Count (and Count-Like) Data in Finance *

(accepted, *Journal of Financial Economics*)

Jonathan Cohn
University of Texas-Austin

Zack Liu
University of Houston

Malcolm Wardlaw
University of Georgia

July 2022

Abstract

This paper assesses different econometric approaches to working with count-based outcome variables and other outcomes with similar distributions, which are increasingly common in corporate finance applications. We demonstrate that the common practice of estimating linear regressions of the log of 1 plus the outcome produces estimates with no natural interpretation that can have the wrong sign in expectation. In contrast, a simple fixed-effects Poisson model produces consistent and reasonably efficient estimates under more general conditions than commonly assumed. We also show through replication of existing papers that economic conclusions can be highly sensitive to the regression model employed.

*Jonathan Cohn: jonathan.cohn@mccombs.utexas.edu, (512) 232-6827. Zack Liu: zliu@bauer.uh.edu, (713) 743-4764. Malcolm Wardlaw: malcolm.wardlaw@uga.edu, (706) 204-9295. Toni Whited was the editor for this article. We would like to thank Jason Abrevaya, Kenneth Ahern, Pat Akey, Andres Almazan, Aydoğan Altı, Tony Cookson, Sergio Correia, John Griffin, Daniel Henderson, Travis Johnson, Praveen Kumar, Sam Krueger, Aaron Pancost, Paul Povel, James Scott, Sheridan Titman, Toni Whited, Jeff Wooldridge, and participants in the Virtual Finance Seminar, the Financial Research Association conference, and seminars at the University of Alabama, University of Houston, and University of Texas at Austin for valuable feedback.

Researchers in finance routinely use regression analysis to model count-based outcomes such and other count-like outcomes that are inherently limited to non-negative values. Examples include number of corporate patents granted, tons of toxic emissions, number of workplace injuries, and miles between cities in which two businesses are located. Outcomes of this type often have highly right-skewed distributions with masses of values at zero – distributional features that present special challenges for regression analysis because they make simple linear regression inefficient. Aware of these challenges, researchers employ a variety of approaches to address them. However, some commonly used approaches lack econometric foundations and produce estimates with unclear interpretations.

In this paper, we use econometric analysis and simulations to assess commonly-used regression models of count and count-like outcomes. We also replicate data sets analyzed in six top-finance journal publications featuring two such outcomes and compare estimates from different regression models. Our main takeaway is that Poisson regression delivers estimates with natural interpretations, requires no special assumptions for valid estimation, typically fits outcomes of this type well, and, crucially for use in corporate finance applications, admits separable group fixed effects. In contrast, the common practice of estimating linear regressions of the log of one plus the outcome ("log1plus" regression) produces estimates that lack meaningful interpretation and suffer from inherent biases that can cause them to have the wrong sign in expectation. While the interpretation of log1plus regression estimates may not be pivotal for understanding a particular paper's conclusions, our replication analysis suggests that the choice between Poisson and log1plus regression typically has a larger effect on estimates than omitting the most important control variable in real-world applications.

A common approach when working with skewed outcomes in general is to log-transform the outcome variable and then estimate a linear regression of the transformed variable. This log-linear regression model corresponds naturally to an underlying constant-elasticity model, with regression coefficients conveniently interpretable as semi-elasticity estimates.

However, Santos Silva and Tenreyro (2006) show that consistent estimation of a log-linear regression model requires that the errors in the corresponding constant-elasticity model be homoskedastic, an assumption that may not hold in practice. We extend their analysis to show that heteroskedasticity in these errors can cause even the sign of a log-linear regression coefficient to be wrong and that controlling for fixed effects can worsen the bias. We also provide novel guidance on the direction and magnitude of the bias.

Log-linear regression may be practically infeasible when the outcome has a mass at zero since the log of zero is undefined. Researchers in finance and other fields often solve this problem by estimating log1plus regressions, which allow for the retention of observations with zero-valued outcomes. 69% of respondents to a recent EconTwitter poll reported that they have either estimated log1plus regressions or used a similar approach involving an inverse hyperbolic sine (IHS) transformation of the outcome.[1] While these approaches allow for retention of observations with zero-valued outcomes, they do not map into natural economic models, and the economic interpretation and econometric properties of the resulting estimators are not well-understood.

We first show that no economically-meaningful quantities can be recovered from log1plus regression coefficients. We then identify two sources of bias likely to be endemic in log1plus regression. First, the homoskedastic constant-elasticity error requirement for consistent estimation of a log-linear regression gives way to a requirement that model errors exhibit a particular and implausible form of heteroskedasticity. Second, the combination of nonlinearities in the relationship between the outcome and covariates and any nonlinear relationships among covariates can bias estimates of average effects. While this second problem is not specific to log1plus regression, any reasonable economic model of the outcome will produce a nonlinear relationship between the log of one plus the outcome and covariates, making the problem endemic in log1plus regression. Simulations show that log1plus regression co-

[1] `https://twitter.com/prof_cookson/status/1462892660545372163`

efficients can easily have the wrong sign in expectation, making it difficult to infer even the direction of a relationship reliably from these coefficients. The same problems arise with the IHS transformation.

An alternative to estimating a linear regression of a transformed outcome is to estimate a generalized linear model (GLM) such as the Poisson model. Like log-linear regression, a Poisson regression corresponds to an underlying constant-elasticity model. However, a Poisson regression can accommodate outcomes with a value of zero and requires no assumptions about higher order model error moments for consistent estimation. While deviation from the well-known conditional mean-variance equality restriction reduces the efficiency of Poisson estimates, it *does not induce any bias*. Of practical importance, Poisson regression admits separable group fixed effects. While other candidate regression models such as the negative binomial model, zero-inflated models, or the Type I Tobit model may produce more efficient estimates than a Poisson model in certain circumstances, they do not admit separable fixed effects, which is a major limitation in corporate finance applications.[2] While Poisson regression explicitly models count data, it produces valid semi-elasticity estimates and standard errors even when the outcome variable is continuous (Santos Silva and Tenreyro, 2011).[3]

It is worth noting that the fixed effects in a Poisson regression are multiplicative rather than additive. An additive fixed effect affects only the mean of the outcome, but a multiplicative fixed effect scales both the mean and standard deviation of the outcome. While multiplicative fixed effects may seem nonstandard, log-linearized regression also implicitly assumes a multiplicative fixed effects structure. Furthermore, multiplicative fixed effects are more natural when working with the types of outcomes that are the focus of this paper. For

[2]Researchers sometimes include group dummy variables as covariates when estimating one of these models. However, doing so gives rise to an incidental parameters problem that results in biased coefficient estimates (Lancaster, 2000). The STATA xtnbreg module allows for group-level variation in the conditional *variance* of the outcome but not in the conditional *mean*, which is generally the object of concern.

[3]Advances in graph theory and computational matrix algebra have produced fast, efficient algorithms for implementing Poisson regression models with multiple fixed effects, including the ppmlhdfe package for Stata (Correia et al., 2020) and feglm package for R.

example, the year-to-year standard deviation in number of patents granted is likely to be approximately 10 times as large for a firm that averages 10 patents per year as it is for a firm that averages 1 patent per year.

To assess the practical relevance of our analysis, we replicate data sets from six papers published in top finance journals that together study two count or count-like outcomes - firm-year corporate patents granted and facility-year toxic release volumes.[4] We choose one regression specification in the main table of each paper, estimate log1plus and Poisson regressions based on that specification, and compare the coefficients of interest. These coefficients differ markedly in all six cases and have *different signs* in three of the six, suggesting that inference about even the direction of a relationship is sensitive to regression model choice in real-world applications. For context, in all five cases involving regressions with control variables, switching from a log1plus to Poisson regression results in a larger change in the coefficient of interest than omitting the most important control variable, generally by a wide margin.

A handful of existing papers in different fields have analyzed the properties of estimators that are commonly-used when working with outcome variables limited to non-negative values. In early work, Cameron and Trivedi (1986) and Cameron and Trivedi (1996) explore the properties of different Poisson and negative binomial estimators [see also Cameron and Trivedi (2013) for a general discussion]. Santos Silva and Tenreyro (2006) show that log-linear regression produces biased estimates in the context of gravity models of trade if errors in the underlying constant elasticity model are heteroskedastic and that Poisson regression avoids this problem. Santos Silva and Tenreyro (2011) show that Poisson models work well even when the outcome variable is continuous or exhibits many zero values. King (1988) and O'Hara and Kotze (2010) show that log1plus regression produces biased estimates, though

[4] 44 papers modeling firm-year corporate patent counts appeared in the *Journal of Finance*, *Journal of Financial Economics*, or *Review of Financial Studies* between 2011 and 2020. Of these 44, 25 estimate log1plus regressions, and 23 of those 25 use this approach exclusively.

neither fully explores the underlying econometric causes or implications.

Our primary contribution to this literature is to identify specific sources of biases likely endemic to log1plus regression and demonstrate that log1plus regression estimates typically differ from Poisson regression estimates by more than the effect of omitting the most important control variable in real-world applications and often have the opposite sign. Thus, the inherent deficiencies in log1plus regression are of practical importance to empirical researchers. We also build on the analysis of Santos Silva and Tenreyro (2006) and others by showing heteroskedasticity in model errors can cause traditional log-linear regression to produce estimates with the wrong sign in expectation. While theory provides a strong prior on the sign of coefficients in gravity models of trade, this is rarely the case in finance, making concerns about correctly signing regression coefficients an important consideration in finance. In addition, we provide guidance on the direction and magnitude of the bias in a log-linearized model coefficient due to heteroskedasticity.

1 Econometrics

Financial economists typically conduct regression analysis to estimate the effect of a set of covariates on an economically meaningful outcome variable. The validity and reliability of the resulting estimates depend on the properties of the underlying regression model. This section examines the properties of estimates from different regression models commonly used when working with count and count-like outcomes. We present a series of takeaways based on our analysis and that of existing work.

1.1 Linear regression

One option when working with these outcomes is to simply estimate a classic linear regression of the form

$$y = \mathbf{x}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of coefficients and ϵ is a mean-zero error. However, if y has a skewed distribution, then ϵ is likely to have a skewed distribution as well, which reduces efficiency and makes appropriate confidence intervals difficult to determine. We show in Section 2 that the efficiency loss can be large.

1.2 Log-linear regression

One solution to this problem is to estimate a log-linear regression - that is, a linear regression where $log(y)$ is the dependent variable. The concavity of the log function reduces skewness and hence can improve efficiency. Formally, log-linear regression takes the form

$$log(y) = \mathbf{x}\boldsymbol{\beta} + \epsilon. \tag{2}$$

A log-linear regression relates y exponentially to a linear combination of covariates and therefore implies an underlying constant-elasticity model, with coefficients interpreted as semi-elasticity estimates. However, Santos Silva and Tenreyro (2006) show that the consistency of log-linear regression estimates depends on the relationship between higher order moments of the errors in the implied model and covariates.

Takeaway 1. *Heteroskedasticity in the implied constant elasticity model can cause log-linear regression estimates to be biased and inconsistent (Santos Silva and Tenreyro, 2006).*

To see why this is the case, observe that $E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}}$ in a constant elasticity model.

Adding an error term yields the regression equation

$$y = e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon'. \tag{3}$$

Since $log(y) = log(e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon') \neq \mathbf{x}\boldsymbol{\beta} + log(\epsilon')$, (3) cannot readily be transformed into (2). A more natural and useful formulation of the underlying model (Weber and Hawkins, 1971) is

$$y = e^{\mathbf{x}\boldsymbol{\beta}}\eta, \tag{4}$$

where η is a multiplicative error. Observe that (3) and (4) are related by $\eta = 1 + \frac{\epsilon'}{e^{\mathbf{x}\boldsymbol{\beta}}}$. Log-transforming (4) yields

$$log(y) = \mathbf{x}\boldsymbol{\beta} + log(\eta). \tag{5}$$

Consistent estimation of $\boldsymbol{\beta}$ requires that $E[\epsilon|\mathbf{x}] = E[log(\eta)|\mathbf{x}] = E[log(1 + \frac{\epsilon'}{e^{\mathbf{x}\boldsymbol{\beta}}})|\mathbf{x}]$ be orthogonal to $\mathbf{x}$. However, it is impossible to factor ϵ' out of this expression even if ϵ' is independent of $\mathbf{x}$ because of the nonlinearity of the *log* function. Thus, orthogonality of an additive error in the implied economic model to $\mathbf{x}$ is insufficient to ensure consistency.

Suppose instead that ϵ' can be represented as $\epsilon' = e^{\mathbf{x}\boldsymbol{\beta}}\nu$, with ν independent of $\mathbf{x}$. This assumption implies that $\eta = 1 + \nu$. It follows that $E[log(\eta)|\mathbf{x}] = E[log(1+\nu)|\mathbf{x}] = E[log(1+\nu)]$ and thus that $E[log(\eta)|\mathbf{x}]$ is uncorrelated with $\mathbf{x}$. Recall that $e^{\mathbf{x}\boldsymbol{\beta}}$ is the conditional mean of y. So, consistent estimation of $\boldsymbol{\beta}$ in (5) requires that the standard deviation of ϵ' scale with the conditional mean of y. This condition is equivalent to assuming that η in (4) is homoskedastic. It follows that heteroskedasticity in η - a plausible scenario in reality - can cause the estimates from log-linear regression to be biased and inconsistent.

Santos Silva and Tenreyro (2006) show that heteroskedastic model errors can bias log-linear regression estimates. However, they do not examine whether the bias can cause regression estimates to have the wrong sign rather than just the wrong magnitude. Santos Silva

and Tenreyro (2006) consider log-linear regressions in the context of a gravity model of trade, where negative semi-elasticities would be nonsensical and only the magnitude of the semi-elasticities is in question. In corporate finance, we typically lack strong priors about the sign of a relationship. In addition, Santos Silva and Tenreyro (2006) do not analyze the direction of the bias in log-linear regressions due to heteroskedastic model error. Knowing the direction of the bias may be useful since this information allows estimates to be interpreted as bounds. We extend the analysis of Santos Silva and Tenreyro (2006) by demonstrating that this bias can cause log-linear regression estimates to have the wrong sign and by linking the direction of the bias to the sign of the relationship between the variance of the model error and a covariate.

