Count Data in Finance*

Jonathan Cohn
University of Texas-Austin

Zack Liu
University of Houston

Malcolm Wardlaw
University of Georgia

August 2021

Abstract

This paper assesses different econometric approaches to working with count-based (and other zero-bounded, right-skewed) outcome variables, which are increasingly common in corporate finance applications. We demonstrate that the common practice of adding a constant to the outcome and then estimating log-linear regressions produces estimates with no natural interpretation that can easily have the wrong sign. 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 that the conclusions of several existing papers change substantially if we estimate Poisson regressions rather than linear regressions on log-transformed data.

*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. We would like to thank Jason Abrevaya, Andres Almazan, Aydoğan Altı, Sergio Correia, John Griffin, Travis Johnson, Sam Krueger, Aaron Pancost, James Scott, Sheridan Titman, Jeff Wooldridge, and participants in the Virtual Finance Seminar and seminar at the University of Texas at Austin for valuable feedback.

Count Data in Finance

Abstract

This paper assesses different econometric approaches to working with count-based (and other zero-bounded, right-skewed) outcome variables, which are increasingly common in corporate finance applications. We demonstrate that the common practice of adding a constant to the outcome and then estimating log-linear regressions produces estimates with no natural interpretation that can easily have the wrong sign. 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 that the conclusions of several existing papers change substantially if we estimate Poisson regressions rather than linear regressions on log-transformed data.

A growing number of papers in empirical corporate finance study outcome variables that are bounded below by zero and highly right-skewed. Many of the leading applications involve count outcomes such as the number of corporate patents a firm is granted in a year. A significant challenge in working with such outcomes is that skewness can make linear estimators inefficient. Researchers employ a variety of econometric approaches in an effort to overcome this challenge. However, little guidance is available on the utility of these different estimators, and the statistical properties of the most common approach, which we call "log1plus-linear" regression, are largely unknown. This paper uses a mix of econometric analysis, simulation, and replication to shed light on the statistical properties of estimators based on commonly-used approaches and provides guidance for future research. While we focus primarily on count-based outcomes, which are discrete, the insights from our analysis also apply to continuous outcome variables that are zero-bounded and right-skewed.

One approach to working with a skewed outcome variable in general is to ignore the skewness and estimate a simple linear regression using ordinary least squares (OLS). OLS produces the best linear unbiased estimator (BLUE) irrespective of the distribution of the errors in the dependent variable. However, a linear approximation of a model producing typical count outcomes is likely to result in severely skewed and heteroskedastic regression errors, reducing efficiency and making it difficult to compute appropriate confidence intervals. A common approach to addressing these problems is to estimate log-linear regressions. Log-transforming the outcome variable implicitly recasts the underlying regression model as a constant elasticity model, with coefficients interpretable as semi-elasticity estimates. A constant elasticity model will fit count-based outcomes better than a linear model would in most circumstances, resulting in more efficient estimates.

Unfortunately, log-linear regression estimates are generally biased and inconsistent if the standard deviation of the error in the untransformed outcome variable does not scale proportionately with the conditional mean of the untransformed outcome - i.e., if the multiplicative

error in the underlying constant elasticity model is heteroskedastic (Santos Silva and Tenreyro, 2006). The bias arises from a Jensen's inequality problem: A semi-elasticity represents the change in the *log of the conditional mean* of an outcome with respect to a covariate, while log-linear regression estimates the *conditional mean of the log*. The two are only equal if the variance of the multiplicative error is independent of the conditional mean - i.e., if the multiplicative error is homoskedastic. Intuitively, a nonlinear transformation of an outcome variable translates a dependence of the *variance* of the *untransformed* regression error on a covariate into a dependence of the *mean* of the *transformed* error on the covariate. We show that the direction of the bias depends on the sign of the relationship between the variance of the error and the covariate. We also show that inclusion of fixed effects in log-linear regressions can either mitigate or exacerbate bias due to heteroskedasticity, depending on the strength of any fixed effect in variance of the error.

An additional difficulty in implementing log-linear regression is that the logarithm of 0 is undefined, and commonly-analyzed count data sets in finance often contain many zero values. For example, 69% of COMPUSTAT firms generate 0 patents in a given year. Observations with 0-valued outcomes are necessarily excluded from the usable data set when estimating log-linear regressions, which could make estimates difficult to interpret. A common solution to this problem is to add a constant - typically 1 - to the outcome variable before log transforming it. This two-step transformation deskews the outcome variable without the need to eliminate zero-valued observations. This log1plus-linear regression approach is the most commonly used approach in finance when working with count-based outcome variables. For example, 25 of 44 papers published in "top three" finance journals between 2011 and 2020 in which corporate patent count is the primary outcome variable use log1plus-linear regression as their primary econometric approach, and 23 use it exclusively. Yet, little is known about the properties of estimates based on log1plus-linear regression and whether these properties allow for a reasonable and accurate test of underlying economic hypotheses.

We describe two serious problems with log1plus-linear regression. First, the estimates have no meaningful economic interpretation. They represent estimates of the semi-elasticities of the outcome variable plus 1. One might imagine that these semi-elasticities are equivalent to semi-elasticities of the outcome – the only plausible objects of interest when estimating a log1plus-linear regression – since the constant is invariant. However, because of another Jensen's inequality problem, they are not equivalent, nor can semi-elasticities of the outcome be recovered from log1plus-linear regression coefficients. Second, we show that even if one were interested in the semi-elasticities of the outcome variable plus 1 for some reason, log1plus-linear regression coefficients are generally biased and inconsistent estimates of these semi-elasticities. We show in simulations that log1plus-linear regressions can lead to incorrect inference in expectation not only about the magnitude of a relationship, but also its sign. That is, even the direction of a relationship cannot reliably be inferred from log1plus-linear regression estimates.[1]

One alternative to linear or transformed-linear models when dealing with a zero-bounded, right-skewed outcome variable is to estimate a Poisson regression. As with log-linear regression, Poisson regression coefficients are conveniently interpretable as semi-elasticity estimates (Wooldridge, 2010, p. 726). However, unlike in the case of log-linear regression, no assumption about the relationship between the higher order moments of the model error and covariates is necessary for Poisson regression to be unbiased. The Poisson model does impose the *restriction* that the conditional variance of the outcome be proportional to the outcome's conditional mean. Departure from conditional variance-mean equality in the data reduces estimation efficiency. However, it does *not* induce bias or inconsistency. In addition, the Poisson model admits separable group fixed effects, including high-dimensional fixed effects - effectively a prerequisite for use in corporate finance applications. While the Poisson

[1]These issues also apply to OLS regressions where the dependent variable is the inverse hyperbolic sine transformation of the outcome, another approach to addressing skewnness and the presence of zero valued-outcomes that is occasionally used in the literature.

regression model formally specifies a count variable for the outcome, Poisson Pseudo Maximum Likelihood (PPML) estimation also yields valid estimates when the outcome variable is continuous.[2]

Other count regression models such as the negative binomial model or zero-inflated models relax the conditional mean-variance restriction of the Poisson model and may produce more efficient estimates under certain conditions. Negative binomial regression allows for overdispersion, where the conditional variance of the outcome exceeds the conditional mean, though it imposes a specific functional form on this overdispersion. Zero-inflated models (e.g., zero-inflated Poisson and zero-inflated negative binomial) estimate intensive and extensive margins separately, allowing for excessive zero values in the distribution of the outcome. A major drawback of all of these alternative count models is that they do not admit separable fixed effects, limiting their usefulness in most corporate finance applications. While one can, in principle, include group dummy variables as covariates when estimating one of these models, the lack of separability can result in biased estimates due to the incidental parameters problem (Lancaster, 2000).[3] The same limitation applies to the Tobit model, which models the outcome variable as censored rather than as coming from a conditional distribution limited to non-negative values.

Finally, if a suitable scaling variable exists, an additional alternative approach is to estimate a linear regression of the outcome rate (e.g., workplace injuries per full-time equivalent employee). Scaling often substantially deskews a count-based outcome variable, as skewness

[2]While computational constraints may have been a practical issue for estimating fixed effects Poisson models in the past, recent advances in graph theory and computational matrix algebra has produced fast, efficient algorithms for implementing PPML models with multiple fixed-effects. Correia et al. (2020) demonstrates the implementation of a fast algorithm for Stata that allows for speedy convergence of even high-dimensional fixed effects Poisson regressions. Hinz et al. (2019) provides a similar algorithm for estimating a Poisson fixed-effects regression in R.

[3]STATA module xtnbreg allows for estimation of a "fixed effects" negative binomial model. However, the fixed effects in this model allow the conditional variance rather than the conditional mean to vary across groups. This model therefore does not address the standard concern in corporate finance regarding unobserved group-level heterogeneity in conditional means.

in the distribution of outcomes is often a product of skewness in the distribution of scale. Linear rate regression coefficients have a simple interpretation - the expected unit change in the rate associated with a one unit change in the explanatory variable. We find in simulations that rate regression may be more efficient than Poisson regression when the outcome exhibits substantial overdispersion. Unfortunately, however, a suitable scaling variable is not available in most settings.

We supplement our simulation analysis with analysis of data sets replicated from existing papers, which allow us to assess the importance of differences in estimates based on different approaches in real-world applications. We replicate data sets from 4 papers analyzing factors affecting corporate patent counts – the leading application involving count-based outcomes in corporate finance. In addition, we replicate data sets from 2 papers analyzing factors affecting an establishment's toxic releases as reported in pounds, an approximately continuous, zero-bounded, right-skewed outcome variable studied in several recent papers. 5 of the 6 of papers estimate log1plus-linear regression models. While we are able to effectively replicate the main result in all 6 replication exercises, we find that Poisson regression estimates differ substantially from log1plus-linear regression estimates in magnitude in all 6 and even differ in sign in 3 of the 6. Further analysis suggests that these differences are attributable to both the addition of the constant in log1plus-linear regression and bias due to heteroskedastic errors.

