regression model (Panel A) have a common pattern, the smaller the value of the parameter, a, the larger the slope estimate. Depending on the value of a, the results present an estimate as small as 0.08 to as large as 0.21. By using the formula for the exact semi-elasticity, an additional child increases childcare expenses as little as 8.3 percent or as much as 23.4 percent.*
- *Small values of a, such as 10^{-8}, cause a significant increase in the slope coefficient as it transforms all the zero values to a value of -18.42 and therefore significantly expands the left end of the distribution of the logged values.*
- *The pattern of the estimates for the parameter of interest are similar in the multivariate regression model (Panel B) as those from the simple regression model (Panel A). The multiple regression estimates show that all the slopes are affected by the choice of the arbitrary value.*
- *From the multivariate regression model and using the exact semi-elasticity formula, an additional child increases childcare expenses as little as 11.1 percent or as much as 97.4 percent, depending on the value of a.*
- *The results in columns 1-5 of Table 2 are not very informative because the choice of the arbitrary value significantly affects the estimates.*
- *The above point can be further highlighted by setting a to an incredibly small number, like $a = 10^{-100}$. As such, the slope estimate on number of children increases to 1.16 and 5.38 for the simple and multiple regression models, respectively (not shown in the table). Or, one could set a to a really large number, like the max in the sample. In this case, the slope estimate on number of children decreases to 0.0013 and 0.0014 for the simple and multiple regression models, respectively.*

Poisson Regression Model

I now turn to fitting a Poisson regression model. Why might the Poisson be used instead of the popular fix of adding a constant within the log transformation? The Poisson model can handle all non-negative values of the dependent variable, i.e. $y \geq 0$. Also, the Poisson regression model is easy to implement with statistical software. Gould (2011) cites two other reasons why the Poisson has a significant advantage over log-linear regressions. First, “small nonzero values, however, they arise, can be influential in log-linear regressions. 0.01, 0.0001, 0.0000001, and 0 may be close to each other, but in the logs they are -4.61, -9.21, -16.12, and -∞ and thus not close at all. Pretending that the values are close would be the same as pretending that that exp(4.61)=100, exp(9.21)=9,997, exp(16.12)=10,019,062, and exp(∞)=∞ are close to each other. Poisson regression understands that 0.01, 0.0001, 0.0000001, and 0 are indeed nearly equal." Second, "when estimating with Poisson, you do not have to remember to apply the $\exp(\sigma^2/2)$ multiplicative adjustment to transform results from ln(y) to y." Particularly, when working with a log-linear regressions, to obtain the predicted values they must first obtain the predicted log values from regressing $\ln(y)$ on x, then exponentiate the predicted log values, and finally, multiply those exponentiated values by $\exp(\sigma^2/2)$, where σ is the root mean squared error (standard error of the regression).

The Poisson model takes the form of

$$y_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i).$$

The assumption of the Poisson model is that the mean is equal to the variance, i.e. $E(y) = var(y)$. However, the estimated coefficients of the maximum-likelihood process do not depend on this assumption. In other words, the estimates of the slope coefficients are unaffected if the assumption holds or not. Rather, only the standard errors are affected by violating this assumption. However, one can overcome this issue by simply calculating Huber-White robust standard errors.[5]

How does one interpret the estimates from the Poisson model? Exponentiating β_k transforms the log effect into a rate ratio (also called a relative risk). The rate ratio interpretation is as follows

- $e^{\beta_k} > 1$: A one-unit increase in x_k is associated with an increase in the expected count.
- $e^{\beta_k} < 1$: A one-unit increase in x_k is associated with a decrease in the expected count.
- $e^{\beta_k} = 1$: No effect of x_k on the expected count.

The practical interpretation is calculated by a one unit increase in x_k is associated with a $\left(e^{\beta_k} - 1\right) \times 100\%$ change in the outcome. In other words, if $\beta_k = .3$, then $e^{0.3} \approx 1.35$, and therefore a one-unit increase in x_k is associated with an 35% increase in the outcome. Similarly, if $\beta_k = -.8$, then $e^{-0.8} \approx 0.45$, and therefore a one-unit increase in x_k is associated with a 65% decrease in the outcome.

Referring to column 6 of Table 2, the 0.295 coefficient in the simple Poisson regression model suggests that an additional child increases childcare expenses by 34.3%. The coefficient of 0.327 from the multivariate Poisson model in Panel B suggests a 38.7% increase in childcare expenses from an additional child.