Takeaway 2. *Bias due to heteroskedastic model error can cause log-linear regression coefficients to have the wrong sign.*

We illustrate this possibility with a bivariate example, which we also use in simulations in Section 2. Suppose that $y = exp(\beta x)\eta$, where x normally distributed with mean 0 and standard deviation σ_x and η is log-normally distributed with mean 1 and standard deviation $\sigma_\eta(x) = exp(\delta x)$ for constant δ. The parameter δ determines the degree of heteroskedasticity in η and can be positive (errors "fanning out" with x) or negative (errors "funnelling in" with x). We show in A that the bias in a log-linear regression coefficient on x in this case is $-\delta/2$. The bias as a proportion of the true coefficient β then is $-\delta/2\beta$. If $\delta/\beta > 2$, then the bias is sufficient to cause the log-linear regression coefficient to have the wrong sign. This analysis also sheds light on the direction of the bias.

Takeaway 3. *All else equal, a positive (negative) relationship between the variance of the error in the implied constant elasticity model and a covariate generally biases the log-linear regression coefficient on that covariate downward (upward).*

We derive this relationship more generally by borrowing the concept of second-order

stochastic dominance from decision theory. From (5), the partial derivative of the conditional expectation of $log(y)$ with respect to covariate x_j is

$$\frac{\partial E[log(y)|\mathbf{x}]}{\partial x_j} = \beta_j + \frac{\partial}{\partial x_j} E[log(\eta)|\mathbf{x}]. \tag{6}$$

Let $F_{x_j}(\eta)$ denote the cumulative distribution of η for a given value of x_j. Suppose, for any pair x_{j1} and x_{j2} of values of x_j satisfying $x_{j2} > x_{j1}$, that $E[\eta|x_{j1}] = E[\eta|x_{j2}]$ but that the variance of η is smaller when $x_j = x_{j2}$ than when $x_j = x_{j1}$ in the sense of second-order stochastic dominance – that is, $\int_0^z [F_{x_{j1}}(\eta) - F_{x_{j2}}(\eta)]d\eta > 0$ for all z, with strict inequality for some z. Since $log(\eta)$ is increasing and concave, $\frac{\partial E[log(\eta)|\mathbf{x}]}{\partial x_j} > 0$ by the definition of second-order stochastic dominance. Thus, the second term on the right-hand side of (6) is positive. As a result, log-linear regression will produce an upward-biased estimate of the true β_j. By the same argument, if the variance of η increases with x_j in the sense of second-order stochastic dominance, then $\frac{\partial E[log(\eta)|\mathbf{x}]}{x_j} < 0$, and log-linear regression will produce a downward-biased estimate of β_j.

1.3 Log1plus regression

Count and count-like outcome variables often have a mass of values at zero. For example, approximately 69% of Compustat firms are granted zero patents in a given year. Because the logarithm of zero is undefined, estimating a traditional log-linear regression requires excluding observations with zero-valued outcomes. The exclusion of these observations raises concerns about efficiency and allows for estimation of only the intensive margin. Researchers in finance frequently circumvent this problem by adding 1 (or some other positive constant) to y before log-transforming. Doing so ensures that the transformed dependent variable is defined for all possible values of y, including 0. The resulting "log1plus" regression equation

is

$$log(1+y) = \mathbf{x}\boldsymbol{\lambda} + \phi. \tag{7}$$

Coefficient λ_j on covariate x_j estimates the semi-elasticity of $1+y$ with respect to x_j. It might be tempting to conjecture that this semi-elasticity is the same as the semi-elasticity of y with respect to x_j (up to an added constant) since the constant added to y is invariant to $\mathbf{x}$. However, this conjecture ignores a Jensen's inequality problem. In fact, log1plus regression coefficients have no economically meaningful interpretation.

Takeaway 4. *Log1plus regression coefficients are not interpretable as semi-elasticities of the outcome variable, nor can any economically meaningful relationship between the outcome variable and a covariate be recovered from a log1plus regression coefficient.*

The coefficient λ_j on covariate x_j in regression equation (7) has the following interpretation:

$$\lambda_j = \frac{1}{E[1+y|\mathbf{x}]}\frac{\partial E[1+y|\mathbf{x}]}{\partial x_j} = \frac{1}{1+E[y|\mathbf{x}]}\frac{\partial E[y|\mathbf{x}]}{\partial x_j} \neq \frac{1}{E[y|\mathbf{x}]}\frac{\partial E[y|\mathbf{x}]}{\partial x_j} = \beta_j, \tag{8}$$

where β_j is the coefficient on covariate x_j in the log-linear regression equation (5). The relationship between the semi-elasticities of $1+y$ and y is

$$\lambda_j = \frac{E[y|\mathbf{x}]}{1+E[y|\mathbf{x}]}\beta_j. \tag{9}$$

Since $E[y|\mathbf{x}]$ is not observable and, indeed, the objective of regression analysis is typically to characterize $E[y|\mathbf{x}]$, the semi-elasticity of y cannot be recovered from λ_j, nor apparently can any other quantity of economic interest. When $E[y|\mathbf{x}]$ is large, $\lambda_j \approx \beta_j$, and log1plus regression coefficient λ_j can be interpreted as an approximation of the semi-elasticity of y with respect to x_j. However, when $E[y|\mathbf{x}]$ is large, y is likely to be zero for few observations, and thus the addition of the constant is unlikely necessary to begin with. In contrast, when

$E[y|\mathbf{x}]$ is small, the difference between between λ_j and β_j is large, and log1plus regression coefficients provide poor approximations of meaningful semi-elasticities.

This deficiency of the log1plus regression approach makes it difficult to determine the economic importance of a relationship. However, it is possible that a researcher is concerned only about establishing the sign of a relationship and not about its economic magnitude. We show next that log1plus regression is likely to be subject to two specific sources of bias that can make inferring even the direction of a relationship difficult.

Takeaway 5. *Log1plus regression is almost certain to suffer from two forms of bias that make even the sign of a relationship difficult to infer from log1plus regression coefficients.*

The first source of bias is almost certain to arise if there are any nonlinear relationships among covariates. In general, the combination of nonlinear relationships among covariates and between the dependent variable and covariates can cause bias, with mis-specification of the relationship between the dependent variable and one covariate contaminating the coefficients on other covariates. This problem is endemic in log1plus regression because any plausible economic model of y would produce a nonlinear relationship between $log(1+y)$ and covariates. For example, a constant elasticity model, which specifies a linear relationship between $log(y)$ and x_j, produces a nonlinear relationship between $log(1+y)$ and x_j.[5] We illustrate the intuition with a simple example.

Suppose that $log(y) = \beta_1 x_1 + \beta_2 x_2$, with $\beta_1 = 1$, $\beta_2 = 0$, and x_1 uniformly distributed over $[-4, 4]$.[6] Let ϵ_{x1} denote the error from a linear regression of $log(1+y)$ on x_1, and consider a linear regression of ϵ_{x1} on x_2. Note that the coefficient on x_2 in the second regression is, by construction, equivalent to the coefficient on x_2 from a regression of $log(1+y)$ on x_2, controlling for x_1.

[5] Even if the relationship between $log(1+y)$ and a covariate is nonlinear, a *univariate* log1plus regression coefficient still represents a valid estimate of the average effect of the covariate on $log(1+y)$, though it is unclear why this object would ever be of interest.

[6] We do not include an error term in y to make this illustration as simple as possible.

Figure 1 illustrates several relationships from this exercise. Figure 1(a) plots $log(1+y)$ against x_1 along with a regression line. While a regression of $log(y)$ on x_1 has no error by construction, a regression of $log(1+y)$ on x_1 does, since $log(1+y)$ has a nonlinear relationship with $log(y)$. Figure 1(b) plots ϵ_{x1} against x_1 along with a regression line. The slope of this line is zero because $corr(\epsilon_{x1}, x_1) = 0$ by assumption. However, even though they are uncorrelated, ϵ_{x1} and x_1 are not independent – ϵ_{x1} has a u-shaped relationship with x_1.

[Insert Figure 1]

Consider now three cases for the realizations of x_2: (i) x_2 independent of x_1 and drawn from a uniform distribution on $[0,1]$, (ii) $x_2 = x_1$, and (iii) $x_2 = x_1^2$. Figures 1(d), 1(e), and 1(f) plot ϵ_{x1} against x_2 along with regression lines for cases (i), (ii), and (iii), respectively. The slope of the regression line is, correctly, zero in Figures 1(c) and 1(d). The former is true because x_2 is unrelated to both x_1 and y by assumption. The latter is true because $corr(\epsilon_{x1}, x_2) = corr(\epsilon_{x1}, x_1) = 0$, which would hold for any linear relationship between x_2 and x_1.

Figure 1(e) shows that ϵ_{x1} is positively correlated with x_2 when $x_2 = x_1^2$. Observe that x_2 is large when x_1 is high or low and small when x_1 is in an intermediate range of values. Because ϵ_{x1} is also large for high and low values of x_1 and small for intermediate values of x_1, ϵ_{x1} and x_2 are indirectly positively correlated. So, the coefficient on x_2 in a regression of $log(1+y)$ on x_1 and x_2 will be positive, even though $log(1+y)$ is independent of x_2 by assumption. Note that the coefficient on x_1 may also be biased. More generally, any nonlinear relationship between two covariates is almost certain to bias the coefficients in a linear regression of $log(1+y)$ on these covariates.

The second reason that log1plus regression is almost certain to produce biased estimates is that unbiased estimation requires an implausible assumption about the relationship between

higher order model error moments and covariates. The closest reasonable economic model to a log1plus regression is a constant elasticity model. Suppose then that (3) is the true model – that is, $y = e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon'$. Adding 1 to both sides yields $1 + y = e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon^{1+}$, where $\epsilon^{1+} = \epsilon' + 1$. Writing the relationship in multiplicative form, we have $1 + y = e^{\mathbf{x}\boldsymbol{\beta}}\eta^{1+}$, where $\eta^{1+} = 1 + \frac{\epsilon^{1+}}{e^{\mathbf{x}\boldsymbol{\beta}}} = 1 + \frac{\epsilon'}{e^{\mathbf{x}\boldsymbol{\beta}}} + \frac{1}{e^{\mathbf{x}\boldsymbol{\beta}}}$. It is immediate that, unlike in the case of log-linear regression, assuming that ϵ' can be written as $\epsilon' = e^{\mathbf{x}\boldsymbol{\beta}}\nu$ with ν independent of $\mathbf{x}$ does not make $E[log(\eta^{1+})|\mathbf{x}]$ independent of x unless $\beta = 0$ for all non-constant coefficients. That is, homoskedasticity in the multiplicative error in a conventional constant-elasticity model is insufficient for consistent estimation of the log1plus model. Instead, what is required for consistent estimation is that $\epsilon' = e^{\mathbf{x}\boldsymbol{\beta}}\nu - 1$, a form or heteroskedasticity unlikely to be satisfied by any reasonable economic model.

As a final point, there is nothing special about the choice to add 1 before log-transforming. Coefficients from the resulting regression are no more interpretable than coefficients from a regression where a different positive constant is added to the outcome variable before log-transforming. Consider the more general logcplus regression equation for constant $c > 0$

$$log(c + y) = \mathbf{x}\boldsymbol{\lambda}^c + \phi^c.$$

The jth coefficient in this regression estimates the semi-elasticity

$$\lambda_j^c = \frac{1}{c + E[y|\mathbf{x}]}\frac{\partial E[y|\mathbf{x}]}{\partial x_j}.$$

Observe that $\frac{\partial \lambda_j^c}{\partial c} = -\frac{1}{(c+E[y|x])^2}\frac{\partial E[y|\mathbf{x}]}{\partial x_j}$. Thus, the coefficient on covariate j mechanically and arbitrarily shrinks in magnitude as c increases.

1.4 IHS regression

Researchers occasionally use a non-logarithmic concave transformation of a count or count-like outcome variable to address concerns about skewness. The most commonly-used of these is the inverse hyperbolic (IHS) transformation, $sinh^{-1}(y)$. It can easily be shown that linear regression of an IHS-transformed outcome suffers from the same problems that log1plus regressions do.

Takeaway 6. *Takeaways 4 and 5 hold for linear regression of an IHS-transformed outcome variable.*

1.5 Poisson regression

We next consider Poisson regression as an alternative when working with a count or count-like outcome variable. Poisson regression assumes that the dependent variable has a Poisson distribution that depends on covariates, with density $f(y|\mathbf{x}) = exp(-\mu(\mathbf{x}))\mu(\mathbf{x})^y/y!$, where $\mu(\mathbf{x}) = E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}}$. Conditional expectation in the Poisson model takes the form $E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}}$ or, equivalently,

$$log(E[y|\mathbf{x}]) = \mathbf{x}\boldsymbol{\beta}. \tag{10}$$

Poisson regression estimates have a number of desirable features.

Takeaway 7. *Poisson regression produces estimates with valid semi-elasticity interpretations and requires no assumptions about the relationship between higher-order model error moments and covariates for consistent estimation.*

A key difference between the Poisson and log-linear regression models is that Poisson regression estimates (10), while log-linear regression estimates $E[log(y)|\mathbf{x}] = \mathbf{x}\boldsymbol{\beta}$. By Jensen's inequality, $log(E[y|\mathbf{x}]) \neq E[log(y)|\mathbf{x}]$. Heteroskedasticity does not bias estimates of (10) because the conditional expectation is inside rather than outside the *log* function. Con-

ceptually, the chief advantage of Poisson regression relative to log-linear regression is that it applies an exponential model to relationships that are likely to be approximately exponential rather than transforming the outcome to make the data fit a linear model.

Takeaway 8. *Poisson regression imposes the restriction that the conditional mean and variance of the outcome are equal. Violation of this restriction reduces efficiency but does not cause any bias.*

Poisson regression does impose the restriction that $E[y|\mathbf{x}] = var(y|\mathbf{x})$. A common critique of Poisson regression is that the conditional mean-variance equality assumption is often violated in practice. In particular, the conditional variance is often larger than than the conditional mean, a situation known as "overdispersion" (the less common converse is known as "underdispersion"). Violations of this restriction reduce efficiency and make it important to report robust standard errors. However, crucially, they do not bias point estimates, so regression coefficients remain valid.[7]

Takeaway 9. *Poisson regression admits separable group fixed effects, and even Poisson models with high-dimensional fixed effects can now be estimated quickly and easily.*

One feature of Poisson regression that is crucial for use in corporate finance applications is that it admits separable group fixed effects. While computational limitations may have limited the usefulness of fixed-effects Poisson regressions in the past, the combination of improvements in computing power and innovations in sparse matrix reduction methods have made even high-dimensional fixed effects Poisson models fast and easy to estimate. Two packages for estimating high-dimensional fixed effects Poisson regressions are ppmlhdfe for Stata (Correia et al., 2020) and glmdhfe for R (Hinz et al., 2019).