Our analysis suggests the following takeaways for working with count outcomes and continuous, zero-bounded, right-skewed outcome variables. Simple linear regression is viable if the degree of skewness is limited and outliers are unlikely. Traditional log-linear regression is viable if zeroes are uncommon and if the researcher has strong *a priori* reason to believe that the multiplicative homoskedasticity assumption is satisfied. Log1plus-linear regression is more problematic, since the estimates it produces have no meaningful interpretation and can easily have the wrong sign in expectation. Poisson regression represents a more ro-

bust approach, produces estimates with simple and meaningful interpretations, and admits separable group fixed effects, making it suitable to the practical needs of corporate finance researchers. Finally, linear rate regression also produces valid, easy-to-interpret estimates and can be more efficient than Poisson regression, but it requires a suitable scaling variable, which does not exist in many applications.

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 regressions models commonly used when outcomes take the form of zero-bounded, right-skewed data such as count data. As an illustration of the distributional features of such data, Figure 1 presents a histogram of firm-year observations of number of patents granted, a commonly used data set in corporate finance research. We top-code the data at 100 to make the figure easier to display. The figure shows that patent counts are highly right-skewed and are 0 in 69% of firm-years.

[Insert Figure 1]

1.1 Linear regression

The simplest approach when working with zero-bounded, right-skewed outcomes is to ignore the nature of the distribution and estimate OLS regressions of the form:

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

where ϵ is assumed to be mean zero. After all, estimates from such regressions have a simple interpretation and are unbiased and consistent under the assumption that ϵ is uncorrelated with each of the covariates in $\mathbf{x}$. However, the linear regression (1) may be a poor choice as a regression model for a count-based y variable because of two features of the implied regression error. Specifically, the error distribution is likely to be both highly right skewed, as expected values of the outcome cannot be negative, and heteroskedastic, as there is more scope for variation in an additive mean-zero error when the conditinal mean of y is large than when it is small.

These features of the implied distribution of the error cause three practical problems. First, both the sharp departure from normality and the presence of heteroskedasticity can substantially reduce the precision of estimates. Second, large outliers due to data errors may be more difficult to detect when the error distribution is skewed and heteroskedastic. Positive outliers are observationally equivalent to large positive linear regression error realizations, and the fat right tail of a right-skewed distribution and large variance when the conditional expectation of y is high make such realizations common. Third, a skewed error makes appropriate confidence intervals difficult to determine.

1.2 Log-linear regression

One common solution to these problems is to deskew y prior to estimation by log-transforming it. Because the log transformation is concave, this transformation generally produces a variable with a distribution that is both more symmetric and less heteroskedastic. With a log-transformed outcome variable, the regression equation becomes:

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

We refer to the OLS estimation of (2) as "log-linear regression."

Log-transformation of the outcome variable implicitly recasts the model as a constant elasticity model. This constant elasticity structure is actually a desirable feature of the log-linear model when working with a zero-bounded, right-skewed outcome variable since it more closely approximates the underlying data generating process. For example, a policy that increases the innovative activity of all firms uniformly is more likely to cause a proportionate rather than additive increase in the number of patents a firm receives. As a result, a constant elasticity model is likely to better fit the data, allowing for more precise estimates. In addition, the log-linear regression model produces estimates that are interpretable as semi-elasticities, which may be more informative about how a covariate affects the outcome than estimates of average linear effects from (1).

The log-linear model has two important drawbacks - one general and one specific to zero-bounded outcomes such as counts that take a value of 0 for a substantial faction of observations. The general drawback is that log-linear regressions can produce biased and inconsistent estimates in the presence of heteroskedasticity (Santos Silva and Tenreyro, 2006). To see why this is the case, observe that a log-linear regression assumes constant elasticity - i.e., that $E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}}$. Adding an error term yields the regression equation:

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

Santos Silva and Tenreyro (2006) show that log-linear regressions only recover consistent estimates of $\boldsymbol{\beta}$ if ϵ can be expressed as $\epsilon = e^{\mathbf{x}\boldsymbol{\beta}}\nu$, with ν independent of $\mathbf{x}$. To see why, observe that we can rewrite the regression equation as

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

where $\eta = 1 + \frac{\epsilon}{e^{\mathbf{x}\boldsymbol{\beta}}}$ is a multiplicative error. Log-transforming (4) yields

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

Consistent estimation of the slope coefficients of this log-linearized regression equation requires that $E[log(\eta)|\mathbf{x}]$ be orthogonal to $\mathbf{x}$. Suppose that ϵ - the error in the natural regression model (3) - is independent of $\mathbf{x}$. Then, $E[log(\eta)|\mathbf{x}] = E[log(1 + \frac{\epsilon}{e^{\mathbf{x}\boldsymbol{\beta}}})]$. It is impossible to integrate ϵ out of this expression if ϵ is independent of $\mathbf{x}$ because of the nonlinearity of the *log* function. Thus, orthogonality of the error in the natural underlying model to $\mathbf{x}$ does not 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 $E[log(\eta)|\mathbf{x}]$ is independent of $\mathbf{x}$. Recall that $e^{\mathbf{x}\boldsymbol{\beta}}$ is the conditional mean of y. So, consistent estimation of $\boldsymbol{\beta}$ requires that the standard deviation of the additive error in the natural model scale with the conditional mean of y. Any other relationship between the standard deviation of ϵ and $\mathbf{x}$ will cause the regression coefficients to be biased and inconsistent estimates of $\boldsymbol{\beta}$.

Observe that the assumption required for consistent estimation is equivalent to assuming that the multiplicative error η in (4) is homoskedastic. Any heteroskedasticity in η will then cause the estimates from log-linear regression to be biased and inconsistent. Since $\epsilon = e^{\mathbf{x}\boldsymbol{\beta}}\nu$ and $E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}}$, homoskedasticity of η is equivalent to the standard deviation of the additive error ϵ scaling proportionately with the conditional mean of the outcome. For example, if Firm A averages 10 patents per year with a standard deviation of 2 and Firm B averages 100 patents per year, then homoskedasticity implies the standard deviation of Firm B's patents will be 20. If this condition is not satisfied, then log-linear regression will produce biased and inconsistent estimates.

While multiplicative homoskedasticity is a natural benchmark in a constant elasticity model, it need not hold in practice. As we show later in simulations, the bias resulting from heteroskedasticity can be large and can even cause estimates to have the wrong sign. Fully characterizing the direction of the bias is difficult since the relationship between the variance of the multiplicative error and the covariates may be non-monotonic. However, we are able to characterize the direction of the bias in the case where the relationship is monotonic. We do so by borrowing the concept of second-order stochastic domination 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}, x_{j2}) of values of x_j satisfying $x_{j2} > x_{j1}$, that $E[\eta|x_{j1}] = E[\eta|x_{j2}]$ but that the conditional mean of η is the same when $x_j = x_{j1}$ and $x_j = x_{j2}$ but the variance of η is lower when $x_j = x_{j2}$ 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, $\frac{\partial E[log(\eta)|\mathbf{x}]}{x_j} > 0$ 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 $\hat{\beta}_j$ 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.

We illustrate the direction of the bias graphically with a simple example. Suppose that $y = e^{\alpha+\beta x}\eta$. Further, suppose that $\alpha = 0$ and $\beta = 1$ so that $y = e^x\eta$. In addition, suppose that $\eta \in \{\eta^-, \eta^+\}$, with $prob(\eta = \eta^-) = prob(\epsilon = \eta^+) = 0.5$, and that the range of x is $(0, 1)$. We consider two cases. In the first case, we assume that $\eta^- = 1 - x$ and $\eta^+ = 1 + x$. In the second case, we assume that $\eta^- = x$ and $\eta^+ = 2 - x$. In both cases, $E[\eta|x] = 1$. In

the first case, $var(\eta) = x^2$, so $corr(var(\eta), x) > 0$. In the second case, $var(\eta) = (1-x)^2$, so $corr(var(\eta), x) < 0$. Figure 2 plots $log(E[y|x])$ and $E[log(y)|x]$ for the two cases.

[Insert Figure 2]

The relationship between $log(E[y|x])$ and x is linear with a slope of $\beta = 1$ – the true semi-elasticity – in both cases since $log(E[y|x]) = log(e^{\alpha+x\beta}) = \alpha + x\beta$. However, a log-linear regression estimates the relationship between $E[log(y)|x]$ and x, and $E[log(y)|x] \neq log(E[y|x])$ by Jensen's inequality. In the first case, where $corr(var(\eta), x) > 0$, the slope of the relationship between $E[log(y)|x]$ and x is less than $\beta = 1$ for all values of x and is actually negative (i.e., has the wrong sign) for $x > 0.6180$. Log-linear regression using a sample of (x, y) values with the properties described here could easily produce a negative slope coefficient in this case, even absent sampling error, despite the fact that $E[y|x]$ increases with x. In the second case, where $corr(var(\eta), x) < 0$, the slope of the relationship between $E[log(y)|x]$ and x is greater than $\beta = 1$ for all values of x. Log-linear regression in this cases will overestimate the semi-elasticity of y with respect to x in expectation.

The second drawback of a log-linear regression model, which is more specific to applications involving zero-bounded outcome variables which take a value of 0 for a large number of observations, is that the logarithm of 0 is undefined. This limitation is of practical importance in many count data sets. For example, approximately 69% of Compustat firms are granted zero patents in a given year. Estimating log-linear regressions requires excluding observations with zero-valued outcomes. The resulting sample shrinkage raises concerns not only about efficiency but also about generality, since it allows for estimation of only the intensive margin.