Teaching Points of Emphasis #4:

- *The Poisson model can handle all non-negative values of the dependent variable, including zeros.*
- *Small non-zero values in log-linear regressions: 0.01, 0.0001, 0.0000001, and 0 may be close to each other, but in the logs they are -4.61, -9.21, -16.12, and -∞ and thus not close at all. Poisson regression understands that 0.01, 0.0001, 0.0000001, and 0 are nearly equal.*
- *The interpretation of the coefficients of the Poisson model is calculated by a one unit increase in x_k is associated with a $\left(e^{\beta_k} - 1\right) \times 100\%$ change in the outcome.*
- *The results in column 6 of Table 2 suggest that an additional child increases household childcare expenses by 34.3% or 38.7%, depending on the choice of the simple or multivariate Poisson model.*

[5] White-Huber robust standard errors can be calculated in Stata using the vce(robust) option within the poisson command.

Inverse Hyperbolic Sine Transformation

In recent years, the inverse hyperbolic sine (IHS) transformation has been used as an alternative to the log transformation, when dealing zeros in regression analysis. The IHS transformation is defined as $IHS(y) = \ln\,(y + \sqrt{y^2 + 1})$. This function is well defined for all real numbers, including $y = 0$. Although this approach has been advocated as an alternative to the log transformation, it is not without its own shortcomings. The inverse hyperbolic sine function is not invariant to scaling. As such, regression results will depend on the units of measurement of the transformed variable. Arbitrary choices regarding the units of measurement will significantly affect the estimated marginal effects (Aihounton and Henningsen, 2021). In addition to the effect of scaling, Chen and Roth (2022) caution estimating elasticities using the IHS transformation because it combines intensive and extensive margin effects. For further discussion, see Chen and Roth (2022).

Conclusion

Given the ease of interpretation, abilities to smooth data, and other practical benefits, it is often appealing for researchers to apply log-linear regression models to data. Unfortunately, in many instances the data includes many zeros. To get around these issues, many researchers apply the popular fix of adding an arbitrarily chosen value to zero to allow for a log transformation, this is despite the well-known problems with this approach. In this paper we demonstrate that the choice of the arbitrary value can have a significant effect on the regression results. Rather than previous approaches, which are largely theoretical, I make use of a real-world, easy-to-understand example. I estimate a model that relates household childcare costs to the number of young children in the home and show that the additional cost of child can be estimated to be quite small by adding a relatively large arbitrary value or quite large by adding an arbitrary value close to zero. Overall, the estimates can be changed to be as large or close to zero as the research may desire by choosing an arbitrary value that works accordingly.

To get around this problem, I suggest that the best method is to introduce the Poisson model to students as it is straightforward, and it allows practitioners to directly deal with zeros in their data. Furthermore, I demonstrate how to interpret the Poisson model coefficients. While the interpretation of the Poisson model is more difficult than the log-linear model, this difficulty is outweighed by the clear benefits of the modified approach. In conclusion, the results on the childcare cost of an additional child are not arbitrary. Rather, by using the Poisson regression model, I find that an additional child that is age five and under increases childcare costs by approximately 35% for the family unit.

References

Aihounton, Ghislain BD, and Arne Henningsen. "Units of measurement and the inverse hyperbolic sine transformation." *The Econometrics Journal* 24.2 (2021): 334-351.

Bartlett, Maurice S. "The use of transformations." Biometrics 3.1 (1947): 39-52.

Bellégo, Christopher, David Benatia, and Louis Pape. "Dealing with logs and zeros in regression models." *arXiv preprint arXiv:2203.11820* (2022).

Chen, Jiafeng, and Jonathan Roth. "Logs with zeros? Some problems and solutions." *The Quarterly Journal of Economics* 139.2 (2024): 891-936.

Flood, Sarah, and Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and

Michael Westberry. Integrated Public Use Microdata Series, Current Population Survey: Version 12.0 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D030.V12.0

Flowerdew, Robin, and Murray Aitkin. "A method of fitting the gravity model based on the Poisson distribution." *Journal of Regional Science* 22.2 (1982): 191-202

Gould, William.22 August 2011. "Use poisson rather than regress; tell at friend" at _The Stata Blog_, https://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/

King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science*, 838-863.