[7]Overdispersion occurs when the *conditional* variance of y exceeds its *conditional* mean. In many cases where the *unconditional* mean of y exceeds its *unconditional* variance, conditioning on observables and, especially, group fixed effects reduces the variance of y relative to the mean.

Letting α_i be the fixed effect for group i, the fixed-effects Poisson model conditional expectation is:

$$E[y|\mathbf{x}] = exp(\alpha_i + \mathbf{x}\boldsymbol{\beta}) = exp(\alpha_i)e^{\mathbf{x}\boldsymbol{\beta}}. \tag{11}$$

Observe that, while the fixed effects in a linear model are additive, they are multiplicative in a Poisson regression, as they are implicitly in a log-linear regression. This feature of a Poisson model is generally desirable. Multiplicative fixed effects are likely to more accurately capture the effect of any fixed group-level factors on a count or count-like outcome than an additive fixed effect would. An additive fixed effect scales the mean of the outcome but not its dispersion. However, the standard deviation of a count or count-like variable is likely to scale with its mean. Thus, multiplicative fixed effects generally allow for better fit of the data, increasing power.

Takeaway 10. *Fixed-effects Poisson regression requires excluding any group for which the outcome variable is zero for all observations. However, this exclusion is not a shortcoming of Poisson regression, as these observations contain no information about regression coefficients in a regression model where the fixed effects are multiplicative.*

Fixed-effects Poisson regression does restrict the usable sample to groups for which the outcome variable is non-zero for at least one observation, which can meaningfully shrink the usable sample. For example, 219 of the 703 firms in the S&P 500 between 1990 and 2010 never patented during this period (Bellstam et al., 2021) and would thus be excluded in a Poisson regression where the dependent variable is patent count. The omission of these observations should not be thought of as a shortcoming in a fixed-effects Poisson model, and their omission does not bias Poisson regression estimates. Rather, these observations simply contain no information about regression coefficients in a model in which the fixed effects are multiplicative. To see why, observe that one possible explanation for why $y = 0$ for all observations in group i is that $exp(\alpha_i) = 0$. However, if $exp(\alpha_i) = 0$, then $E[y|\mathbf{x}] = 0 * e^{\mathbf{x}\boldsymbol{\beta}}$,

and the $\boldsymbol{\beta}$ coefficients are unidentified. Observe that the same is true in any multiplicative model. The lack of information in observations for groups where $y = 0$ for all observations is an example of a more general phenomenon that Correia et al. (2021) term "statistical separation."[8]

Takeaway 11. *Poisson regression produces valid estimates even when the outcome variable is continuous, admits an exposure variable that acts as a scaling variable for the outcome, and can be used in IV regression.*

This takeaway highlights three other useful features of Poisson regression. The standard approach to estimating a Poisson regression is to compute the Poisson Pseudo Maximum Likelihood (PPML) estimator by numerically solving the series of first-order conditions (Gourieroux et al., 1984):

$$\sum_{i=1}^{n}[y_i - exp(\mathbf{x}_i\boldsymbol{\beta})]\mathbf{x}_i = 0. \tag{12}$$

Examination of (12) shows that Poisson regression estimation imposes no restriction on the domain of y other than requiring $y \geq 0$ (Santos Silva and Tenreyro, 2011). Thus, Poisson regression can be estimated even if the distribution of y is continuous. Poisson regression allows for the specification of an "exposure" variable that captures the baseline exposure to the Poisson arrival process underpinning the Poisson model and serves as a scaling variable. When an exposure variable is specified, Poisson regression coefficients represent estimates of the semi-elasticities of the rate of outcome per unit of exposure (e.g., workplace injuries per employee). Finally, Poisson models can be used in instrumental variables (IV) regression (Mullahy, 1997; Windmeijer and Santos Silva, 1997).[9]

[8]To validate the lack of bias due to the exclusion of these observations, we conduct a series of tests using the replicated data sets that we describe in Section 3. Specifically, we add 0.01 to the outcome for one randomly-chosen observation for any group (firm or establishment) for which the outcome variable is zero for all observations. Doing so allows us to retain all observations for the group. In untabulated results, we find that all coefficients from Poisson regressions estimated on this altered data set differ from those estimated using the original data set by less than 0.5% in all cases, despite the substantial increase in reported sample size.

[9]See Karolyi and Taboada (2015) for a finance application.

1.6 Other count-based regression models

Other regression models that may be appropriate when working with count or count-like outcome variables include the negative binomial model and zero-inflated Poisson and negative binomial models. The negative binomial model has the same conditional expectation as the Poisson model but relaxes the conditional mean-variance equality restriction by explicitly modeling the variance as a separate gamma process that allows for overdispersion (but not underdispersion). Because it relaxes the mean-variance equality restriction, negative binomial regression may be more efficient than Poisson regression in some cases, especially if the true variance is approximately gamma distributed. Zero-inflated models account for the possibility that some observations are not exposed to the underlying process that drives y by explicitly modeling the relationship between exposure and observables. These models may be suitable when working with count data that has an excessive number of zero values, though factors affecting exposure but not the outcome conditional on exposure are generally difficult to identify. These alternative models have useful features. However, they all have one critical limitation – they do not admit separable group fixed effects.

Takeaway 12. *Negative binomial or zero-inflated Poisson/negative binomial regression may be more efficient than Poisson regression but do not admit separable group fixed effects. They are subject to an incidental parameters problem if group dummies are included as covariates, potentially biasing all of the estimates.*

In principle, one could include group dummy variables as additional covariates to approximate fixed effects, and researchers often do so when estimating these models. However, the inclusion of such dummies gives rise to an incidental parameters problem that causes the estimated coefficients on all variables to be biased and inconsistent (Lancaster, 2000). Asymptotically, estimates converge to the true coefficient values as the number of observations per group becomes large but not as the number of groups becomes large. Since

controlling for group fixed effects is often considered essential for identification in corporate finance applications, the inability of these alternative models to readily accommodate fixed effects limits their usefulness in the field.[10]

1.7 Rate regressions

Log-linear, log1plus, and inverse hyperbolic sine regression all decrease skewness in an outcome variable by applying a concave transformation to it. Poisson and negative binomial regressions both fit models that assume skewed outcomes. A third possibility is to scale the outcome variable, since skewness in an outcome is often partly attributable to skewness in scale. Let s denote a suitable scaling variable, and note that s is equivalent to an exposure variable in the context of a Poisson regression. Then, the following linear regression estimates the effect of a one-unit change in each covariate on the rate y/s:

$$y/s = \mathbf{x}\boldsymbol{\beta} + \epsilon \tag{13}$$

As an example, Cohn and Wardlaw (2016) and Cohn et al. (2020) estimate linear regressions of the number of workplace injuries at an establishment in a given year scaled by the average number of employees working at the establishment in the year. In Section 2.3, we compare the efficiency of Poisson regression with an exposure variable and OLS rate regression. Unfortunately, in most finance applications, a scaling variable that faithfully captures exposure does not exist. In principle, a noisy measure of exposure (for example, total assets as a measure of exposure for corporate patenting) can be used as a scaling variable. However, any correlation between the noise in the scaling variable and the outcome would contaminate estimation. It is also worth noting that one could, in principle, estimate a Poisson regression of the rate y/s, though we have not seen this approach used previously. We show in the next

[10]The same issue applies to the Type I Tobit model.

section that a Poisson regression of a rate dependent variable may produce marginally more efficient estimates than a comparable linear regression of the rate.

2 Simulations

This section presents three simulation exercises that further illustrate the econometric properties of different estimators when working with a skewed outcome limited to non-negative values. The first simulation examines the degree of bias that heteroskedasticity introduces into log-linear regression estimates and whether this bias can cause estimates to have the wrong sign in expectation. The second simulation examines how the addition of the constant in log1plus regression distorts estimates. The third simulation examines the statistical power of estimates from different regression models under various conditions.

2.1 Log-linear Regressions and Heteroskedasticity

In the first set of simulations, we illustrate the effect of heteroskedasticity on log-linear regression coefficients. While prior papers have demonstrated that heteroskedasticity can create estimation bias in regressions with logged dependent variables (Manning and Mullahy, 2001; Santos Silva and Tenreyro, 2006), they have typically focused on inaccuracies in the predicted value of y or in the quantitative relationships implied by regression coefficients. However, researchers in finance may be more interested in determining the direction of a relationship than in predicting outcomes or establishing quantitative relationships. In this set of simulations, we demonstrate that the bias can cause estimates to have the wrong sign in expectation, making them unreliable at estimating even the direction of a relationship.

We simulate data sets of observations (x, y), with $y = exp(\beta x)\eta$, where η is a mean-1 multiplicative error.[11] We set $\beta = 0.2$ and evaluate the effects of heteroskedasticity in two

[11] We write y as a function of a multiplicative error for convenience, though, as we describe in Section 1.2,

different scenarios - one where x is an i.i.d. random variable drawn from a standard normal distribution truncated at the 1st and 99th percentiles, and one where x is an i.i.d. binary random variable equal to 0 or 1 with equal probability. In both scenarios, we draw η from a lognormal distribution with a mean of 1 and standard deviation of $\sigma_\eta(x)$.[12] For the scenario where x is binary, we define $\sigma_0 = \sigma_\eta(0)$ and $\sigma_1 = \sigma_\eta(1)$. The error in this scenario is homoskedastic only if $\sigma_1 = \sigma_0$.

Within each scenario, we evaluate three specific cases - one where the variance of the error is positively related to x (fanning out), one where it is unrelated to x, and one where it is negatively related to x (funnelling in). Thus, we evaluate six specific cases altogether. For the continuous x scenario, we evaluate the following cases: (i) $\sigma_\eta(x) = exp(x)$, (ii) $\sigma_\eta(x) = exp(1/2)$, and (iii) $\sigma_\eta(x) = exp(-x)$. Note that we choose $\sigma_\eta(x) = exp(1/2)$ for case (ii) because $E[\sigma_\eta(x)] = exp(1/2)$ in cases (i) and (iii), so doing so keeps the unconditional variance the same across all three cases. For the binary x scenario, we evaluate the following cases: (i) $\sigma_1 = 2$ and $\sigma_0 = 1$, (ii) $\sigma_1 = \sigma_0 = 1.5$, and (iii) $\sigma_1 = 1$ and $\sigma_0 = 2$.

For each of the six cases, we generate 10,000 simulated data sets of 5,000 observations. We then estimate and compare Poisson and log-linear regressions using each data set. For completeness, we also estimate log1plus, log0.1plus, log10plus, and IHS regressions, though, as Section 1 makes clear, there is no reason to expect these regressions to recover the true coefficient. We estimate all linear regressions throughout the paper, including those with transformed outcome variables, using OLS. For each regression coefficient, we compute White-corrected robust standard errors. Finally, we compute the mean coefficient and standard error over the 10,000 simulations for each regression model in each of the six cases. Table 1, Panel A reports these means.

[Insert Table 1]

we can recast the error as an additive mean-0 error term.

[12] In untabulated results, we generally find that the bias caused by heteroskedasticity is larger for errors drawn from other distributions.

The mean coefficient from Poisson regression is approximately 0.2 in all six cases. That is, Poisson regression recovers the true model coefficients, on average, despite the presence of heteroskedasticity. As expected, log-linear regressions also recover the true model coefficients when the multiplicative error is homoskedastic (where $\sigma_\eta = exp(1/2)$ in the continuous x scenario and $\sigma_1 = \sigma_0 = 1.5$ in the binary x scenario). When $\sigma_\eta = exp(x)$ ($\sigma_\eta = exp(-x)$) in the continuous x scenario or $\sigma_1 > \sigma_0$ ($\sigma_1 < \sigma_0$) in the binary x scenario, the log-linear regression coefficient is less (greater) than the true parameter. The directions of the biases are consistent with our conclusion in Section 1 that a positive (negative) relationship between the variance of the error and a covariate generally downward (upward) biases log-linear regression estimates.

In case (i) in both scenarios, where the variance of the error increases with x in the simulations, the mean log-linear regression coefficient is negative, even though the true coefficient is positive. That is, heteroskedasticity in the implied economic model error can cause estimates to have the wrong sign in expectation. Similarly, if we assumed $\beta = -0.2$, then log-linear regression coefficients can incorrectly have a positive sign in expectation in case (iii), where the variance of the error decreases with x. Note that the bias in the continuous x scenarios is exactly the magnitude predicted by the formula we provide in Section 1.2.

To understand the practical implications of these conclusions, we compare the simulated data set from case (iii) of the continuous x scenario to a replicated corporate patent data set based on Fang et al. (2014), which we describe and further analyze in Section 3. Figure 2(a) presents a scatterplot of $log(y)$ against x based on simulated data from the continuous outcome scenario where $\sigma_\eta = exp(-x)$, with a sequence of bars depicting the range from from -3 to +3 standard deviations of $log(y)$ for different bins of x. Figure 2(b) presents a scatterplot of residuals from a linear regression of the log of the outcome (patents granted) on controls against the covariate of interest. This figure shows that the degree of funneling in observed in the replicated data set is comparable to that from the simulated data.

[Insert Figure 2]

Most regressions in corporate finance include group-level fixed effects such as firm fixed effects. We show next that the inclusion of fixed effects in a log-linear regression can either mitigate or exacerbate bias due to heteroskedasticity, depending on how much of the variation in the outcome is at the group level. To do so, we extend the heteroskedastic error framework above by specifying two components to the variance of the error term – one that is fixed at the group level and one that varies within group.

Let i index group and it denote observation t within group i. We assume that $x_{it} = 0.5\mu_i + 0.5\nu_{it}$, where μ_i and ν_{it} are i.i.d. random variables each drawn from a normal distribution truncated at the 1st and 99th percentiles. We then assume that $\sigma_\eta = exp(\gamma\mu_i + (1-\gamma)\nu_{it})$. We examine five cases, each corresponding to a different value of $\gamma \in [0, 1]$. For each case, we generate 10,000 data sets of 5,000 observations apiece, with 500 independent groups and 10 observations per group in each data set. For each simulation, we estimate Poisson and log-linear regression models, each with and without group fixed effects. Panel B of Table 1 reports the mean coefficient and standard error across simulated data sets for each regression, where we cluster standard errors at the group level.