1.3 Log1plus-linear regression

The most common approach in finance to addressing the 0-value problem is to add an arbitrary constant – typically 1 – to y before log-transforming. Doing so ensures that the transformed dependent variable is defined for all possible values of y, including 0. When the constant is 1, the resulting regression equation is:

$$log(1+y) = \mathbf{x}\boldsymbol{\beta}^{1+} + \epsilon^{1+}. \tag{7}$$

We refer to OLS estimation of (7) as log1plus-linear regression. We focus on the case where the constant added before log-transformation is 1 since this is the most common case. However, any positive constant will allow for a log transformation while preserving observations with zero-valued outcomes. We refer to the more general form of OLS regression where the dependent variable is $log(c+y)$ for constant $c > 0$ as logcplus-linear regression. Despite its routine use in finance, the interpretation and econometric properties of log1plus-linear regression are not well understood.

There are two significant problems with log1plus-linear regressions. The first is that the coefficients in log1plus-linear regression do not represent estimates of the semi-elasticities of y or any other quantity typically of interest. The regression coefficient in the log-linear regression (5) estimates the semi-elasticity of y with respect to x. The regression coefficient in the log1plus-linear regression (7) estimates the semi-elasticity of $1+y$ with respect to covariate x_j. It might be tempting to assume that these two semi-elasticities are the same since the constant added to y is invariant to $\mathbf{x}$. However, this assumption overlooks a Jensen's inequality problem.

Formally, the $j$the regression coefficient in (7) estimates the semi-elasticity:

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

The problem is that $\frac{1}{1+E[y|\mathbf{x}]} \neq \frac{1}{E[y|\mathbf{x}]}$, so $\beta_j^{1+} \neq \beta_j$. The relationship between the semi-elasticities of $1+y$ and y is:

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

In principle, knowledge of $E[y|\mathbf{x}]$ would allow one to infer $\boldsymbol{\beta}$ directly from $\boldsymbol{\beta}^{1+}$. Unfortunately, $E[y|\mathbf{x}]$ is not observable. Indeed, the entire objective of the empirical exercise is to characterize $E[y|\mathbf{x}]$.

The second problem with log1plus-linear regression is that, even if one wanted to estimate an empirical relationship between $1+y$ and $\mathbf{x}$, log1plus-linear regression is likely to produce biased estimates of the semi-elasticity of $1+y$ with respect to $\mathbf{x}$, for two reasons. First, if the true model is (3), then $log(1+y)$ will be nonlinearly related to $\mathbf{x}$. One might assume that this nonlinearity is not a major concern, since the regression coefficients will still estimate the best linear relationship between $log(1+y)$ and $\mathbf{x}$. This argument is valid in a univariate regression. However, it is not generally valid in a multivariate regression. If y is nonlinearly related to a covariate that is in turn nonlinearly related to another covariate, then the coefficients on both x variables will be biased. This issue is not unique to log1plus-linear regression. Nonlinear relationships can cause biased estimates of average effects in any multivariate linear regression, and we speculate that this issue occurs frequently. However, the problem is endemic in the context of log1plus-linear regression. A standard constant elasticity model of y, the closest plausible standard model to a log1plus-linear regression model, produces a linear relationship between $log(y)$ and $\mathbf{x}$ but induces a nonlinear relationship between $log(1+y)$ and $\mathbf{x}$.

To provide further intuition for why log1plus-linear regression yields biased estimates of the relationship between $log(1+y)$ and $\mathbf{x}$, consider the following simple example, on which we build further in simulations in Section 2. Let $y = e^{x_1 - 0.1x_2}$. In this constant elasticity model, $log(y)$ is linearly related to x_1 and x_2. Since the relationship between $log(y)$ and

$log(1+y)$ is nonlinear, the relationships between $log(1+y)$ and x_1 and x_2 are also nonlinear. Figure 3 plots $log(y)$ and $log(1+y)$ against x_1.

[Insert Figure 3]

Because of the nonlinear relationship between $log(1+y)$ and x_1, log1plus-linear regression will provide only a linear approximation of the true relationship between $log(1+y)$ and x_1. If x_1 and x_2 are nonlinearly related to each other, then x_2 will be correlated with the residual of $log(1+y)$ after partialling out the effect of x_1. As a result, the coefficient on x_2 in a log1plus-linear regression will be a biased estimate of the true average relationship. The converse is true as well: The coefficient on x_1 in a log1plus-linear regression will pick up part of the relationship between y and x_2. We demonstrate in Section 2 that the bias can be large enough that regression coefficients may have the wrong sign, even absent sampling error. Note that the nonlinearity that Figure 3 illustrates is sharpest when y is small. The bias due to nonlinearity, then, is likely to be most problematic in data sets in which the value of y is frequently zero or near zero. That is, the problem is largest in precisely the settings where researchers are most likely to use the log1plus-linear regression approach.

The second reason that log-linear regression may produce biased estimates is that the heteroskedasticity problem associated with log-linear regression that we described in Section 1.2 is even more vexing in the context of a log1plus-linear regression. Suppose that (3) is the true model – that is, $y = e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon$. Adding 1 to both sides yields $1 + y = e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon + 1$, or $1 + y = e^{\mathbf{x}\boldsymbol{\beta}} + \epsilon^{1+}$, where $\epsilon^{1+} = \epsilon + 1$. Writing this relationship in its multiplicative form, $1 + y = e^{\mathbf{x}\boldsymbol{\beta}}\eta^{1+}$, we have $\eta^{1+} = 1 + \frac{\epsilon^{1+}}{e^{\mathbf{x}\boldsymbol{\beta}}} = 1 + \frac{\epsilon}{e^{\mathbf{x}\boldsymbol{\beta}}} + \frac{1}{e^{\mathbf{x}\boldsymbol{\beta}}}$. It is immediately apparent 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 $\mathbf{x}$ no longer makes $E[log(\eta^{1+})|\mathbf{x}]$ 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-linear model. Instead, what is required for consistent estimation is a particular form of heteroskedasticity for which there is likely to be no theoretical justification.

1.4 Inverse hyperbolic sine regression

An alternative to log-transforming the outcome that is occasionally used is to transform the outcome variable using the inverse hyperbolic sine function. While the exact transformation is different, the rationale is the same as it is for log-transforming the outcome variable. One advantage of the inverse hyperbolic sine transformation is that it is defined for $y = 0$, so addition of an arbitrary constant is unnecessary. However, as with log-linear regression, if the true model is multiplicative, then consistency depends on the relationship between higher moments of the true error and the covariates. Unlike the case of log-linear regression, the exact requirement for consistent estimation is unclear. Moreover, coefficients from inverse hyperbolic sine regression do not have economically meaningful interpretations.

1.5 Poisson regression

We next consider Poisson regression as an alternative when working with a zero-bounded, right-skewed outcome variable. Poisson regression represents a generalized linear model and assumes that the dependent variable has a Poisson distribution that depends on covariates, with density $f(y|\mathbf{x}) = e^{-\mu(\mathbf{x})}\mu(\mathbf{x})^y/y!$, where $\mu(\mathbf{x}) = e^{\mathbf{x}\boldsymbol{\beta}}$. The domain of y in the Poisson model is the counting numbers (0, 1, 2, ...). The conditional expectation in the Poisson model takes the form:

$$E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}} \tag{10}$$

or, equivalently, $log(E[y|\mathbf{x}]) = \mathbf{x}\boldsymbol{\beta}$. As with log-linear regression, Poisson coefficients represent semi-elasticity estimates. Unlike log-linear regression, Poisson regression produces

estimates that also have simple interpretations in terms of partial effects:

$$\frac{\partial E[y|\mathbf{x}]}{\partial x_j} = \beta_j e^{\mathbf{x}\boldsymbol{\beta}}. \tag{11}$$

The estimated average partial effect is simply $\bar{y}\beta_j$, where $\bar{y}$ is the mean value of y for the sample.

Unlike log-linear regression, consistent estimation of Poisson regression does not require homoskedasticity of the multiplicative error. The difference between the two regression models is that Poisson regression estimates $log(E[y|\mathbf{x}])$, while log-linear regression estimates $E[log(y)|\mathbf{x}]$. The variance of the error does not affect $log(E[y|\mathbf{x}])$ since $E[y|\mathbf{x}] = e^{\mathbf{x}\boldsymbol{\beta}}$. However, there is no way to integrate the error out of $E[log(y)|\mathbf{x}]$ unless the error satisfies the homoskedasticity outlined in Section 1.2. Conceptually, the chief advantage of Poisson regression relative to log-linear regression is that it applies an inherently multiplicative model to outcomes that are likely to be multiplicatively related to covariates rather than transforming the data to make it fit a linear model, which introduces complications.

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 variance-mean 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"). Poisson regression does not fit the data as well when the data generating process does not conform to the conditional variance-mean equality restriction, resulting in a loss of efficiency. However, crucially for the usefulness of Poisson regression, departures from conditional variance-mean equality does *not* induce any bias or inconsistency in the regression coefficients. That is, they remain consistent semi-elasticity estimates as long as the standard orthogonality condition for consistent estimation holds.

There may be circumstances in which the universal baseline arrival rate is a function of an

observable "exposure" variable. For example, in their analysis of annual establishment-level workplace injuries, Cohn and Wardlaw (2016) and Cohn et al. (2021) identify an establishment's average number of employees and total hours worked as natural exposure variables. The exposure variable enters the Poisson regression equation as a covariate with coefficient constrained to 1. When an exposure is specified, the regression coefficients become estimates of rate semi-elasticities - e.g., the percent change in workplace injuries *per employee* associated with a one-unit change in a covariate.