OpenAI (2024). ChatGPT (June 14 version) [Large language model]. https://chat.openai.com/chat

Williams, C. B. "The use of logarithms in the interpretation of certain entomological problems." *Annals of Applied Biology* 24.2 (1937): 404-414

Zhao, Qingyuan, and Trevor Hastie. "Causal interpretations of black-box models." *Journal of Business & Economic Statistics* 39.1 (2021): 272-281.

Appendix

Below is the Stata code for the analysis. Stata 18 SE was used for the analysis. The data can be found here: https://www.dropbox.com/scl/fi/qcq2hidvqcqvvbkvi6uqq/data.dta?rlkey=211tn88bcob8p7sty6sc1eofr&dl=0

```
**********************************
*** Setup
**********************************
use data.dta, clear

keep if pernum==1
keep if eldch<=5
keep if nchild>=1

gen cc=spmchxpns
gen cc1=cc+1
gen lncc=ln(cc)
gen lncc1000=ln(cc+1000)
gen lncc1=ln(cc+1)
gen lncc01=ln(cc+.01)
gen lncc00001=ln(cc+.00001)
gen lncc100m=ln(cc+.00000001)
gen married=marst<=2
replace ftotval=ftotval/10000
forvalues i=0(1)5{
gen yn`i'=yngch==`i'
}
gen yred=.
replace yred=0 if educ==2
replace yred=1 if educ==11
replace yred=2 if educ==12
replace yred=2.5  if educ==10
replace yred=3 if educ==13
replace yred=4 if educ==14
replace yred=5 if educ==21
replace yred=5.5  if educ==20
replace yred=6 if educ==22
replace yred=7 if educ==31
replace yred=7.5  if educ==30
replace yred=8 if educ==32
replace yred=9  if educ==40
replace yred=10  if educ==50
replace yred=11  if educ==60
replace yred=11  if educ==71
replace yred=11 if educ==72
replace yred=12  if educ==73
replace yred=13 if educ==80
replace yred=13  if educ==81
replace yred=14 if educ==90
replace yred=14  if educ==91
replace yred=14  if educ==92
replace yred=15 if educ==100
replace yred=16 if educ==110
replace yred=16  if educ==111
replace yred=17 if educ==121
replace yred=18 if educ==122
replace yred=18  if educ==123
replace yred=19  if educ==124
replace yred=21  if educ==125
**********************************
*** Figure 2: Log Function
**********************************
preserve
clear
set obs 2000
gen x=_n
replace x=x/1000
gen y=ln(x)
gen y1=ln(x+1)
gen ydot1=ln(x+.1)
gen ydotz1=ln(x+.01)
label var y "log(y)"
label var y1 "log(y+1)"
label var ydot1 "log(y+.1)"
label var ydotz1 "log(y+.01)"
twoway (line y x) (line y1 x) (line ydot1 x ) (line ydotz1 x)
restore
**********************************
*** Figure 2: Histograms
**********************************
hist cc,name(cc,replace)
hist cc1,name(cc1,replace)
hist lncc,name(lncc,replace)
hist lncc1,name(lncc1,replace)
**********************************
*** Table 1: Summary Statistics
**********************************
sum cc nchild yred married ftotval yngch
sum cc nchild yred married ftotval yngch if cc!=0
sum cc nchild yred married ftotval yngch if cc==0
capture gen z=cc==0
foreach var of varlist  cc nchild yred married ftotval yngch   {
            xi:  reg `var'  z
}
**********************************
```

```
*** Table 2: Regression Results
*********************************
foreach var of varlist lncc1000 lncc1 lncc01 lncc00001 lncc100m {
        reg `var'                nchild,vce(robust )
        reg `var'        yred married ftotval   yn1-yn5 nchild i.statefip ,vce(robust )
}
        poisson cc               nchild,vce(robust )
        poisson cc        yred married  ftotval   yn1-yn5 nchild  i.statefip ,vce(robust )

*********************************
*** Additional Results
*********************************
gen lnccsmall=ln(cc+1.000e-100)
sum cc,d
gen lnccbig=ln(cc+r(max))
        reg lnccsmall               nchild,vce(robust )
        reg lnccsmall        yred married ftotval   yn1-yn5 nchild i.statefip ,vce(robust )
        reg lnccbig                 nchild,vce(robust )
        reg lnccbig         yred married ftotval   yn1-yn5 nchild i.statefip ,vce(robust )
```