As expected, Poisson regression without group fixed effects consistently estimates a mean coefficient of approximately 0.2 – the true value – in all cases. Also as expected, log-linear regression without group fixed effects generally results in a negatively biased coefficient in all cases since the variance of η increases with x by construction. However, when heteroskedasticity is driven entirely by variation at the group level ($\gamma = 1$), the fixed-effects log-linear regression recovers the true parameter value of 0.2, despite the heteroskedasticity in η. In contrast, when most of the relationship between the variance of η and x reflects a relationship with the within-group variation in x ($\gamma < 1/2$), then including group fixed effects in the log-linear regression magnifies the bias.

To see why this is the case, observe that, in a constant-elasticity setup, a group fixed

effect impacts both the level of the outcome and the variance of the error in the outcome since the fixed effect and error are both multiplicative. Controlling for fixed effects reduces (increases) bias due to heteroskedasticity if it removes proportionately more (less) of the variation driving the relationship between the variance of the error in y and x than it does the variation driving the relationship between the level of y and x. Controlling for fixed effects removes exactly half of the variation driving the relationship between the level of y and x by construction. If $\gamma > 1/2$, then controlling for fixed effects removes more than half of the variation driving the relationship between the variance of the error in y and x and therefore decreases bias due to heteroskedasticity. If $\gamma < 1/2$, then controlling for fixed effects removes less than half of this variation and therefore increases bias due to heteroskedasticity.

2.2 The effect of adding the constant in log1plus regression

As we demonstrate in Section 1, the addition of the constant in a log1plus regression causes two problems. First, estimates of the semi-elasticity of $1+y$ lack meaningful interpretation. Second, nonlinearities introduced into the relationship between the logged dependent variable and covariates by the addition of the constant may result in biased estimates of this semi-elasticity if covariates are nonlinearly related to each other. Our second set of simulations examines the extent to which these problems make it difficult to learn about the true relationship between the outcome variable and covariates using log1plus regression.

We simulate data sets of observations (x_1, x_2, y), with $y = kexp(\beta_1 x_1 + \beta_2 x_2)$, where k is a positive constant, $\beta_1 = 1$, and $\beta_2 = -0.1$. We do not include an error term in order to isolate the effect of adding a constant to log1plus regression coefficients from the effect of the relationships between the variance of the error and covariates that arises in any reasonable underlying model. The k parameter scales the conditional mean, which may be important for the quality of estimates produced by log1plus regression since the relationship between $log(y)$ and $log(1 + y)$ becomes approximately linear at higher values of y. For each observation,

we draw the value of x_1 from a standard normal distribution. We analyze three different specifications for x_2: (i) x_2 drawn from an independent standard normal distribution, (ii) $x_2 = .5x_1 + .5z$, where z is drawn from an independent standard normal distribution, and (iii) $x_2 = \max\{x_1, 0\}$. The second specification makes x_2 a linear function of x_1, while the third makes x_2 a nonlinear function of x_1.

We simulate six data sets of 5,000 observations each – one for each combination of $k = 1, 10$ and specification for x_2. Because there is no sampling error in our simulated data, we only need one data set for each combination. For each of the six simulated data sets, we estimate Poisson, log-linear, log1plus, log0.1plus, log10plus, and inverse hyperbolic sine (IHS) regressions of y on x_1 and x_2 plus a constant. Table 2 presents the regression results.

[Insert Table 2]

Panels A, B, and C report results for cases (i), (ii), and (iii), respectively. Each panel shows results for the cases where $k = 1$ on the left and $k = 10$ on the right. Poisson and log-linear regression both recover the true values of β_1 and β_2 in all three cases. Panels A and B show that log1plus regression produces coefficients with the correct sign but incorrect magnitudes when x_2 is independent of x_1 or linear in x_1. However, in Panel C, where x_2 is a nonlinear function of x_1, the estimated coefficient β_2 is positive while the true value is negative. Thus, it appears that the bias in log1plus regression due to nonlinearities among covariates that we describe in Section 1 can cause coefficients to have the incorrect sign, even absent sampling error. As a result, a researcher estimating a log1plus regression using this simulated data would conclude that x_2 has a positive effect on y, while the true effect is negative.

Log1plus, log0.1plus, and log10plus regressions all yield sharply different coefficients on both x_1 and x_2. While expected and purely mechanical, the sensitivity of the coefficients to a purely arbitrary choice of constant highlights the lack of meaning in logcplus regressions in

general.[13] The coefficient on x_2 in the IHS regression also has the wrong sign when x_1 and x_2 are nonlinearly related. Finally, it is worth noting that the logcplus and IHS regression coefficients are all closer to the true values when k is larger.

2.3 Efficiency of three unbiased estimators

In our final set of simulations, we explore the efficiency of different regression models when confronted with count or count-like outcome variables. We evaluate three regression models – linear, Poisson, and linear rate. These three models admit fixed effects and do not require assumptions about higher order moments of model error for consistent estimation. We conduct analysis for both count and continuous outcomes.

We simulate panels of observations (x_1, x_2, y). For each observation it, where i denotes a group and t an observation within the group, we draw two random variables, μ_i and ν_{it}, each from an independent standard normal distribution truncated at the 1st and 99th percentiles. We then set $x_{1,it} = 0.5\mu_i + 0.5\nu_{it}$. This structure produces a group fixed effect in x_{it}. We assume a panel structure so that we can evaluate the efficiency of different estimators in a setting that approximates real-world applications in finance. We independently draw $x_{2,it}$ from a normal distribution with a mean of 0 and a standard deviation of 2.

To produce count outcomes, we draw y_{it} from a negative binomial distribution with conditional mean $E[y_{it}|\mathbf{x}] = exp(\beta_1 x_{1,it} + x_{2,it})$ and overdispersion parameter α_{NB}, which captures deviations from the conditional mean-variance equality restriction imposed by Poisson regression. Since it has a coefficient of 1, $x_{2,it}$ plays the role of an exposure variable, which is suitable for scaling the outcome. To produce continuous outcomes that are skewed, limited to non-negative values, and potentially have a mass at zero, we model y_{it} as a continuous variable using the mixture model approach of Santos Silva and Tenreyro (2011). In

[13] One potential approach to addressing the distortions caused by the added constant is to estimate the value of the constant necessary to recover the correct relationship between the outcome and covariates. See Bellego et al. (2021) for a clever implementation of this approach.

this formulation, y_{it} is the sum of m_i random variables z_{it}, where m_i is a negative binomial-distributed random variable with mean $exp(\beta_1 x_{1,it}+x_{2,it})$, and z_{it} is a $\chi^2_{(1)}$-distributed random variable. Again, $x_{2,it}$ plays the role of an exposure variable. We set the variance of m_i to $E[m_i|\mathbf{x}] + bE[m_i|\mathbf{x}]^2$, which implies that $Var(y_{it}|x_{it}) = 3 * E[y_{it}|x_{it}] + b * E[y_{it}|x_{it}]^2$, where b is a parameter that determines the conditional variance and hence degree of overdispersion in the data.

For both the count and continuous outcomes, we evaluate 12 different cases, each a different combination of β_1 parameter and degree of overdispersion as captured by α_{NB} or b. We consider four values of β_1: -2, -0.2, 0.2, and 2. For count (continuous) outcomes, we consider three values of α_{NB} (b): 0.001, 0.5, and 2. An α_{NB} or b of 0.001 approximates the case where there is no overdispersion. An α_{NB} or b of 0.5 approximates the degree of overdispersion in common skewed data sets such as firm-year corporate patents. An α_{NB} or b of 2 represents extreme overdispersion.

For each of the two outcome types (count and continuous) and each of the 12 cases we evaluate, we generate 10,000 simulated panels of 5,000 observations. Each panel consists of 500 groups, each with its own value of μ_i, with 10 observations per group. We then estimate four regressions, each with group fixed effects, using each simulated panel. These are linear and Poisson regression where the dependent variable is y_{it} and linear and Poisson regression where the dependent variable is $y_{it}/x_{2,it}$. We set $x_{2,it}$ as the exposure variable in the Poisson regression where y_{it} is the dependent variable. For each combination of outcome type, overdispersion level, coefficient β_1, and regression model, we compute two quantities. The first is the percentage of the 10,000 simulated panels in which the regression coefficient on x_1 has the same sign as the true value of β_1 and is statistically different from 0 at the 5% level. The second is the root mean squared error (RMSE). Table 3 reports these quantities.

[Insert Table 3]

Panel A reports results for count outcomes, while Panel B reports results for continuous outcomes. Not surprisingly, linear regression where the dependent variable is y_{it} exhibits the least power in all scenarios. Poisson regression where y_{it} is the dependent variable exhibits more power than linear regression where $y_{it}/x_{2,it}$ is the dependent variable when the degree of overdispersion is moderate but not when the degree is large. In fact, linear rate regression exhibits more power when overdispersion is large in the count outcome case (Panel A).

When y_{it} is the dependent variable, Poisson regression fits the data substantially better than linear regression, as indicated by the considerably smaller RMSEs. It is interesting to note that, while Poisson and linear regression of the rate $y_{it}/x_{2,it}$ have similar rejection rates, Poisson regression again generally fits the data better than OLS regression. The superior fit reflects the fact that, even after scaling by $x_{2,it}$ to compute a rate outcome, the outcome remains significantly skewed.

3 Analysis of Six Real-World Finance Data Sets

In this section, we analyze six data sets with two different count or count-like outcome variables, both of which exhibit masses at zero, that we replicate based on existing papers in top finance journals. One of the outcome variables is a count variable, while the other is continuous. We first illustrate the distributional properties of the outcome variables in these data sets. We then compare estimates from log1plus and Poisson regressions using each data set. To provide context for the magnitudes of the differences, we compare these differences to the effect of excluding control variables from the regressions. Finally, we explore the causes of differences between log1plus and Poisson regression estimates.

3.1 Replicated Data Sets

We replicate data sets from four papers in the large innovation literature analyzing factors driving the number of corporate patents granted to firms - those by Hirshleifer, Low, and Teoh (2012), He and Tian (2013), Fang, Tian, and Tice (2014), and Amore, Schneider, and Žaldokas (2013). These four papers collectively have 4,081 Google Scholar citations and 1,634 Web of Science citations as of the time of this writing. We also replicate data sets from two papers in the newer literature analyzing factors driving firms' volume of toxic releases - those by Akey and Appel (2021) and Xu and Kim (2022). We choose these six papers because they are influential and easy to replicate with publicly-available data sets.

The main patent data sets that finance researchers use are the NBER patent database, the HBS patent database, and the KPSS patent database. We use these sources to replicate the main data set in each of the four patent papers, following the data preparation outlined in each paper as best we can, including any adjustments for patent truncation (Dass, Nanda, and Xiao, 2017).[14] We use data from the EPA's Toxic Release Inventory (TRI) program to replicate the main data set in each of the two toxic release papers, following the data preparation outlined in the paper and the published replication packages. We begin by analyzing the distributions of the two outcome variables - number of patents granted and tons of toxic ground releases. Figure 3 presents histograms of firm-year observations of number of patents granted and establishment-year observations of tons of toxic ground releases. We top-code the data in both subfigures at 100 to make them easier to display. The figure shows that patent counts and toxic releases are both highly skewed and are 0 for 69% and 87.6% of observations, respectively.

[Insert Figure 3]

[14] See Lerner and Seru (2021) for an analysis of potential bias due to truncation methods when working with patent data.

3.2 Comparisons of Regressions Using Replicated Data Sets

For each of the six replicated data sets, we estimate one regression specification from the paper using the data set.[15] We choose the specification based on the ease of collecting all of the necessary control variables. We then estimate log-linear, log1plus, and Poisson regressions based on the chosen specification. In addition, we estimate a log1plus regression using a subsample where we exclude observations belonging to firms/establishments for which the outcome is zero in all years – the usable sample in Poisson regression.

Appendix Tables B1 through B6 present estimates from these regressions, with one table for each replicated data set. Each table reproduces the actual estimates from the paper as well as the estimates from the four regressions we estimate using the replicated data. The first explanatory variable listed in each data set is the explanatory variable of interest. Comparison of the estimates from each paper to our estimation of the same regression model shows that we are able to approximately replicate the results in all six papers.[16] For the sake of brevity, we summarize the main findings in Table 4. That table reports, for each paper, where the specification that we replicate is located in the paper, the outcome variable, the explanatory variable of interest, the type of regression model that the paper uses to estimate that specification, coefficients and standard errors for the explanatory variable of interest from Poisson and log1plus regressions using the replicated data, and the ratio of the Poisson and log1plus coefficients (if positive).

[Insert Table 4]

The log1plus and Poisson regression estimates for each replication exercise show substantial differences, suggesting that regression model choice has a first-order impact on the

[15] Of the six papers, four estimate log1plus regressions, one estimates a log-linear regression, and one estimates a Poisson regression for the specification we choose.

[16] The replication is exact for the paper by Akey and Appel (2021) (Table B5) since we use the data set from their replication package.

conclusions one would reach from the regression analysis. The log1plus and Poisson regression coefficients on the main variable of interest have opposite signs in three of the six replication exercises. In two of these three, the opposing coefficients on the log1plus and Poisson regression coefficients are of approximately the same magnitude. Of the three where the signs agree, the Poisson regression estimate is 239%, 309%, and 233% larger than the log1plus regression estimate.

3.3 Importance of regression model choice

To provide further insight into the importance of model choice when working with a count or count-like outcome variable, we compare the impact of model choice on the coefficient of interest to the impact of the choice of control variables to include in the regression using the specification analysis approach of Simonsohn et al. (2020).[17] Specifically, for each of the five replicated data sets that include control variables, we estimate a series of log1plus regressions covering every possible combination of control variables used in the replication specification, where each regression represents a different combination of controls. We repeat this exercise estimating Poisson regressions. So, if a data set contains n control variables, we generate a series of 2^n log1plus coefficients and a series of 2^n corresponding Poisson coefficients.

Each panel in Figure 4 plots the series of log1plus and the series of Poisson coefficient estimates in order from lowest to highest within each series for one replicated data set. Each point represents a single estimate of the coefficient of interest. The log1plus coefficients are depicted as blue diamonds, and the Poisson coefficients are depicted as grey diamonds. The red diamond in each series represents the coefficient from the regression including all of the control variables in the replicated specifications - i.e., the coefficients reported in the second or fourth column of Tables B1 through B6.

[Insert Figure 4]

[17] We thank Tony Cookson for suggesting this analysis.

Jumps within a series (log1plus or Poisson) tend to be small, suggesting that including or excluding any individual control variable has little effect on the coefficient of interest. However, there is virtually no overlap between the log1plus and Poisson coefficient series in any of the panels. Thus, it appears that model choice generally has a much larger impact on the coefficient of interest than even the most important control variable. We demonstrate this more formally by comparing the absolute average difference between the coefficients of interest from log1plus and Poisson regressions for all 2^n specifications to the absolute average difference between the coefficients of interest excluding each control variable one at a time. Table 5 presents the results.