One useful feature of the Poisson model is that, unlike most nonlinear models, it admits separable fixed effects. That is, the researcher can specify that each unit of observation (e.g., firm) have a different baseline outcome arrival rate. Let α_i be the fixed effect associated with firm i. Including this fixed effect, the Poisson model conditional expectation becomes:

$$E[y|\mathbf{x}] = e^{\alpha_i + \mathbf{x}\boldsymbol{\beta}} = e^{\alpha_i} e^{\mathbf{x}\boldsymbol{\beta}}. \tag{12}$$

Observe that, while the fixed effects in a linear model are additive, they are multiplicative in a Poisson regression, as they are in a log-linear regression. Group fixed-effects Poisson regression does restrict the usable sample to groups in which at least one observation has a non-zero value for the dependent variable.[4] However, this sample restriction may be desirable, as a lack of non-zero values for a group may indicate that the group is not subject to the data-generating process, in which case observations from the group may be uninformative about the relationship between y and $\mathbf{x}$.

Poisson models can also be used in conjunction with instruments in an instrumental variable (IV) estimation (Mullahy, 1997; Windmeijer and Santos Silva, 1997). The properties of the resulting IV estimates are similar to those of OLS-based IV estimates. One complication is that orthogonality is not generally sufficient to ensure consistent estimation in a Poisson IV

[4]This is a result of a more general phenomenon known as statistical separation as detailed by Correia et al. (2019).

regression. However, Mullahy (1997) provide a technique that allows for consistent estimation under the assumption of orthogonality. Moreover, the robustness of fixed-effects Poisson models can also be leveraged to deal with both time-constant and time-varying sources of endogeneity, as described by Lin and Wooldridge (2019).

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

$$\sum_{i=1}^{n}[y_i - e^{\mathbf{x}_i\beta}]x_i = 0. \tag{13}$$

Recent innovations in sparse matrix reduction methods have made the estimation of Poisson models via PPML maximization fast and reliable. The PPMLHDFE module for Stata based on (Correia et al., 2020) allows for speedy convergence of even high-dimensional fixed effects Poisson regression models.

Finally, examination of (13) shows that PPML estimation imposes no restriction on the domain of y other than requiring $y \geq 0$. Thus, even though the domain of the Poisson distribution is restricted to the counting numbers, Poisson regression can be estimated even if the distribution of y is continuous. The PPML estimator has been shown to perform well for a wide range of distributional assumptions, including those where the outcome contains a large number of zero values (Santos Silva and Tenreyro, 2011) or is continuous (Santos Silva and Tenreyro, 2006). Ultimately, fitting a true constant elasticity model when the data has a constant elasticity structure introduces fewer potential distortions and generally provides a better fit than forcing the data to fit a linear model by transforming it.

1.6 Other nonlinear models

Several other alternative regression models are available when working with outcome variables limited to non-negative values. The negative binomial model is similar to the Poisson model but relaxes the conditional variance-mean equality restriction by explicitly modeling the variance as a separate gamma process that allows for overdispersion (but not underdispersion). Specifically, negative binomial regression estimates a separate parameter α_{NB} that reflects the degree of overdispersion in the data. As $\alpha_{NB} \to 0^+$, the conditional variance and mean converge, and the negative binomial model distribution converges to the Poisson model distribution. Statistical packages generally report the estimate of α_{NB}, which is informative about the degree of overdispersion in the data. Because it relaxes the variance-mean 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 (zero-inflated Poisson or zero-inflated negative binomial) represent another alternative class of models. These models account for the possibility that some observations are not exposed to the underlying process that potentially gives rise to positive values of y. In addition to the regression equation (Poisson or negative binomial), they specify a second equation that relates exposure to observables. These models may be suitable when working with count data that has an excessive number of zero values relative to a Poisson or negative binomial distribution. Of course, estimating such a model requires explicitly specifying the process determining exposure, which may be difficult to intuit.

Another alternative is the Type I Tobit regression model. Applied to outcomes bounded below by 0, this model specifies $y = \max(0, \mathbf{x}\boldsymbol{\beta} + \epsilon)$, where ϵ is mean-zero, normally distributed, and homoskedastic. An important difference between Tobit regression and log-linear, Poisson, or negative binomial regression is that Tobit regression assumes an additive rather than multiplicative relationship between the outcome and covariates. As we discussed in Section 1.2, a multiplicative relationship is often more natural when working with count

outcomes or other zero-bounded, right-skewed outcomes. As a result, Tobit regression may have poor efficiency when working with outcomes of these types.

Each of these alternative models has useful features. However, they all have one critical limitation – they do not admit separable fixed effects. 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 can cause the estimated coefficients on all variables to be biased and inconsistent (Lancaster, 2000). Asymptotically, estimates converge to the true coefficient values as T increases but not as N increases. Since controlling for group (e.g., firm or establishment) and time fixed effects is often considered essential for identification in corporate finance applications, the inability of these alternative models to accommodate fixed effects severely limits their usefulness in the field.

1.7 OLS rate regression model

The log-linear, log1plus-linear, and inverse hyperbolic sine regressions all decrease skewness in the outcome variables through concave transformation of the data. Poisson and negative binomial regressions both fit multiplicative models that effectively assume skewed outcomes. A third approach to addressing concerns about right-skewed outcomes is to scale the outcome variable by an appropriate scaling variable s, transforming it into a rate $r = y/s$, and then to estimate a linear regression:

$$r = \mathbf{x}\boldsymbol{\beta} + \epsilon \tag{14}$$

For example, Cohn and Wardlaw (2016) and Cohn et al. (2021) scale the number of workplace injuries at an establishment in a given year by the average number of employees working at the establishment in the year, which naturally determines the baseline level of

workplace injury exposure at the establishment. Scaling in this manner may not mitigate skewness due to coding errors or peculiarities of the data set. However, it does mitigate skewness due to differences in scale, especially if scale itself has a skewed distribution, as is typically the case in corporate finance applications. Note that a scaling variable of the type we describe here is equivalent to an exposure variable in the context of a Poisson model.

If an appropriate exposure variable is available, then scaling by the exposure variable and estimating a linear rate regression may be preferable to estimating a Poisson regression. Specifically, linear rate regression may be more efficient since, unlike Poisson regression, it does not impose any restriction on the conditional variance of the count outcome. In Section 2.3, we compare the efficiency of Poisson regression and OLS rate regression. Unfortunately, in many finance applications, an appropriate exposure variable does not exist. For example, total assets and total sales, which are common measures of firm scale, are probably poor approximations of a firm's baseline exposure to an activity such as patenting.[5]

2 Simulations

The section presents three sets of simulations that further illustrate the econometric properties of different estimators when working with count-based and continuous zero-bounded, right-skewed outcome variables. The first simulation examines how the addition of the constant affects estimates from log1plus-linear regression specifically and logcplus-linear regression generally. The second simulation examines the degree of bias that heterogeneity introduces into log-linear regression estimates. The third simulation examines the efficiency of linear regression, log-linear regression, and under various conditions.

[5]While still noisy, research and development expenditures may more closely approximate a firm's exposure to patenting activity specifically. However, treating patents per dollar of R&D spending as the outcome changes the interpretation of regression coefficients. These coefficients represent estimates of the effect of covariates on the patenting efficiency associated with each dollar spent on R&D rather than the effect of the covariates on overall patenting activity. If a researcher's objective is to test a theory that a policy affects patenting in part by inducing research activity, then scaling by R&D will produce uninformative estimates.

2.1 The effect of adding the constant in log1plus-linear regression

As we argue in Section 1, the addition of the constant in a log1plus-linear regression is pernicious under the fairly natural and seemingly benign assumption that the log of the outcome variable is conditionally linear in the covariates if, as is generally the case, covariates are correlated with each other. We now demonstrate the potential severity of this effect with a simple simulation and show that the addition of the constant can easily cause a log1plus-linear regression coefficient to have the wrong sign. Secondarily, we show that the magnitudes of logcplus-linear regression coefficients vary sharply with the value of c, the choice of which is arbitrary.

We simulate three sets of 5,000 observations (x_1, x_2, y), with $y = e^{\beta_1 x_1 + \beta_2 x_2} * k$, each for a different constant k. We use k as a shifter of the conditional mean of y, which allows us to assess how the properties of the different estimators change as the conditional mean changes. We set $\beta_1 = 1$ and $\beta_2 = -0.1$. For each observation, we draw the value of x_1 from a standard normal distribution. We consider two different specifications for x_2. In the first, we also draw x_2 from an independent standard normal distribution. In the second, we set $x_2 = \max\{x_1, 0\}$. Our intent in making x_2 a kinked function of x_1 in the second specification is to induce a nonlinear relationship between x_1 and x_2. There is nothing special about the kinked nature of the function we choose. Other, more complex nonlinear relationships yield the same insight.[6] Within each specification of x_2, we consider three different values of the shifter k – 1, 5, and 25. For each combination of specification of x_2 and value of k, we estimate Poisson, log-linear, log1plus-linear regressions, log0.1plus-linear, log10plus-linear, and inverse hyperbolic sine (IHS) regressions of y on x_1 and x_2 using the simulated data.[7]

[6] In some scenarios, a kinked relationship is a reasonable approximation to a real relationship. For example, suppose that x_1 is the size of a firm measured by total assets and x_2 is the number of stock analysts covering the firm. In reality, the number of analysts covering a firm generally increases with firm size, though a fairly large set of smaller firms has no analyst coverage. If there is a minimum size threshold below which brokerages do not assign analysts to firms, then the relationship between number of analysts and firms size will be approximately kink shaped.

[7] Note that we do not include an error term in y in this simulation as we do not need one to demonstrate

Table 1 presents the regression results.

[Insert Table 1]

Panel A reports results for the model specification where x_2 is independent of x_1, while Panel B reports results for the specification where x_2 in a kinked function of x_1. Each panel shows results for the cases where $k = 1$, $k = 5$, and $k = 25$. Panel A shows that the Poisson and log-linear regressions both recover the true values of β_1 and β_2 when x_1 and x_2 are uncorrelated. In contrast, log1plus-linear regression produces coefficients with the same signs as the true values but smaller values, as our analysis in Section 1.3 suggests that we should when x_1 and x_2 are independent. These differences increase as k increases from 1 to 5 and then to 25. Intuitively, as the conditional mean of y gets larger, the ratio of $1 + E[y|x]$ to $E[y|x]$ becomes closer to 1, and the semi-elasticities of y and $1 + y$ converge.