[Insert Table 5]

Panel A shows the absolute average difference in the coefficient of interest between the log1plus and Poisson regressions for each replicated specification. Panels B and C show the absolute average effect of omitting each control variable on the log1plus and Poisson regression coefficient of interest, respectively. The table shows that changing from a log1plus to Poisson regression model changes indeed changes the coefficient of interest by a greater magnitude than excluding even the most important control variable in all five of the replication exercises that include control variables. In two cases, the former effect is an order of magnitude greater than the latter.

3.4 Explaining differences in log1plus and Poisson coefficients

The results summarized in Table 4 suggest that log1plus and Poisson regression estimates based on the same data set can differ substantially in both magnitude and sign. These estimates could differ for three reasons. The first is the addition of the constant in the log1plus regression. The second is the possibility that relationships between higher order model moments and covariates could bias the log1plus regression coefficients. The third is the

difference in the usable samples because of the necessary exclusion of firms/establishments where the outcome is zero in every period in Poisson regression. To shed light on the importance of each of these three possibilities, we estimate two auxiliary regressions using each of the six replicated data sets that allow us to disaggregate the differences into three parts.

To assess the effect of the addition of the constant in log1plus regression, we estimate a log1plus regression where the dependent variable is the fitted values of y from a Poisson regression, which we label $\hat{y}$, and where we restrict the sample to observations included in the Poisson regression. The difference between these estimates and the Poisson regression estimates captures the effect of changing the regression model, holding fixed the sample and filtering out potential bias in log1plus regression due to relationships between higher order error moments and covariates (by removing the noise completely).[18] To assess the effects of the Poisson sample restriction, we again substitute $\hat{y}$ for y and estimate a log1plus regression, this time without imposing the Poisson sample restriction. The difference between the estimates with and without the sample restriction captures the effect of the sample differences. Finally, the difference between log1plus estimates using the full sample where y is the dependent variable and where $\hat{y}$ is the dependent variable captures the effect of any relationships between higher order model error moments and covariates.

Table 6 reports the coefficient of interest from the Poisson and log1plus regressions from Table 4 in the first and fourth columns and as well as the two auxiliary regressions in the second and third columns, with each panel representing a different replicated data set. Comparing the first and second columns, the effect of adding the constant in log1plus regression appears to be large in four of the six replications (Panels A, B, C, and D), reversing the sign of the coefficient of interest in two (Panels B and C). Comparing the second and third columns,

[18]Note that log-linear regression (no constant added) using $\hat{y}$ as the dependent variable would produce coefficients identical to those from Poisson regression if there were no 0-valued observations.

the effect of sample differences due to the exclusion of observations for firms/establishments for which the outcome variable is zero in every year appears generally to be small. Comparing the third and fourth columns, the correlation behind higher order error moments and covariates appears to account for substantial differences in four of the six replications (Panels C, D, E, and F). Overall, then, both the addition of the constant in log1plus regression and the effects of correlations between higher order error moments and covariates appear to account for substantial differences between log1plus and Poisson regression estimates in these six real-world applications.

[Insert Table 6]

4 Conclusion

This paper highlights issues surrounding model choice when analyzing count-variable outcomes and other outcome variables inherently limited to non-negative values with skewed distributions, an increasingly common scenario in corporate finance. Our analysis suggests that researchers should rely primarily on Poisson regression. Poisson regression produces unbiased and consistent estimates under standard exogeneity conditions, admits separable fixed effects, and can now be estimated quickly, even with high-dimensional fixed effects. In contrast, commonly-used linear regressions where the dependent variable is the log of 1 plus the outcome produce estimates that have no economic meaning and can have the opposite sign of the true relationship being estimated, even absent sampling error. Our replications of data sets in six published papers modeling corporate patent counts and toxic release volumes suggest that regression model choice is a first-order decision when working with outcomes limited to non-negative values.

References

Akey, P. and I. Appel (2021). The limits of limited liability: Evidence from industrial pollution. *Journal of Finance 76*(1), 5–55.

Amore, M. D., C. Schneider, and A. Žaldokas (2013). Credit supply and corporate innovation. *Journal of Financial Economics 109*(3), 835–855.

Bellego, C., D. Benatia, and L.-D. Pape (2021). Dealing with logs and zeros in regression models. Working paper.

Bellstam, G., S. Bhagat, and J. A. Cookson (2021). A text-based analysis of corporate innovation. *Management Science 67*(7), 4004–4031.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of Applied Econometrics 1*(1), 29–53.

Cameron, A. C. and P. K. Trivedi (1996). 12 count data models for financial data. *Handbook of Statistics 14*, 363–391.

Cameron, A. C. and P. K. Trivedi (2013). *Regression analysis of count data*, Volume 53. Cambridge University Press.

Cohn, J. B., N. Nestoriak, and M. Wardlaw (2020). Private equity buyouts and workplace safety. *Review of Financial Studies 34*(10), 4832—-4875.

Cohn, J. B. and M. I. Wardlaw (2016). Financing constraints and workplace safety. *Journal of Finance 71*(5), 2017–2058.

Correia, S., P. Guimarães, and T. Zylkin (2020). Fast poisson estimation with high-dimensional fixed effects. *Stata Journal 20*(1), 95–115.

Correia, S., P. Guimarães, and T. Zylkin (2021). Verifying the existence of maximum likelihood estimates for generalized linear models. Working paper.

Dass, N., V. Nanda, and S. C. Xiao (2017). Truncation bias corrections in patent data: Implications for recent research on innovation. *Journal of Corporate Finance 44*, 353–374.

Fang, V. W., X. Tian, and S. Tice (2014). Does stock liquidity enhance or impede firm innovation? *Journal of Finance 69*(5), 2085–2125.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica 52*(3), 681–700.

He, J. J. and X. Tian (2013). The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics 109*(3), 856–878.

Hinz, J., A. Hudlet, and J. Wanner (2019). Separating the wheat from the chaff: Fast estimation of glms with high-dimensional fixed effects. Working Paper.

Hirshleifer, D., A. Low, and S. H. Teoh (2012). Are overconfident ceos better innovators? *Journal of Finance 67*(4), 1457–1498.

Karolyi, G. A. and A. G. Taboada (2015). Regulatory arbitrage and cross-border bank acquisitions. *Journal of Finance 70*(6), 2395–2450.

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 32*(3), 838–863.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics 95*(2), 391–413.

Lerner, J. and A. Seru (2021). The use and misuse of patent data: Issues for finance and beyond. *Review of Financial Studies Forthcoming*.

Manning, W. G. and J. Mullahy (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics 20*(4), 461–494.

Mullahy, J. (1997). Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics 79*(4), 586–593.

O'Hara, R. and J. Kotze (2010). Do not log-transform count data. *Nature Precedings*, 1–1.

Santos Silva, J. M. C. and S. Tenreyro (2006). The log of gravity. *Review of Economics and Statistics 88*(4), 641–658.

Santos Silva, J. M. C. and S. Tenreyro (2011). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters 112*(2), 220–222.

Simonsohn, U., J. P. Simmons, and L. D. Nelson (2020). Specification curve analysis. *Nature Human Behaviour 4*(11), 1208–1214.

Weber, J. E. and C. A. Hawkins (1971). The estimation of constant elasticities. *Southern Economic Journal*, 185–192.

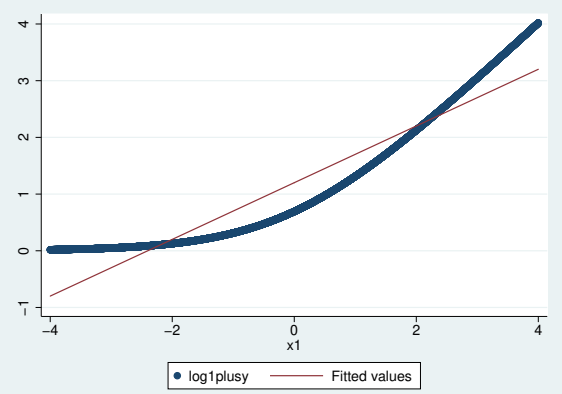
Windmeijer, F. A. G. and J. M. C. Santos Silva (1997). Endogeneity in count data models: An application to demand for health care. *Journal of Applied Econometrics 12*(3), 281–294.

Xu, Q. and T. Kim (2022). Financial constraints and corporate environmental policies. *Review of Financial Studies 35*(2), 576–635.

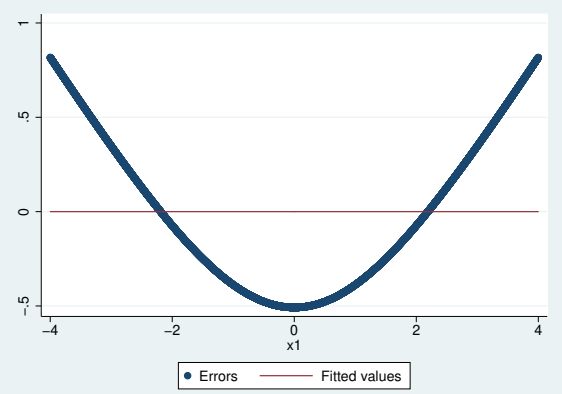
Figure 1: Bias due to nonlinear covariate relationships in log1plus regression

This figure presents an example in which $y = exp(\beta_1 x_1 + \beta_2 x_2)$, with $\beta_1 = 1$, $\beta_2 = 0$, and x_1 uniformly distributed on $[-4, 4]$. Each subfigure plots one variable against another and the associated regression line. Subfigure (a) plots $log(1+y)$ against x_1. The signed distance between the true relationship curve in blue and the regression line in red represents the error from a linear regression of $log(1+y)$ on x_1. We denote this error ϵ_{x1}. Subfigure (b) plots ϵ_{x1} against x_1. Subfigure (c) plots ϵ_{x1} against x_2, where x_2 is uniformly distributed on $[0, 1]$ and independent of x_1. Subfigure (d) plots ϵ_{x1} against x_2, where x_2 is linearly related to x_1 ($x_2 = x_1$). Subfigure (e) plots ϵ_{x1} against x_2, where $x_2 = x_1^2$.

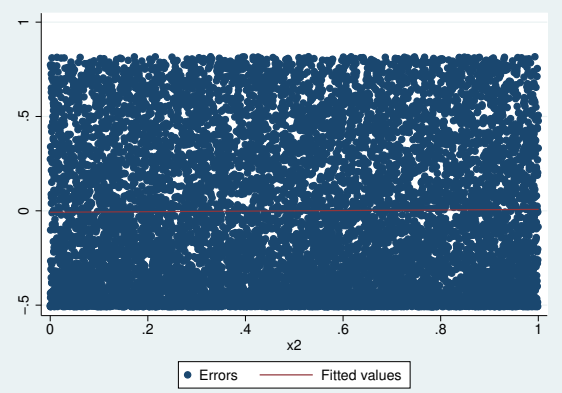
(a) $log(1+y)$ vs x_1



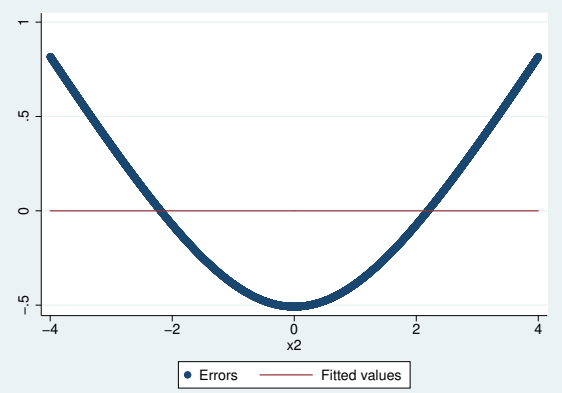
(b) ϵ_{x1} vs x_1



(c) ϵ_{x1} vs x_2, x_2 independent of x_1



(d) ϵ_{x1} vs x_2, $x_2 = x_1$



(e) ϵ_{x1} vs x_2, $x_2 = x_1^2$

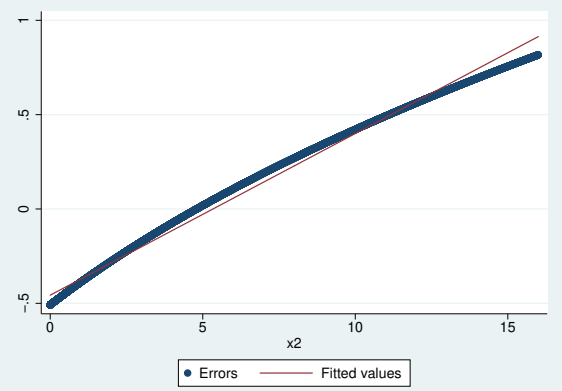
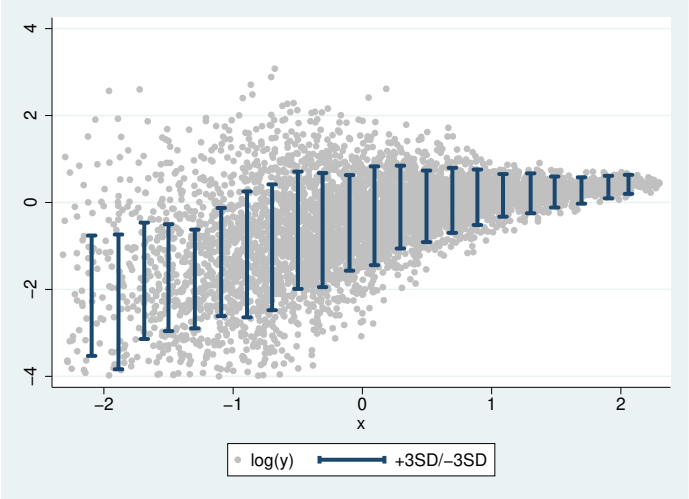


Figure 2: Dispersion and covariates: simulation vs real-world application

This figure presents scatter plots of simulated and real-world data. Subfigure (a) plots $log(y)$ against x for simulated data with 5,000 observations, where $y = e^{\mathbf{x}\beta}\eta$, with x drawn from a standard normal distribution truncated at the 1st and 99th percentiles and error η drawn from a lognormal distribution with mean 1 and variance $exp(-x)$. Subfigure (b) plots the residuals from a linear regression of the log of patent count on control variables for the specification in Table 2 column (1) of Fang et al. (2014) against $ILLIQ$, the covariate of interest in that specification, using replicated data. We normalize $ILLIQ$ by centering and scaling by the sample standard deviation to make it directly comparable to the normally distributed x variable in the simulated data depicted in Subfigure (a). Range plots are overlaid for bins of width 0.2 showing a range of plus or minus 3 standard deviations around the mean.