Panel B shows that Poisson and log-linear regression continue to recover the true values of β_1 and β_2 when x_1 and x_2 are nonlinearly related – that is, they produce the same (correct) estimates of β_1 and β_2 as in the case where x_1 and x_2 are independent. In contrast, log1plus-linear regression coefficients change when x_2 is a nonlinear function of x_1 rather independent. Thus, the coefficients can no longer even be interpreted as the semi-elasticities of $1 + y$ with respect to x, confirming our conclusions in Section 1.3. In fact, the coefficient on x_2 in the $k = 1$ case is positive, while the true value of β_2 is negative. Thus, a researcher estimating a log1plus-linear regression using this simulated data would conclude that x_2 has a positive effect on y, while the true effect is negative, even though there is no sampling error. As in Panel A, the log1plus-linear regression coefficients are closer to the true values of β_1 and β_2 when k is larger, though they appear to converge at a slower rate than when x_1 and x_2 are

our main points. Because y is a deterministic function of x_1 and x_2, regression estimates obtained from the simulated data are nearly perfectly precise as long as the sample is sufficiently large.

independent.

Comparing the log1plus-linear, log0.1plus-linear, and log10plus-linear regressions all yield sharply different coefficients on both x_1 and x_2. While expected and purely mechanical, the strong sensitivity of the coefficients to a purely arbitrary choice of constant highlights the lack of meaning in logcplus-linear regressions in general. The coefficient on x_2 in the IHS regression also has the wrong sign when x_1 and x_2 are nonlinearly related and $k = 1$ or 5. As Bellemare and Wichman (2020) argue, the IHS regression performs better when the mean value of y is higher.

2.2 Log-linear Regressions and Heteroskedasticity

In our second 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 the exact point estimate of a coefficient in a constant elasticity model. However, researchers in finance may be more interested in the direction of a relationship than in interpreting an exact point estimate. In this set of simulations, we demonstrate that the bias can actually cause estimates to have the wrong sign in expectation, making them uninformative about even the direction of a relationship.[8]

We simulate a set of observations (x, y), with $y = \exp(\beta x)\eta$, where η represents a mean-1 multiplicative error. We set $\beta = 0.2$. We write y as a function of a multiplicative error for convenience, though we can recast the error as an additive mean-0 error term. We evaluate the effects of heteroskedasticity in two different scenarios - one where we x is an i.i.d. binary random variable equal to 0 or 1, each with equal probability, and the other where x is

[8]In applications such as the gravity model of trade that (Santos Silva and Tenreyro, 2006) analyze, a negative coefficient would not be sensible. However, in corporate finance applications, we often lack strong priors on the sign of a relationship.

an i.i.d. random variable drawn from a standard normal distribution truncated at the 1st and 99th percentiles. In both scenarios, we draw η from an lognormal distribution with an independent mean of 1 and standard deviation $\sigma_\eta(x)$. For the scenario where x is binary, we define $\sigma_0 = \sigma_\eta(0)$ and $\sigma_1 = \sigma_\eta(1)$. The error is homoskedastic in the special case where $\sigma_1 = \sigma_0$ but heteroskedastic otherwise. For the scenario where x is continuous, we assume that σ_η is an exponential function of x.

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

For each of the six cases, we generate 10,000 simulated data sets of 5,000 observations. We then estimate Poisson, log-linear, and log1plus-linear regressions, log0.1plus-linear, log10plus-linear, and IHS regressions for each data set. For each regression coefficient estimate, 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 2, Panel A reports these means.

[Insert Table 2]

The mean coefficients from the Poisson regressions are approximately 0.2 in all six cases. That is, the Poisson regression recovers the true model coefficients, on average, despite the presence of heteroskedasticity. The log-linear regressions also recover the true model

coefficients in two cases where the multiplicative error is homoskedastic - the binary x case where $\sigma_1 = \sigma_0 = 1.5$ and the continuous x case where $h = 0$. When $\sigma_1 > \sigma_0$ ($\sigma_1 < \sigma_0$) in the binary x scenario or $h = 1$ ($h = -1$) in the continuous x scenario, the log-linear regression coefficient is greater (less) than the true parameter. The directions of the bias are consistent with our conclusion in Section 1 that a positive relationship between the variance of the error and a covariate generally downward biases log-linear regression estimates, while a negative relationship generally upward biases these estimates. Indeed, when the variance of the error increases with x in the simulations ($\sigma_1 = 2$ and $\sigma_0 = 1$ in the binary case; $\sigma_\eta = e^x$ in the continuous case), the log-linear regression coefficient is negative, even though the true coefficient is positive.

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 is at the group level, even if there is no actual fixed effect in y. 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 denote the group level and it denote observation t within group i. We assume that $x_{it} = .5\mu_i + .5\nu_{it}$, where μ_i and ν_{it} are i.i.d. randome variables each drawn from a normal distribution truncated at 1st and 99th percentiles. We then assume that $\sigma_\eta = e^{\gamma\mu_i + (1-\gamma)\nu_{it}}$. We examine five cases, each corresponding to a different value of γ. For each case, we generate 10,000 data sets of 5,000 observations apiece, with 500 independent groups and 10 observations within each group. For each simulation, we estimate Poisson and log-linear regression models, each both with and without group fixed-effects. Panel B of Table 2 reports the mean coefficient and standard error across simulated data sets for each regression. 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, while log-linear regression without group fixed effects results in a negatively biased coefficient in all cases. When the heteroskedasticity is driven entirely by variation at the group level ($\gamma = 1$), the fixed-effects log-linear regression recovers the true parameter value 0.2, despite the heteroskedasticity in η. However, when most of the relationship between the variance of η and x reflects a relationship with the within-group variation in x, then including group fixed effects in the log-linear regression amplifies the bias.

Intuitively, log-transforming a random variable translates a dependence of the variance of the variable on x into a dependence of the mean of the transformed variable on x, which results in biased estimates. However, if the variance of the untransformed error only depends on the group fixed effect, then absorbing group fixed effects in the log-linearized regression sweeps out the group mean of the transformed error, so that the conditional expectation of the transformed error is independent of x. Thus, controlling for group fixed effects "solves" the heteroskedasticity problem in log-linear regression, even if there is no fixed effect in y, as long as the variance of the error is only related to the group-level component of x.

2.3 Efficiency of three unbiased estimators

In our final set of simulations, we explore the efficiency of the three regression models discussed in Section 1 that produce unbiased and consistent estimates under standard exogeneity conditions when confronted with count outcomes and continuous, zero-bounded, right-skewed outcomes – OLS regression where the dependent variable is the raw outcome, Poisson regression, and OLS rate regression. More specifically, we compare rejection rates for the different regression models for different assumed magnitudes of effect and degrees of overdispersion. We simulate panels of observations $(x_{1,it}, x_{2,it}, y_{it})$, where i indexes groups and t observations within group. For each observation, we draw two random variables, μ_i and ν_{it}, each from an independent standard normal distribution truncated at the 1st and

99th percentiles, and set $x_{1,it} = .5\mu_i + .5\nu_{it}$. This structure produces a group fixed effect in x_{it}. We independently draw $x_{2,it}$ from a normal distribution with a mean of 0 and a standard deviation of 2.

For the count data simulation, we draw y_{it} from a negative binomial distribution with conditional mean $E[y_{it}|\mathbf{x}] = e^{\beta_1 x_{1,it} + x_{2,it}}$ and overdispersion parameter $\alpha_{i,NB} = (1 + \frac{\mu_i}{4.654}) * \alpha_{sim,NB}$.[9] For the continuous data simulation, we model y_{it} as a continuous variable using the mixture model approach of Santos Silva and Tenreyro (2011). In this formulation, y_{it} is the sum of a random number m_i of random variables z_{it}, where m_i is a negative binomial random variable with mean $e^{\beta_1 x_{1,it} + x_{2,it}}$, and z_{it} is a $\chi^2_{(1)}$ distributed random variable. This formulation allows for continuously distributed outcomes with a mass at 0. 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. This mixture model generates a continuous outcome variable that is zero-bounded and right-skewed. Because $x_{2,it}$ lacks a coefficient, we interpret it as an exposure variable reflecting the baseline exposure of an observation to the process that drives the magnitude of the outcome.

For both the count and continuous outcome simulations, we evaluate 12 different cases, each of which is a different combination of β_1 parameter and conditional variance of y. We consider four values of β_1: -0.3, -0.1, 0.1, and 0.3. For the count outcome simulations, where the conditional variance is determined by $\alpha_{sim,NB}$, we consider three values $\alpha_{sim,NB}$: 0.001, 3, and 8. For the continuous outcome simulations, where the conditional variance is determined by b, we consider the same three values for b (0.001, 3, and 8). As noted previously, the negative binomial distribution converges to the Poisson model distribution as $\alpha_{sim,NB} \to 0$. Therefore, the case where $\alpha_{sim,NB} = 0.001$ is approximately the case where

[9] μ_i has a range of [-2.327,2.327] because it is a standard normal distribution truncated at 1% and 99% tails.

the data is generated by a Poisson model. The case where $\alpha_{sim,NB} = 3$ approximates the distribution of common count data sets such as firm-year corporate patent data. The case where $\alpha_{sim,NB} = 8$ represents extreme overdispersion. Similarly, small values of b result in an approximately Poisson distributed outcome, while large values of b result in significant overdispersion.

For each of the two approaches and each of the 12 cases within each approach, 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 different regression models using each simulated panel. Finally, for each combination of outcome type (count and continuous), overdispersion level, coefficient β_1, and regression model, we compute 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. Table 3 reports these percentages.

[Insert Table 3]

Panels A, B, and C report results for the count outcome simulations where $\alpha_{sim,NB} = 0.001$, $\alpha_{sim,NB} = 3$, and $\alpha_{sim,NB} = 8$, respectively. Panels D, E, and F report results for the continuous outcome simulations where $b = 0.001$, $b = 3$, and b, respectively. Not surprisingly, OLS regression where the dependent variable is the raw outcome (y) exhibits the least power in all scenarios, rejecting at much lower rates than the other two regression models. This lack of power is not surprising given the skewed distribution of y, which, as we have already noted, is why researchers often resort to log-transforming the dependent variable before estimating OLS regressions.

Poisson regression exhibits more power than OLS rate regression when the data is approximately Poisson distributed (Panels A and D). It also exhibits more power with moderate and

large levels of overdispersion (Panels B, C, E, and F) when the true effect being estimated is small ($\beta_1 = -0.1$ or $\beta_1 = 0.1$). The same is true when the true effect to be estimated is large ($\beta_1 = -3$ or $\beta_1 = 3$) in the continuous outcome simulations (Panel F). However, OLS rate regression exhibits more power than Poisson regression in the count outcome simulations when β_1 is large (Panels B and D). These findings suggest that OLS rate regression may outperform Poisson regression in terms of efficiency, at least in some cases, when the effect that the researcher is attempting to estimate is likely to be first order. However, Poisson regression still performs well in these cases unless overdispersion is extreme, and it outperforms OLS rate regressions, in some cases by large margins, when the effect to be estimated is likely to be relatively small.

3 Replications and Decomposition

In this section, we replicate six data sets analyzed in existing papers published in top-3 finance journals and present log-linear, log1plus-linear, and Poisson regression estimates using each data set. We then introduce and apply a technique to quantify the roles of the addition of the constant and heteroskedasticity in driving differences between log1plus-linear and Poisson regression estimates.

3.1 Comparisons of Regression Estimates Using 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). The first three papers rely on log1plus-linear regression, while the fourth relies primarily on Poisson regression. These four papers collectively have 3,419 Google

Scholar 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), who rely on log1plus-linear regression, and Xu and Kim (2020), who rely on log-linear regression. We choose these six papers because they are 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 the data preparation outlined in the paper as best we can, including any adjustments for patent truncation (Dass, Nanda, and Xiao, 2017). 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. For each paper, we tabulate the results that the authors report in the paper for the main regression specification as well as results from log1plus-linear, log-linear, and Poisson regressions that we estimate ourselves using the replicated data set. Two pieces of evidence give us confidence in the fidelity of our data set replications. First, we are able to effectively replicate the main results from all six papers. Second, our sample summary statistics (untabulated) line up with those presented in the papers.

Table 4 presents the analysis for the Hirshleifer, Low, and Teoh (2012) paper. We replicate Table V column (1) from this paper. The main explanatory variable is *Confident CEO (Options)*, an indicator variable equal to one if a firm's CEO holds options that are at least 67% in the money and zero otherwise. The coefficient on *Confident CEO (Options)* from the log1plus-linear regression that the paper reports is positive and statistically significant at the ten percent level. The log1plus-linear estimate from our replication is also positive and statistically significant, and it is close in magnitude to the estimate reported in the paper. Our log-linear and Poisson estimates are positive and statistically significant as well, though

they are approximately twice as large as the estimate from the log1plus-linear regression. The coefficients on the control variables are also larger in the log-linear and Poisson regressions than in the log1plus-linear regression.

[Insert Table 4]

Table 5 presents the analysis for the He and Tian (2013) paper. We replicate Table 2 column (4) from this paper. The main explanatory variable is $lnCoverage$, which is the natural log of the number of stock analysts covering a firm in a given year. The coefficient on $lnCoverage$ from the log1plus-linear regression that the paper reports is negative and statistically significant at the one percent level. Our replication of this log1plus-linear regression also yields a negative coefficient on $lnCoverage$ with statistical significance at the one percent level, though the magnitude is smaller than the estimate reported in the paper. In contrast, log-linear and Poisson regression both yield *positive* coefficients on $lnCoverage$. Some of the control variables have the same sign across all three replication specifications, while others, such as the coefficient on $LnAge$, differ in sign.

[Insert Table 5]

Table 6 presents the analysis for the Fang, Tian, and Tice (2014) paper. We replicate Table 2 column (1) from this paper. The main explanatory variable is $ILLIQ$, which is the natural logarithm of the annual relative effective spread (the absolute value of the difference between the execution price and the midpoint of the prevailing bid-ask quote), divided by the midpoint of the prevailing bid-ask quote. The coefficient on $ILLIQ$ from the log1plus-linear regression that the paper reports is positive and statistically significant at the one percent

level. Our replication of this log1plus-linear regression also yields a positive and statistically coefficient on *lnCoverage* that is similar in magnitude. In contrast, log-linear regression yields a coefficient that, while positive, is an order of magnitude smaller and statistically insignificant, and Poisson regression yields a negative coefficient.

[Insert Table 6]

Table 7 presents the analysis for the Amore, Schneider, and Žaldokas (2013) paper. We replicate Table 3 column (4) from this paper. The main explanatory variable is *Interstate deregulation*, an indicator variable equal to one if a firm is headquartered in a state that has passed an interstate banking deregulation and zero otherwise. The coefficient on *Interstate deregulation* from the Poisson regression that the paper reports is positive and statistically significant at the one percent level. Our replication of this Poisson regression also yields a positive and statistically significant coefficient on *Interstate deregulation* that is similar in magnitude. Log-linear and log1plus-linear regression also yield positive coefficients, but these coefficients are much smaller in magnitude and are statistically insignificant.

[Insert Table 7]

Table 8 presents the analysis for the Akey and Appel (2021) paper. We replicate Table 3 column (1) from this paper. The main explanatory variable is *Bestfoods*, an indicator variable equal to one if an establishment is located in a jurisdiction where parent company liability for subsidiary debt declines as a result of the Bestfoods court ruling. The coefficient on *Bestfoods* from the log1plus-linear regression that the paper reports is positive and statistically significant at the one percent level. Because we use the authors' published replication

package to construct the data set, our replication of this log1plus-linear regression yields identical results. In contrast, log-linear regression yields a coefficient that, while positive, is an order of magnitude smaller, and Poisson regression yields a negative coefficient.

[Insert Table 8]

Finally, Table 9 presents the analysis for the Xu and Kim (2020) paper. We replicate Table 2 column (4) from this paper. The main explanatory variable is *HM Debt*, a text-based measure of debt-related financing constraints developed by Hoberg and Maksimovic (2015). Because this paper estimates log-linear regressions, the sample is constrained to firm-years with a positive number of pollutants. The coefficient on *HM Debt* from the log-linear regression that the paper reports is positive and statistically significant. Our replication of this log-linear regression yields similar results, though the coefficient on *HM Debt* falls just short of statistical significance (t-statistic of 1.55). The log1plus-linear regressions also yields a positive but statistically insignificant coefficient. In contrast, the Poisson regression coefficient on *HM Debt* is both positive and statistically significant, at the 5% level. It is also 59% larger than the log-linear regression coefficient.

[Insert Table 9]

Based on this small sample of replication exercises, it appears that the choice of regression model frequently makes a big difference in terms of conclusions in real-world applications involving count outcomes and continuous, zero-bounded, right-skewed data. The coefficients on the main variable of interest from our estimates of log1plus-linear and Poisson regressions have the same signs in three of the six replications and the opposite signs in the other three.

Of the three where the signs agree, the Poisson regression estimate is 68%, 309%, and 233% larger than the log1plus-linear regression estimate (Tables 4, 7, and 9, respectively).

3.2 Explaining differences in log1plus-linear and Poisson coefficients

The results in Tables 4 through 9 suggest that log1plus-linear 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. First, the addition of the constant could cause differences between the coefficients for two reasons – they inherently measure different quantities and are thus not comparable, and non-linearities in the relationship between $1+y$ and covariates could cause estimation bias. Second, relationships between higher order moments of a multiplicative regression error and covariates could bias the log1plus-linear regression coefficients. Third, the samples being used in estimating the two regressions differ because Poisson regression excludes from the sample firms/establishments for which the outcome variable is zero in every period.

We conduct a decomposition exercise in an effort to estimate how much of the difference between the log1plus-linear and Poisson estimates from each replication exercise is attributable to each of the three possible drivers of differences just described. To estimate the effect of the addition of the constant in log1plus-linear regression, we first fit a Poisson regression to compute the predicted values of y, which we label $\hat{y}$. We then estimate a log1plus-linear regression where we substitute $\hat{y}$ for y and 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 removing the effects of heteroskedasticity (by removing the noise completely).[10]

[10]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 zero-valued observations.

To estimate the effects of the Poisson sample restriction on the difference between log1plus-linear and Poisson regression estimates, we again substitute $\hat{y}$ for y and estimate a log1plus-linear regression, this time without imposing the Poisson sample restriction. The difference between these estimates and the estimates where we impose the Poisson sample restriction capture the effect of the sample differences, holding fixed the regression model and again removing the effects of heteroskedasticity. Finally, to estimate the effects of heterkoskedasticity, we compare the estimates from log1plus-linear regression on the full sample, with $\hat{y}$ as the dependent variable, to the actual log1plus-linear regression, where y is the dependent variable. The difference between the two captures the effect of any relationships between higher order moments of the log1plus-linear regression error and covariates. Our approach here represents a decomposition in the sense that the sum of the three estimated effects equals the total difference between Poisson and log1plus-linear regression estimates.