(a) Dispersion of log(y) with respect to x in simulation where $\sigma_\nu = exp(-x)$



(b) Dispersion of log-linear residuals with respect to the variable of interest in Fang et al. (2014)

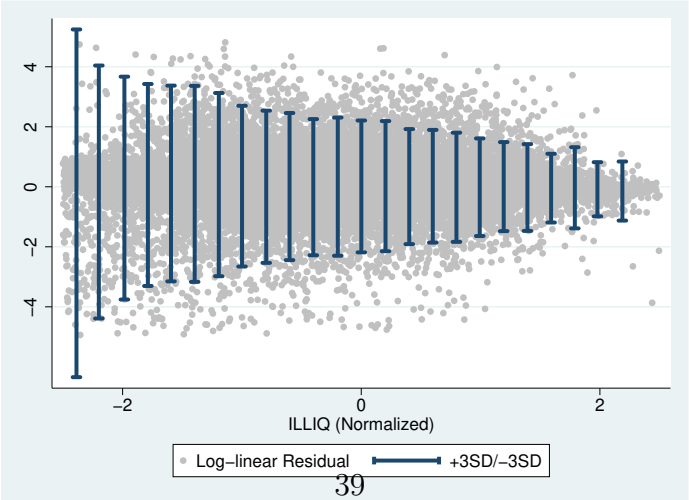
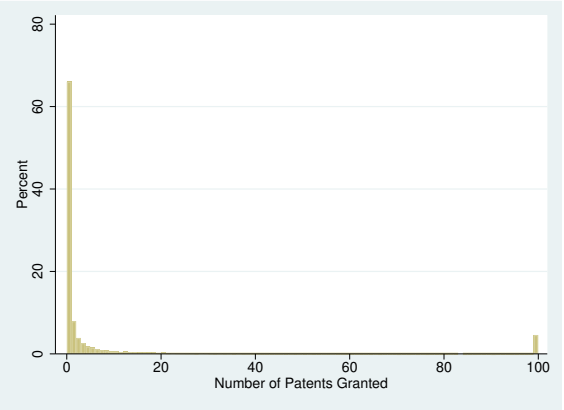


Figure 3: Histograms of common count datasets

This figure presents histograms for (a) the number of patents granted in a firm-year and (b) tons of pollutants in a facility-year, both for Compustat firms. Each bar in the histogram has a width of 1. We top-code each variable at 100 to make the figure easier to read. Hence, the left-most bar represents the percent of observations with 0 patents or tons of pollutants, and the right-most bar represents the percent of observations with more than 99 patents or tons of pollutants.

(a) Patent data histogram



(b) Toxic release data histogram

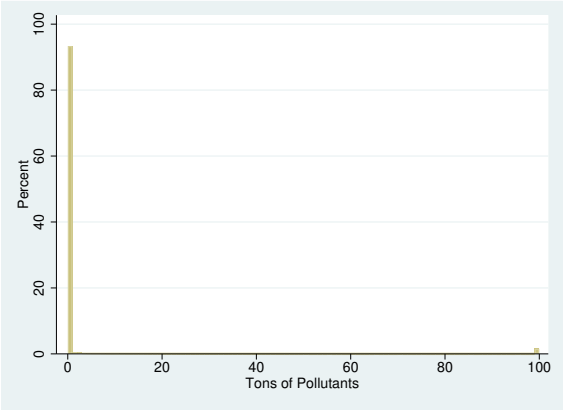
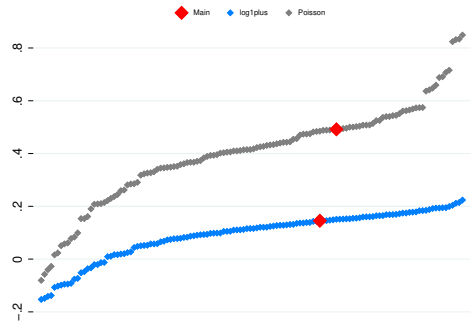


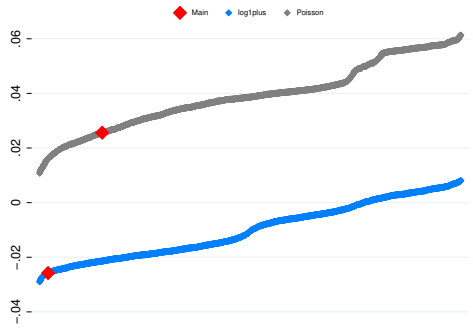
Figure 4: Specification Curves

This figure implements the specification curve analysis of Simonsohn et al. (2020) to examine the degree to which log1plus and Poisson regressions yield different coefficient estimates. For each replicated data set, we estimate a set of log1plus regressions covering every possible combination of control variables used in the replicated specification, where each regression represents a different combination of controls. We repeat this exercise estimating Poisson regressions. In each subfigure, log1plus estimates are plotted in blue, Poisson estimates are plotted in grey, and estimates from the specifications where all controls are included are plotted in red.

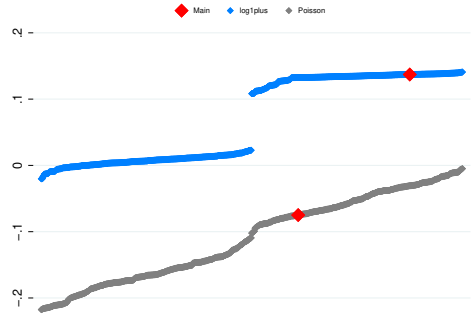
(a) Hirshleifer, Low, and Toeh (2012)



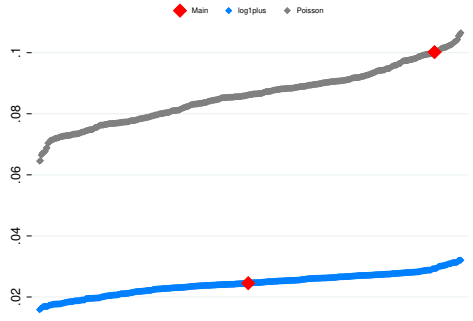
(b) He and Tian (2013)



(c) Fang, Tian, Tice (2014)



(d) Amore, Schneider, and Žladokas (2013)



(e) Xu and Kim (2021)

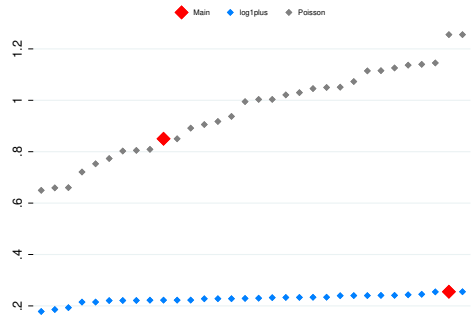


Table 1: Heteroskedasticity simulation

This table presents results from regressions based on simulated data sets in which we introduce various forms of heteroskedasticity. Each simulation involves 10,000 simulated data sets of 5,000 observations (x, y) each, with $y = exp(\beta x)\eta$, where η is a lognormally distributed error with mean 1 and standard deviation $\sigma_\eta(x)$ and $\beta = 0.2$. Panel A presents the mean coefficient and standard error for Poisson, log-linear, log1plus, log0.1plus, log10plus and Inverse Hyperbolic Sine (IHS) regressions of y on x plus a constant in two scenarios - one where x is continuous variable drawn from a standard normal distribution truncated at the 1st and 99th percentiles and one where x is a binary variable drawn from a Bernoulli distribution with $p = .5$. The table displays three cases for each of these two scenarios involving different values of $\sigma_\eta(x)$. Panel B presents results using simulated data sets where we introduce a panel structure. We simulate this data as a balanced panel of 500 individuals (i) and 10 time units (t). The variable $x_{1,it}$ is composed of a fixed part μ_i and a time-varying part ν_{it} such that $x_{it} = 0.5\mu_i + 0.5\nu_{it}$. The standard deviation of the error is given by $\sigma_\eta = \exp(\gamma\mu_1 + (1-\gamma)\nu_{it})$, where γ is a parameter reflecting the importance of the fixed component in the dispersion of the error.

Panel A: Comparisons of different estimators with heteroskedastic model errors

$\sigma_\eta =$	Continuous x					
	$\exp(x)$		$\exp(1/2)$		$\exp(-x)$	
	Coef(x)	Bias	Coef(x)	Bias	Coef(x)	Bias
Poisson	0.199	−0.5%	0.200	0.0%	0.201	0.5%
Log-linear	−0.300	−250.0%	0.200	0.0%	0.700	250.0%
Log1plus	−0.017	−108.5%	0.075	−62.5%	0.171	−14.5%
Log0.1plus	−0.164	−182.0%	0.158	−21.0%	0.473	136.5%
Log10plus	0.008	−96.0%	0.016	−92.0%	0.023	−88.5%
IHS	−0.020	−110.0%	0.098	−51.0%	0.229	14.5%
$\sigma_\eta =$	Binary x					
	1 if $x = 0$ 2 if $x = 1$		1.5		2 if $x = 0$ 1 if $x = 1$	
	Coef(x)	Bias	Coef(x)	Bias	Coef(x)	Bias
Poisson	0.200	0.0%	0.200	0.0%	0.201	0.5%
Log-linear	−0.258	−229.0%	0.200	0.0%	0.659	229.5%
Log1plus	−0.015	−107.5%	0.078	−61.0%	0.178	−11.0%
Log0.1plus	−0.153	−176.5%	0.161	−19.5%	0.481	140.5%
Log10plus	0.010	−95.0%	0.017	−91.5%	0.025	−87.5%
IHS	−0.021	−110.5%	0.103	−48.5%	0.237	18.5%

Panel B: Heteroskedasticity in model errors and fixed effects

Estimator:	Poisson		Poisson FE		Log-linear		Log-linear FE	
	Coef(x)	Bias	Coef(x)	Bias	Coef(x)	Bias	Coef(x)	Bias
$\gamma = 100\%$	0.199	−0.5%	0.200	0.0%	−0.299	−249.5%	0.200	0.0%
$\gamma = 75\%$	0.199	−0.5%	0.199	−0.5%	−0.300	−250.0%	−0.050	−125.0%
$\gamma = 50\%$	0.199	−0.5%	0.198	−1.0%	−0.300	−250.0%	−0.300	−250.0%
$\gamma = 25\%$	0.199	−0.5%	0.199	−0.5%	−0.300	−250.0%	−0.550	−375.0%
$\gamma = 0\%$	0.199	−0.5%	0.197	−1.5%	−0.301	−250.5%	−0.800	−500.0%

Table 2: Constant added simulation

This table presents results from regressions estimated on a simulated data set of 5,000 observations, where each observation takes the form (x_1, x_2, y), with $y = kexp(\beta_1 x_1 + \beta_2 x_2)$. We set $\beta_1 = 2$ and $\beta_2 = -0.2$. For each observation, we draw the value of x_1 from a standard normal distribution. For the analysis reported in Panel A, we draw the value of x_2 from an independent standard normal distribution. For the analysis reported in Panel B, we set $x_2 = 0.5x_1 + 0.5z$, where z is drawn from an independent standard normal distribution. For the analysis reported in Panel C, we set $x_2 = \max\{x_1, 0\}$. In each panel, we set $k = 1$ in the left columns and $k = 10$ in the right columns. In each case, we report coefficients and the percentage bias in each coefficient from Poisson, log-linear, log1plus, log0.1plus, log10plus, and Inverse Hyperbolic Sine (IHS) regressions of y on x_1 and x_2 plus a constant.

Panel A: x_1 and x_2 independent

	k = 1				k = 10			
	Coef(x_1)	Bias(x_1)	Coef(x_2)	Bias(x_2)	Coef(x_1)	Bias(x_1)	Coef(x_2)	Bias(x_2)
Poisson	2.000	0.0%	-0.200	0.0%	2.000	0.0%	-0.200	0.0%
Log-linear	2.000	0.0%	-0.200	0.0%	2.000	0.0%	-0.200	0.0%
Log1plus	1.003	-49.9%	-0.117	-41.7%	1.614	-19.3%	-0.169	-15.5%
Log0.1plus	1.614	-19.3%	-0.169	-15.5%	1.912	-4.4%	-0.192	-3.9%
Log10plus	0.388	-80.6%	-0.051	-74.3%	1.003	-49.9%	-0.117	-41.7%
IHS	1.222	-38.9%	-0.139	-30.5%	1.798	-10.1%	-0.183	-8.4%

Panel B: $x_2 = 0.5 * x_1 + 0.5 * z$

	k = 1				k = 10			
	Coef(x_1)	Bias(x_1)	Coef(x_2)	Bias(x_2)	Coef(x_1)	Bias(x_1)	Coef(x_2)	Bias(x_2)
Poisson	2.000	0.0%	-0.200	0.0%	2.000	0.0%	-0.200	0.0%
Log-linear	2.000	0.0%	-0.200	0.0%	2.000	0.0%	-0.200	0.0%
Log1plus	1.016	-49.2%	-0.128	-35.8%	1.634	-18.3%	-0.174	-12.8%
Log0.1plus	1.634	-18.3%	-0.174	-12.8%	1.921	-4.0%	-0.193	-3.3%
Log10plus	0.384	-80.8%	-0.060	-70.2%	1.016	-49.2%	-0.128	-35.8%
IHS	1.243	-37.9%	-0.151	-24.3%	1.818	-9.1%	-0.186	-7.0%

Panel C: $z_2 = \max\{x_1, 0\}$

	k = 1				k = 10			
	Coef(x_1)	Bias(x_1)	Coef(x_2)	Bias(x_2)	Coef(x_1)	Bias(x_1)	Coef(x_2)	Bias(x_2)
Poisson	2.000	0.0%	-0.200	-0.0%	2.000	0.0%	-0.200	-0.0%
Log-linear	2.000	0.0%	-0.200	0.0%	2.000	0.0%	-0.200	0.0%
Log1plus	0.302	-84.9%	1.209	-704.5%	1.167	-41.7%	0.693	-446.4%
Log0.1plus	1.167	-41.7%	0.693	-446.4%	1.785	-10.8%	0.054	-127.0%
Log10plus	-0.039	-102.0%	0.721	-460.7%	0.302	-84.9%	1.209	-704.5%
IHS	0.435	-78.2%	1.368	-783.8%	1.486	-25.7%	0.423	-311.3%

Table 3: Regression rejection rates and root mean square errors

This table presents results from a series of simulations that compare the rejection rates and root means square errors of linear and Poisson regressions. We simulate a set of observations $(x_{1,it}, x_{2,it}, y)$, where $E[y|x] = exp(\beta_1 x_1 + x_2)$. The data is simulated as a balanced panel of 500 groups, i, and 10 time units, t. The variable $x_{1,it}$ is composed of a fixed part μ_i and a time-varying part ν_{it} such that $x_{1,it} = 0.5\mu_i + 0.5\nu_{it}$, where both μ_i and ν_{it} are drawn from a standard normal distribution truncated at the 1st and 99th percentiles. The variable x_2 is drawn from a normal distribution with a mean of 0 and a standard deviation of 2 and is independent of x_1. We vary β_1 in each set of simulations to be -2, -0.2, 0.2, and 2.