Table 10 reports the coefficient estimates when we apply the procedures described above, where each panel reports results for the replication from a different paper. Based on a comparison of the first and second columns, the effect of adding the constant in log1plus-linear regression is large in all four patent replications, reversing the sign of the coefficient of interest in two, but is small in the two toxic release replications. Based on comparison of the second and third columns, the effect of sample differences appears to be minimal in general. Based on a comparison of the third and fourth columns, the correlation behind higher order error moments and covariates causes substantial differences in 4 of the 6 regressions (Panels C, D, E, and F). Overall, then, both the addition of the constant in log1plus-linear regression and the effects of correlation between higher order error moments and covariates appear to cause substantial differences between log1plus-linear and Poisson regression estimates in real-world applications.

[Insert Table 10]

4 Conclusion

This paper highlights the issues surrounding model choice when working with outcome variables based on count data or continuous, zero-bounded data, the analysis of both of which is increasingly common in corporate finance. Our analysis suggests that researchers should rely primarily on either Poisson regression or, if a suitable exposure variable is available and the researcher is concerned about overdispersion, OLS rate regressions when working with such data. 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, using the Stata module PPMLHDFE. In contrast, commonly-used OLS regressions where the dependent variable is the log of 1 plus the count produce estimates that have no economic meaning and are generally subject to multiple sources of inherent bias, even if errors in the count are uncorrelated with covariates. Our replications of data sets in six published papers using patent and toxic pollution data suggest that this bias produces substantially different inferences from Poisson regression estimates.

References

Akey, P. and I. Appel (2021). The limits of limited liability: Evidence from industrial pollution. *The 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.

Bellemare, M. F. and C. J. Wichman (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics 82*(1), 50–61.

Cohn, J., N. Nestoriak, and M. Wardlaw (2021). Private Equity Buyouts and Workplace Safety. *The Review of Financial Studies* Forthcoming, hhab001.

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

Correia, S., P. Guimarães, and T. Zylkin (2019, August). Verifying the existence of maximum likelihood estimates for generalized linear models. arXiv: 1903.01633.

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

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? *The 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. Publisher: [Wiley, Econometric Society].

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? *The journal of finance 67*(4), 1457–1498.

Hoberg, G. and V. Maksimovic (2015). Redefining financial constraints: A text-based analysis. *The Review of Financial Studies 28*(5), 1312–1352.

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

Lin, W. and J. M. Wooldridge (2019, January). Chapter 2 - Testing and Correcting for Endogeneity in Nonlinear Unobserved Effects Models. In M. Tsionas (Ed.), *Panel Data Econometrics*, pp. 21–43. Academic Press.

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

Mullahy, J. (1997, November). Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior. *The Review of Economics and Statistics 79*(4), 586–593.

Santos Silva, J. M. C. and S. Tenreyro (2006, November). The Log of Gravity. *The Review of Economics and Statistics 88*(4), 641–658. Publisher: MIT Press.

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

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. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-1255%28199705%2912%3A3%3C281%3A%3AAID-JAE436%3E3.0.CO%3B2-1.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Xu, Q. and T. Kim (2020). Financial constraints and corporate environmental policies. *The Review of Financial Studies*.

Figure 1: Histogram of firm-year patents granted

This figure presents a histogram of number of patents granted by firm-year using replicated dataset of He and Tian (2013). Each bar in the histogram has a width of 1. We top-code the count at 100 to make the figure easier to read. Hence, the left-most bar represent the percent of firm-year observations with 0 patents, and the right-most bar represents the percent of firm-year observations with 100 or more patents.

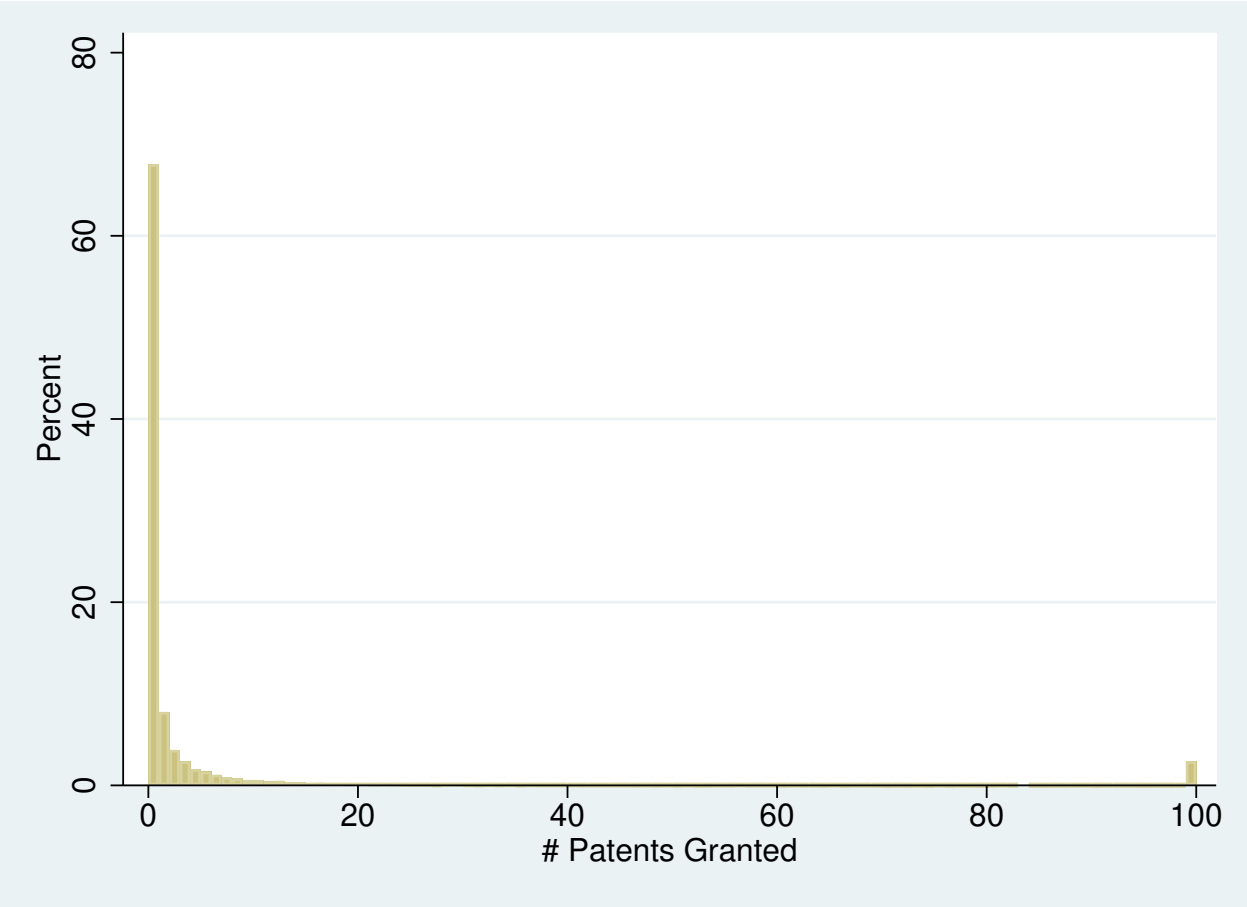
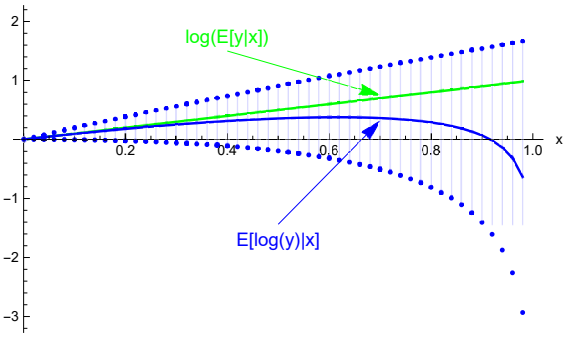


Figure 2: Log-linear regressions and heteroskedasticity

This figure presents two cases of an example involving correlations between the variance of errors in y and a covariate. In both examples, $y = e^{\alpha+\beta x}\eta$, with $\alpha = 0$ and $\beta = 1$, so that $y = e^{x}\eta$, and $\eta \in \{\eta^{-}, \eta^{+}\}$, with $prob(\eta = \eta^{-}) = prob(\epsilon = \eta^{+}) = 0.5$. In the first case (subfigure a), $\eta^{-} = 1 - x$ and $\eta^{+} = 1 + x$. In this case, the variance of η increases with x. In the second case, $\eta^{-} = x$ and $\eta^{+} = 2 - x$. In this case, the variance of η decreases with x. In each subfigure, the solid green line depicts $log(E[y|\mathbf{x}])$. The two sets of blue points for each x value represent the values of $log(y)$ when $\epsilon = \epsilon^{+}$ (top point) and when $\epsilon = \epsilon^{-}$ (bottom point). The solid blue line depicts $E[log(y)|x] = 0.5(e^{x}\epsilon^{+} + e^{x}\epsilon^{-})$ as a function of x.

(a) $corr(var(\epsilon), x) > 0$



(b) $corr(var(\epsilon), x) < 0$

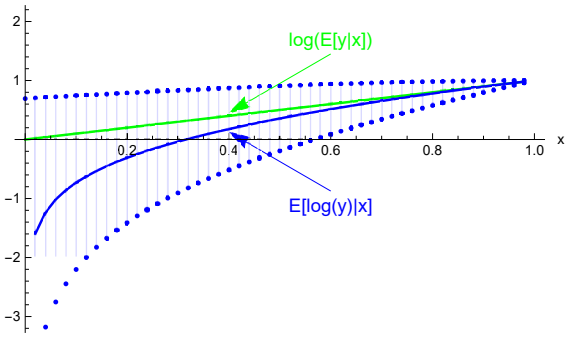


Figure 3: Simulated Log(y) and Log(1+y)

This figure plots the data from the first simulation. In this simulation, we generate x_1 and x_2 as standard random variables and define $y = e^{x_1 - .1 * x_2}$. We then take the log of y and $1 + y$ and generate scatter plot. This figure illustrates how $log(1 + y)$ converges to 0 and differs significantly from $log(y)$ when y is close to 0.