Panel A simulates discrete outcomes by using a negative binomial data generating process, with the overdispersion parameter, α_{NB}, set to 0.001 (negligible overdispersion), 0.5 (medium overdispersion), and 2 (high overdispersion). Panel B simulates continuous outcomes using the mixture model of Santos Silva and Tenreyro (2011). In this formulation, y_{it} is the sum of m_i random variables z_{it}, where m_i is a negative binomial-distributed random variable with mean $exp(\beta_1 x_{1,it} + x_{2,it})$, and z_{it} is a $\chi^2_{(1)}$-distributed random variable. We set the variance of m_i to $E[m_i|\mathbf{x}] + bE[m_i|\mathbf{x}]^2$, which implies that $Var(y_{it}|x_{it}) = 3 * E[y_{it}|x_{it}] + b * E[y_{it}|x_{it}]^2$, where b is a parameter that determines the conditional variance and hence degree of overdispersion in the data.

Panel A: Discrete

Overdispersion	Low				Medium				High			
β	-2	-0.2	0.2	2	-2	-0.2	0.2	2	-2	-0.2	0.2	2
Rejection Rate:												
Linear (y)	0.981	0.279	0.274	0.984	0.992	0.208	0.212	0.992	0.992	0.140	0.137	0.993
Poisson (y)	1.000	1.000	1.000	1.000	1.000	0.878	0.876	1.000	1.000	0.436	0.429	1.000
Linear (rate)	1.000	0.999	0.999	1.000	1.000	0.748	0.750	1.000	1.000	0.649	0.644	1.000
Poisson (rate)	1.000	0.999	0.999	1.000	1.000	0.750	0.754	1.000	1.000	0.652	0.649	1.000
Average Root Mean Square Error:												
Linear (y)	164.522	42.169	42.360	162.373	154.996	49.126	49.616	154.364	153.171	61.871	61.351	153.610
Poisson (y)	0.223	0.345	0.345	0.222	0.561	0.636	0.635	0.568	1.107	1.101	1.101	1.104
Linear (rate)	3.973	1.268	1.268	3.977	5.809	2.613	2.596	5.835	8.162	2.884	2.911	8.136
Poisson (rate)	0.868	1.281	1.281	0.868	1.248	1.620	1.619	1.248	1.602	1.892	1.893	1.602

Panel B: Continuous

Overdispersion	Low				Medium				High			
β	-2	-0.2	0.2	2	-2	-0.2	0.2	2	-2	-0.2	0.2	2
Rejection Rate:												
Linear (y)	0.982	0.281	0.269	0.982	0.973	0.210	0.207	0.977	0.959	0.150	0.150	0.955
Poisson (y)	1.000	1.000	1.000	1.000	1.000	0.851	0.847	1.000	1.000	0.470	0.466	1.000
Linear (rate)	1.000	0.523	0.518	1.000	1.000	0.506	0.502	1.000	1.000	0.453	0.452	1.000
Poisson (rate)	1.000	0.535	0.529	1.000	1.000	0.520	0.515	1.000	1.000	0.470	0.470	1.000
Average Root Mean Square Error:												
Linear (y)	163.427	42.313	42.500	167.192	185.063	49.813	49.876	186.745	222.000	63.031	63.609	228.433
Poisson (y)	0.400	0.605	0.605	0.398	0.596	0.777	0.776	0.596	0.867	1.032	1.030	0.864
Linear (rate)	6.740	3.824	3.770	6.649	7.532	3.941	3.863	7.493	9.493	4.076	4.033	9.397
Poisson (rate)	1.334	1.781	1.780	1.332	1.448	1.856	1.854	1.448	1.652	1.994	1.993	1.649

Table 4: Replicated regression specifications

This table details the regression specifications that we analyze using replicated data sets from six papers and the outcomes from Poisson and log1plus regression models. The first column reports the paper. The second column indicates the specification that we analyze from the paper. The third column indicates the outcome variable. The fourth column indicates the explanatory variable of interest. The fifth column indicates the model that the paper estimates. The sixth and seventh columns report our estimates of the coefficients and standard errors for the explanatory variable of interest from Poisson and log1plus regressions, respectively. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test. The eighth column reports the ratio of the coefficients from the Poisson and log1plus regressions. A value of "Neg" indicates that the ratio is negative (i.e., the coefficients have opposite signs).

Paper	Specification	Outcome	Expl variable of interest	Model in paper	Our estimates Poisson	Log1plus	Ratio
Hirshleifer et al. (2012)	Table V col (3)	Patents	Overconfident CEO (options)	log1plus	0.492*** (0.183)	0.145** (0.065)	3.39
He and Tian (2013)	Table 2 col (4)	Patents	lnCoverage	Log1plus	0.026 (0.031)	-0.026*** (0.010)	Neg
Fang et al. (2014)	Table 2 col (1)	Patents	ILLIQ	Log1plus	-0.075 (0.057)	0.137*** (0.020)	Neg
Amore et al. (2013)	Table 3 col (4)	Patents	Interstate deregulation	Poisson	0.1002** (0.0401)	0.0245 (0.0241)	4.09
Akey and Appel (2021)	Table 3 col (1)	Toxic releases	Bestfoods	Log1plus	-0.050 (0.138)	0.047*** (0.014)	Neg
Xu and Kim (2022)	Table 2 col (4)	Toxic releases	HM Debt	Log-linear	0.850** (0.351)	0.255 (0.175)	3.33

Table 5: Importance of model choice vs control variables

This table compares differences in coefficients of interest between log1plus and Poisson regression estimates to changes in coefficients of interest when different control variables are omitted. Specifically, for each of the five data sets that we replicate that include control variables, we estimate a series of log1plus regressions covering every possible combination of control variables used in the replication specification, where each regression represents a different combination of controls. We repeat this exercise estimating Poisson regressions. Panel A reports the average absolute difference between the log1plus and Poisson regression coefficients of interest across all regressions. Panel B reports the absolute average difference in log1plus regression coefficients of interest for specifications that include and exclude the specified control variable. Panel C reports the absolute average difference in Poisson regression coefficients of interest for specifications that include and exclude the specified control variable.

Panel A: Absolute average coefficient difference between log1plus and Poisson

Hirshleifer, Low, and Toeh (2012)	He and Tian (2013)	Fang, Tian, and Tice (2014)	Amore, Schneider, and Žaldokas (2013)	Xu and Kim (2021)
0.310	0.049	0.179	0.062	0.726

Panel B: Absolute average log1plus coefficient difference including or excluding each control variable

Hirshleifer, Low, and Toeh (2012)		He and Tian (2013)		Fang, Tian, and Tice (2014)		Amore, Schneider, and Žaldokas (2013)		Xu and Kim (2021)	
Excluded	|diff|	Excluded	|diff|	Excluded	|diff|	Excluded	|diff|	Excluded	|diff|
Log(1+delta)	0.125	lnAssets	0.018	$LNMV_t$	0.126	Ln(R&D)	0.004	CAPEX/PPE	0.016
Log(sales)	0.078	lnAge	0.006	$CAPTEXTA_t$	0.006	Ind. Trend	0.004	Tangible	0.014
Log(1+tenure)	0.034	Leverage	0.003	$KZINDEX_t$	0.005	HIndex	0.002	Cash/Assets	0.007
Log(PPE/Emp)	0.031	ROA	0.002	LEV_t	0.004	ROA	0.002	Log(assets)	0.006
Log(1+vega)	0.017	HIndex	0.002	$PPETA_t$	0.004	Ln(K/L)	0.001	Tobin Q	0.000
Stock return	0.007	PPEAssets	0.001	Q_t	0.003	Ln(sales)	0.001		
Inst. holdings	0.005	HIndex^2	0.001	ROA_t	0.002	Cash	0.001		
		CapexAssets	0.001	$RDTA_t$	0.002	Tangibility	0.000		
		Tobin Q	0.000	$LNAGE_t$	0.001	Ln(age)	0.000		
		KZIndex	0.000	$HINDEX_t^2$	0.000				
		RDAssets	0.000						

Panel C: Absolute average Poisson coefficient difference including or excluding each control variable

Hirshleifer, Low, and Toeh (2012)		He and Tian (2013)		Fang, Tian, and Tice (2014)		Amore, Schneider, and Žaldokas (2013)		Xu and Kim (2021)	
Excluded	|diff|	Excluded	|diff|	Excluded	|diff|	Excluded	|diff|	Excluded	|diff|
Log(1+delta)	0.249	lnAssets	0.016	$LNMV_t$	0.115	Ln(K/L)	0.010	Log(assets)	0.261
Log(1+tenure)	0.097	lnAge	0.015	$KZINDEX_t$	0.042	Ln(sales)	0.006	Tangible	0.109
Log(sales)	0.075	HIndex	0.004	$LNAGE_t$	0.018	HIndex	0.004	CAPEX/PPE	0.037
Log(PPE/Emp)	0.066	PPEAssets	0.004	$CAPTEXTA_t$	0.006	Ln(age)	0.003	Cash/Assets	0.028
Log(1+vega)	0.005	HIndex^2	0.003	Q_t	0.005	Ln(R&D)	0.002	Tobin Q	0.002
Stock return	0.004	Leverage	0.002	LEV_t	0.003	ROA	0.002		
Inst. holdings	0.020	ROA	0.002	$PPETA_t$	0.002	Ind. Trend	0.001		
		Tobin Q	0.002	ROA_t	0.002	Cash	0.001		
		KZIndex	0.001	$HINDEX_t^2$	0.001	Tangibility	0.001		
		CapexAssets	0.001	$RDTA_t$	0.000				
		RDAssets	0.000						

Table 6: Poisson and Log1plus Reconciliation

This table decomposes the differences in the Poisson and log1plus regression estimates reported in Tables B1 through B6. Each of Panels A through D provides the decomposition for one paper. The first column reproduces the Poisson estimates from the corresponding table. The second column presents estimates from log1plus regression, where we replace the dependent variable y with the fitted value of y from the Poisson regression in the first column and limit the sample to the sample used for the Poisson regression. The third column presents estimates from the same log1plus regression as the second column, but using the full sample. The fourth column reproduces the log1plus regression (using the actual value of y) from the corresponding table. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

Panel A: Hirshleifer, Low, and Toeh (2012)

Model	Poisson	Log1plus ($\hat{y}$)	Log1plus ($\hat{y}$)	Log1plus
Sample	Poisson	Poisson	Full	Full
Overconfident CEO	0.492	0.263	0.261	0.145
Observations	6,326	6,326	6,482	6,482
Controls, FEs	Yes	Yes	Yes	Yes

Panel B: He and Tian (2013)

Model	Poisson	Log1plus ($\hat{y}$)	Log1plus ($\hat{y}$)	Log1plus
Sample	Poisson	Poisson	Full	Full
lnCoverage	0.026	-0.014	-0.025	-0.026
Observations	15,857	15,857	27,058	27,064
Controls, FEs	Yes	Yes	Yes	Yes

Panel C: Fang, Tian, and Tice (2014)

Model	Poisson	Log1plus ($\hat{y}$)	Log1plus ($\hat{y}$)	Log1plus
Sample	Poisson	Poisson	Full	Full
$ILLIQ_t$	-0.075	0.028	0.024	0.137
Observations	15,970	15,970	39,000	39,000
Controls, FEs	Yes	Yes	Yes	Yes

Panel D: Amore, Schneider, and Žaldokas (2013)

Model	Poisson	Log1plus ($\hat{y}$)	Log1plus ($\hat{y}$)	Log1plus
Sample	Poisson	Poisson	Full	Full
Interstate dereg	0.100	0.050	0.047	0.025
Observations	14,920	14,920	18,424	18,424
Controls, FEs	Yes	Yes	Yes	Yes

Panel E: Akey and Appel (2021)

Model	Poisson	Log1plus ($\hat{y}$)	Log1plus ($\hat{y}$)	Log1plus
Sample	Poisson	Poisson	Full	Full
Bestfoods	-0.050	-0.041	-0.047	0.047
Observations	182,454	182,454	501,259	501,259
Controls, FEs	Yes	Yes	Yes	Yes

Panel F: Xu and Kim (2021)

Model	Poisson	Log1plus ($\hat{y}$)	Log1plus ($\hat{y}$)	Log1plus
Sample	Poisson	Poisson	Full	Full
HM Debt	0.850	0.835	0.835	0.255
Observations	38,365	38,365	39,951	39,951
Controls, FEs	Yes	Yes	Yes	Yes

Appendices

A Bias in log-linear regression due to heteroskedasticity in example

Suppose that $y = \eta exp(\beta x)$, where x is normally distributed with mean 0 and variance σ_x^2, and η is log-normally distributed with mean 1 and standard deviation $\sigma_\eta(x) = exp(\delta x)$ for constant δ. Then, $log(E[y|x]) = \beta x + log(E[\eta|x]) = \beta x + log(E[exp(\epsilon)|x])$, where ϵ is normally distributed with mean $\mu_\epsilon(x)$ and standard deviation $\sigma_\epsilon(x)$. In this case, $log(E[y|x]) = \beta x + \frac{1}{2}\sigma_\epsilon^2(x)$. Assuming that η is log-normally distributed with mean 1 and standard deviation $\sigma_\eta(x) = exp(2\delta x)$ is equivalent to assuming $\mu_\epsilon(x) = log(\frac{1}{\sqrt{1+exp(2\delta x)}})$ and $\sigma_\epsilon^2(x) = log(1 + exp(2\delta x))$. Thus, we have:

$$log(E[y|x]) = \beta x + \frac{1}{2} log(1 + exp(2\delta x)).$$

Taking the derivative with respect to x, we have:

$$\frac{dlog(E[y|x])}{dx} = \frac{dE[y|x]}{dx}\frac{1}{E[y|x]} = \beta + \frac{\delta exp(2\delta x)}{1 + exp(2\delta x)},$$

or, equivalently:

$$\beta = \frac{dE[y|x]}{dx}\frac{1}{E[y|x]} - \frac{\delta exp(2\delta x)}{1 + exp(2\delta x)}.$$

Let $f(x)$ denote the (normal) probability density function of x. Integrating over x and denoting the expectation over x by E_x, we have

$$\begin{aligned}
\beta &= E_x\left[\frac{dE[y|x]}{dx}\frac{1}{E[y|x]}\right] - \int_{-\infty}^{\infty}\frac{\delta exp(2\delta x)}{1+exp(2\delta x)}f(x)dx \\
&= E_x\left[\frac{dE[y|x]}{dx}\frac{1}{E[y|x]}\right] - \delta\int_{-\infty}^{\infty}\left[g(x)+\frac{1}{2}\right]f(x)dx \\
&= E_x\left[\frac{dE[y|x]}{dx}\frac{1}{E[y|x]}\right] - \delta\int_{-\infty}^{\infty}g(x)f(x)dx - \frac{1}{2}\delta\int_{-\infty}^{\infty}f(x)dx \\
&= E_x\left[\frac{dE[y|x]}{dx}\frac{1}{E[y|x]}\right] - \delta\int_{-\infty}^{\infty}g(x)f(x)dx - \frac{1}{2}\delta, \qquad (14)
\end{aligned}$$

where $g(x) = \frac{exp(2\delta x)}{1+exp(2\delta x)} - \frac{1}{2}$. Observe that

$$\begin{aligned}
g(-x) &= \frac{exp(-2\delta x)}{1+exp(-2\delta x)} - \frac{1}{2} = \frac{1}{1+exp(2\delta x)} - \frac{1}{2} \\
&= \left(\frac{1}{1+exp(2\delta x)} - 1\right) + \frac{1}{2} = -\frac{exp(2\delta x)}{1+exp(2\delta x)} + \frac{1}{2} = -g(x),
\end{aligned}$$

so $g(x)$ is odd. As a result, the second expression in (14) equals zero, and (14) simplifies to:

$$\beta = E_x\left[\frac{dE[y|x]}{dx}\frac{1}{E[y|x]}\right] - \frac{1}{2}\delta. \qquad (15)$$

The log-linear regression model is $log(y) = \beta x + log(\eta)$, which yields the relationship $E[log(y)|x] = \beta x$. The objective in estimating a log-linear regression of y on x is to recover an estimate of the semi-elasticity of y with respect to x, which is the first term on the right-hand side of (15). Thus, the log-linear regression coefficient β is biased by $-\frac{1}{2}\delta$ due to heteroskedasticity in η.

B Full Replication Tables

Table B1: Replication: Hirshleifer, Low, and Teoh (2012)

This table presents a series of regressions based on the regression specification in Table V column (3) of Hirshleifer, Low, and Teoh (2012). The unit of observation is a firm-year. The outcome variable is the number of patents a firm generates in a given year. The first column reproduces the results from the original paper, which estimates a log1plus regression. The next three columns present results from log1plus, log-linear, and Poisson regressions, based on our replication of the original data set. The final column presents results from log1plus regression where the sample is restricted to the sample usable in Poisson regression. Standard errors clustered at the firm level are presented below each coefficient. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Original	Replication	Replication	Replication	Replication
	Log1plus	Log1plus	Log-Linear	Poisson	Log1plus Poisson Sample
Overconfident CEO	0.111**	0.145**	0.288***	0.492***	0.155**
	(0.053)	(0.065)	(0.091)	(0.183)	(0.066)
Log(sales)	0.640***	0.402***	0.518***	0.818***	0.406***
	(0.051)	(0.038)	(0.042)	(0.072)	(0.038)
Log(PPE/Emp)	0.218***	0.211***	0.321***	0.688***	0.218***
	(0.051)	(0.053)	(0.083)	(0.175)	(0.054)
Stock return	0.052***	0.085***	0.060	0.175***	0.086***
	(0.012)	(0.023)	(0.037)	(0.058)	(0.023)
Institutional holdings	-0.113***	-0.503***	-0.467**	-0.855**	-0.518***
	(0.030)	(0.177)	(0.237)	(0.362)	(0.179)
Log(1+tenure)	-0.051**	-0.058*	-0.135***	0.044	-0.062*
	(0.025)	(0.035)	(0.046)	(0.080)	(0.035)
Log(1+delta)	0.014	0.034	0.110**	-0.075	0.034
	(0.036)	(0.031)	(0.044)	(0.092)	(0.031)
Log(1+vega)	0.218***	0.161***	0.208***	0.284***	0.165***
	(0.039)	(0.034)	(0.046)	(0.092)	(0.034)
Observations	8,939	6,482	3,121	6,326	6,326
Adjusted R2	0.507	0.483	0.516		0.479

Table B2: Replication of He and Tian (2013)

This table presents a series of regressions based on the regression specification in Table 2 column (4) of He and Tian (2013). The unit of observation is a firm-year. The outcome variable is the number of patents a firm generates in a given year. The first column reproduces the results from the original paper, which estimates a log1plus regression. The next three columns present results from log1plus, log-linear, and Poisson regressions, based on our replication of the original data set. The final column presents results from log1plus regression where the sample is restricted to the sample usable in Poisson regression. T-statistics based on standard errors clustered at the firm level are presented below each coefficient. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Original Log1plus	Replication Log1plus	Replication Log-Linear	Replication Poisson	Replication Log1plus Poisson Sample
lnCoverage	-0.053***	-0.026***	0.036*	0.026	0.000
	(0.016)	(0.010)	(0.020)	(0.031)	(0.013)
lnAssets	0.050**	0.079***	0.107***	0.093	0.086***
	(0.020)	(0.022)	(0.039)	(0.062)	(0.026)
RDAssets	0.100**	0.405***	0.305	0.246	0.217
	(0.048)	(0.128)	(0.204)	(0.462)	(0.154)
lnAge	0.180**	0.352***	0.057	-0.215*	0.090*
	(0.072)	(0.046)	(0.070)	(0.111)	(0.050)
ROA	0.693***	0.239***	0.035	0.204	0.170**
	(0.200)	(0.059)	(0.112)	(0.276)	(0.076)
PPEAssets	0.330***	0.455***	0.790***	0.901**	0.437***
	(0.105)	(0.135)	(0.244)	(0.358)	(0.159)
Leverage	-0.324***	-0.346***	-0.294**	-0.369**	-0.329***
	(0.067)	(0.069)	(0.119)	(0.179)	(0.082)
CapexAssets	-0.051	0.063	-0.221	-0.115	-0.037
	(0.113)	(0.171)	(0.325)	(0.487)	(0.224)
TobinQ	0.019***	0.029***	0.012	0.009	0.021***
	(0.005)	(0.005)	(0.007)	(0.010)	(0.005)
KZIndex	-0.001**	-0.001	-0.001	-0.002	-0.000
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)
HIndex	0.226	0.504	-0.241	-1.786**	0.451
	(0.163)	(0.318)	(0.507)	(0.768)	(0.357)
$HIndex^2$	-0.128	-0.132	0.423	1.659**	-0.051
	(0.139)	(0.264)	(0.448)	(0.774)	(0.307)
Observations	25,860	27,064	8,263	15,857	15,857
R2	0.833	0.730	0.869		0.790

Table B3: Replication: Fang, Tian, and Tice (2014)

This table presents a series of regressions based on the regression specification in Table 2 column (1) of Fang, Tian, and Tice (2014). The unit of observation is a firm-year. The outcome variable is the number of patents a firm generates in a given year. The first column reproduces the results from the original paper, which estimates a log1plus regression. The next three columns present results from log1plus, log-linear, and Poisson regressions, based on our replication of the original data set. The final column presents results from log1plus regression where the sample is restricted to the sample usable in Poisson regression. Standard errors clustered at the firm level are presented below each coefficient. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Original	Replication	Replication	Replication	Replication
	Log1plus	Log1plus	Log-Linear	Poisson	Log1plus Poisson Sample
$ILLIQ_t$	0.141***	0.137***	0.014	-0.075	0.220***
	(0.020)	(0.020)	(0.071)	(0.057)	(0.045)
$LNMV_t$	0.160***	0.149***	0.343***	0.165***	0.315***
	(0.018)	(0.017)	(0.057)	(0.054)	(0.037)
$RDTA_t$	0.283***	0.316***	0.560**	0.948***	0.240
	(0.089)	(0.091)	(0.236)	(0.345)	(0.151)
ROA_t	-0.032	0.033	-0.266*	-0.563*	-0.130
	(0.068)	(0.028)	(0.158)	(0.307)	(0.093)
$PPETA_t$	0.287***	0.052*	0.130	-0.072	0.093
	(0.094)	(0.031)	(0.195)	(0.246)	(0.117)
LEV_t	-0.256***	-0.226***	0.064	0.399	-0.337**
	(0.075)	(0.065)	(0.214)	(0.281)	(0.149)
$CAPTEXTA_t$	0.175	0.235***	0.600	0.396	0.584*
	(0.119)	(0.085)	(0.520)	(0.574)	(0.316)
$HINDEX_t$	0.106	0.098	0.082	-0.300	0.097
	(0.086)	(0.083)	(0.281)	(0.418)	(0.184)
$HINDEX_t^2$	-0.112	-0.094	0.191	0.589	0.032
	(0.150)	(0.141)	(0.477)	(0.873)	(0.313)
Q_t	-0.006	0.001	-0.027***	-0.013	-0.015**
	(0.007)	(0.003)	(0.008)	(0.009)	(0.006)
$KZINDEX_t$	-0.000*	0.001*	0.000	0.004	0.002
	(0.000)	(0.000)	(0.008)	(0.011)	(0.005)
$LNAGE_t$	0.168***	0.267***	0.252*	0.438**	0.285***
	(0.035)	(0.050)	(0.151)	(0.209)	(0.108)
Observations	39,469	39,000	8,205	15,970	15,970
Adjusted R2	0.839	0.809	0.817		0.783

Table B4: Replication: Amore, Schneider, and Žaldokas (2013)

This table presents a series of regressions based on the regression specification in Table 3 column (4) of Amore, Schneider, and Žaldokas (2013). The unit of observation is a firm-year. The outcome variable is the number of patents a firm generates in a given year. The first column reproduces the results from the original paper, which estimates a log-linear regression. The next three columns present results from log1plus, log-linear, and Poisson regressions, based on our attempt replication of the original data set. The final column presents results from log1plus regression where the sample is restricted to the sample usable in Poisson regression. Standard errors clustered at the firm level are presented below each coefficient. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Original	Replication	Replication	Replication	Replication
	Poisson	Log1plus	Log-Linear	Poisson	Log1plus Poisson Sample
Interstate dereg	0.1188***	0.0245	0.0749	0.1002**	0.0289
	(0.0397)	(0.0241)	(0.0515)	(0.0401)	(0.0274)
Ln (sales)	0.5360***	0.1615***	0.3271***	0.6741***	0.1946***
	(0.0901)	(0.0234)	(0.0558)	(0.0845)	(0.0283)
Ln (K/L)	0.1969**	0.0148	0.0403	0.2734***	0.0089
	(0.0789)	(0.0211)	(0.0369)	(0.0900)	(0.0301)
Ln (R&D stock)	0.3264***	0.0918***	0.1289***	0.2124***	0.1082***
	(0.1196)	(0.0164)	(0.0326)	(0.0584)	(0.0211)
Industry trends	Yes	Yes	Yes	Yes	Yes
Additional Controls	Yes	Yes	Yes	Yes	Yes
Observations	18,066	18,424	9,040	14,920	14,920
R2		0.877	0.867		0.862

Table B5: Replication: Akey and Appel (2021)

This table presents a series of regressions based on the regression specification in Table 3 column (1) of Akey and Appel (2021). The unit of observation is a chemical-facility-firm-year. The outcome variable is the pounds of ground pollutants a facility releases in a given year. We use their replication kit to generate the data. The first column reproduces the results from the original paper, which estimates a log1plus regression. The next three columns present results from log1plus, log-linear, and Poisson regressions, based on the data provided by the authors in a replication kit. The final column presents results from log1plus regression where the sample is restricted to the sample usable in Poisson regression. Standard errors clustered at the circuit level are presented below each coefficient. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Original	Replication	Replication	Replication	Replication
	Log1plus	Log1plus	Log-Linear	Poisson	Log1plus Poisson Sample
Bestfoods	0.047***	0.047***	0.119	-0.050	0.118**
	(0.014)	(0.014)	(0.078)	(0.138)	(0.046)
Observations	501,259	501,259	61,510	182,454	182,454
Adjusted R2	0.541	0.541	0.570		0.448

Table B6: Replication: Xu and Kim (2021)

This table presents a series of regressions based on the regression specification in Table 2 column (4) of Xu and Kim (2022). The unit of observation is a facility-year. The outcome variable is the amount of pollution a facility generates in a given year in tons. The first column reproduces the results from the original paper, which estimates a log-linear regression. The next three columns present results from log1plus, log-linear, and Poisson regressions, based on our replication of the original data set. The final column presents results from log1plus regression where the sample is restricted to the sample usable in Poisson regression. Standard errors clustered at the firm level are presented below each coefficient. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Original	Replication	Replication	Replication	Replication
	Log-Linear	Log1plus	Log-Linear	Poisson	Log1plus Poisson Sample
HM Debt	0.654*	0.255	0.533	0.850**	0.267
	(0.360)	(0.175)	(0.344)	(0.351)	(0.179)
Log(assets)	0.039	0.057*	0.028	-0.143*	0.058*
	(0.039)	(0.032)	(0.061)	(0.083)	(0.032)
Cash/Assets	0.194	0.002	0.013	0.085	0.003
	(0.296)	(0.022)	(0.031)	(0.256)	(0.022)
CAPEX/PPE	0.008	0.051	0.083	-0.550*	0.051
	(0.130)	(0.051)	(0.082)	(0.304)	(0.052)
Tangible	0.012	-0.123	-0.186	0.492	-0.118
	(0.369)	(0.075)	(0.122)	(0.343)	(0.077)
Tobin Q	0.082**	0.020	0.017	0.109	0.021
	(0.032)	(0.017)	(0.029)	(0.116)	(0.018)
Observations	36,562	39,951	35,835	38,365	38,365
R2	0.860	0.883	0.864		0.879