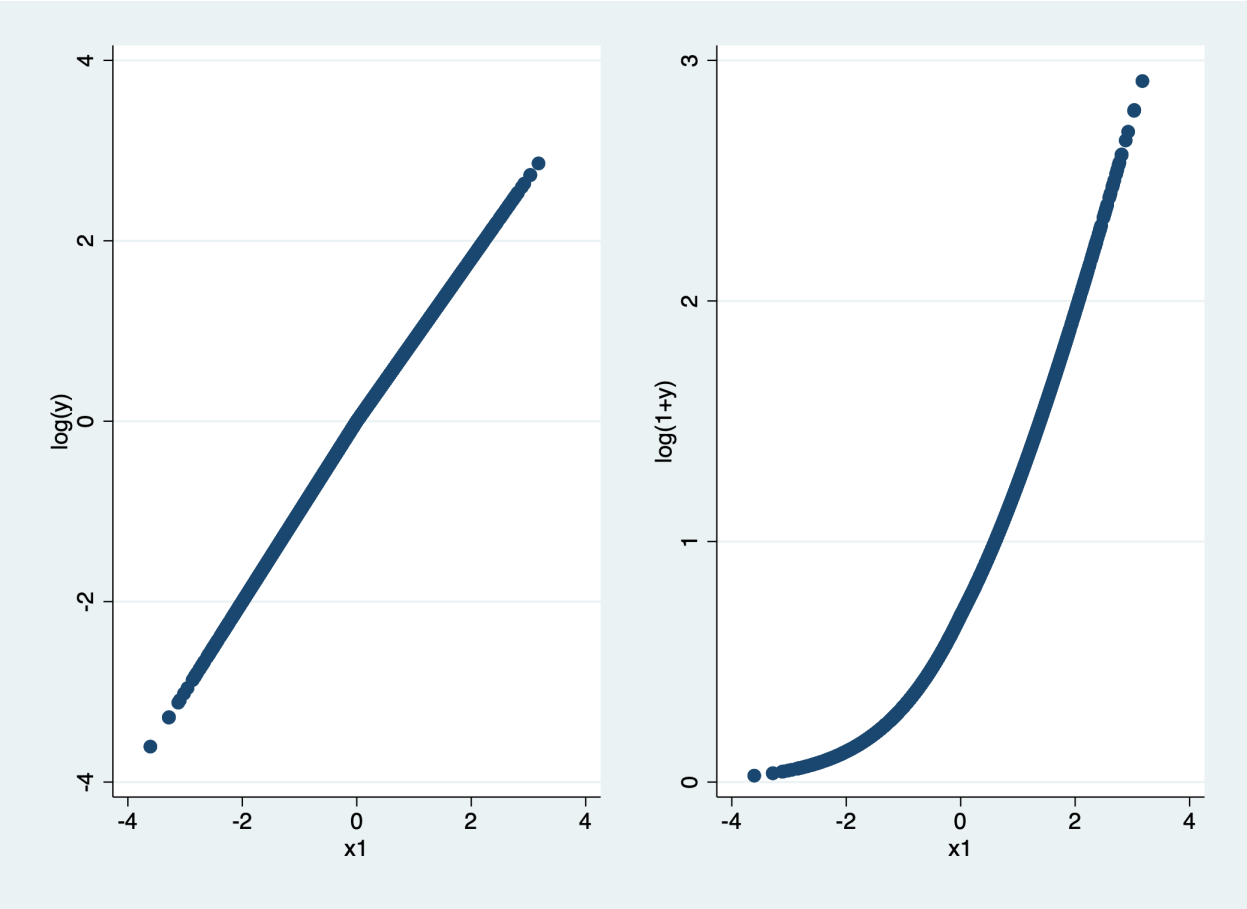


Table 1: Constant Added Simulation

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

$$y = e^{\beta_1 x_1 + \beta_2 x_2} * k.$$

We set $\beta_1 = 1$ and $\beta_2 = -0.1$. For each observation, we draw the value of x_1 from a standard normal distribution. For the analysis reported in Panel A, we also draw the value of x_2 from a standard normal distribution. For the analysis reported in Panel B, we set $x_2 = \max\{x_1, 0\}$. In each case, we report estimates from Poisson, log-linear, log1plus-linear, log0.1plus-linear, log10plus-linear, and IHS regressions of y on x_1 and x_2. Because we do not include noise in y, the regressions estimates are all perfectly precise, so there are no standard errors to report. Panel C reports summary statistics for the y values in each of the four simulated data sets.

Panel A: x_1 and x_2 independent

	(1) Poisson	(2) Log-linear	(3) Log1plus	(4) Log0.1plus	(5) Log10plus	(6) IHS
$k = 1$:						
x_1	1.000	1.000	0.498	0.875	0.122	0.656
x_2	-0.1000	-0.1000	-0.0504	-0.0878	-0.0128	-0.0661
Constant	0.000	0.000	0.808	0.142	2.444	1.029
$k = 5$:						
x_1	1.000	1.000	0.793	0.969	0.359	0.933
x_2	-0.1000	-0.1000	-0.0796	-0.0970	-0.0365	-0.0936
Constant	1.609	1.609	1.864	1.642	2.813	2.347
$k = 25$:						
x_1	1.000	1.000	0.942	0.993	0.680	0.995
x_2	-0.1000	-0.1000	-0.0943	-0.0994	-0.0684	-0.0995
Constant	3.219	3.219	3.281	3.225	3.654	3.915

Panel B: $x_2 = \max\{x_1, 0\}$

	(1) Poisson	(2) Log-linear	(3) Log1plus	(4) Log0.1plus	(5) Log10plus	(6) IHS
$k = 1$:						
x_1	1.000	1.000	0.272	0.772	0.0259	0.373
x_2	-0.1000	-0.1000	0.374	0.110	0.160	0.472
Constant	0.000	0.000	0.627	0.0589	2.369	0.802
$k = 5$:						
x_1	1.000	1.000	0.640	0.940	0.151	0.845
x_2	-0.1000	-0.100	0.213	-0.0410	0.347	0.0785
Constant	1.609	1.609	1.741	1.618	2.648	2.275
$k = 25$:						
x_1	1.000	1.000	0.890	0.987	0.480	0.987
x_2	-0.1000	-0.1000	0.00668	-0.0870	0.311	-0.0843
Constant	3.219	3.219	3.238	3.220	3.493	3.908

Panel C: Summary Stats

	mean	sd	p10	p50	p90
x_1 & x_2 independent, $k = 1$	1.661948	2.035685	.2771289	1.004888	3.711035
x_1 & x_2 independent, $k = 5$	8.309741	10.17842	1.385644	5.024441	18.55518
x_1 & x_2 independent, $k = 25$	41.5487	50.89212	6.928223	25.1222	92.77588
$x_2 = \max\{x_1,0\}$, $k = 1$	1.493545	1.594144	.2775079	1.017342	3.232467
$x_2 = \max\{x_1,0\}$, $k = 5$	7.467726	7.970722	1.38754	5.086708	16.16233
$x_2 = \max\{x_1,0\}$, $k = 25$	37.33863	39.85361	6.937699	25.43354	80.81167

Table 2: Heteroskedacity Simulation

This table reports results from simulations in which we introduce heteroskedasticity into the outcome variable. We simulate a set of observations (x, y), where $y = exp(\alpha + \beta x)\eta_i$. We set $\alpha = 0$ and $\beta = 0.2$. For each observation, we draw x either as a continuous variable from a truncated standard normal distribution (truncated at the 1st and 99th percentile) or as a binary indicator from a Bernoulli distribution with $p(x = 0) = p(x = 1) = .5$. We draw η from an independent lognormal distribution with a mean of 1 and a standard deviation given in each panel. In each panel, we generate 5,000 observations and run 10,000 simulations.

Panel A presents the mean coefficient and standard error for Poisson, LOLS, LOGcPLUS, and Inverse Hyperbolic Sine (IHSY) regressions of y on x, with the standard deviation of the error η_i given at the top.

Panel B presents an alternative model where $\beta = 0.2$. The data is simulated 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 $x_{it} = .5\mu_i + .5\nu_{it}$. The heteroskedasitic error is simulated as a weighted function of each part with a weight on the fixed part γ and a weight on the time-varying part $1 - \gamma$.

Panel A

$\sigma_\eta =$	Continuous x						Binary x					
	exp(x)		exp $-(x)$		exp(1/2)		1 if $x = 0$ 2 if $x = 1$		2 if $x = 0$ 1 if $x = 1$		1.5	
	Coef	S.E.	Coef	S.E.	Coef	S.E.	Coef	S.E.	Coef	S.E.	Coef	S.E.
Poisson	0.200	0.044	0.201	0.044	0.200	0.046	0.199	0.047	0.201	0.031	0.200	0.025
Log-linear	-0.258	0.030	0.659	0.030	0.200	0.032	-0.300	0.019	0.700	0.019	0.200	0.017
Log1plus-linear	-0.015	0.013	0.178	0.013	0.078	0.014	-0.017	0.007	0.171	0.005	0.075	0.007
Log0.1plus-linear	-0.153	0.025	0.481	0.024	0.161	0.026	-0.164	0.013	0.473	0.012	0.158	0.014
Log10plus-linear	0.010	0.003	0.025	0.003	0.017	0.003	0.008	0.002	0.023	0.001	0.016	0.002
IHS	-0.021	0.017	0.237	0.017	0.103	0.018	-0.020	0.009	0.229	0.007	0.098	0.009

Panel B: $x_{it} = .5\mu_i + .5\nu_{it}$; $\sigma_\eta = \exp(\gamma\mu_i + (1 - \gamma)\nu_{it})$

Estimator:	Poisson		Poisson FE		Log-linear		Log-linear FE	
	β_1	S.E.	β_1	S.E.	β_1	S.E.	β_1	S.E.
$\gamma = 100\%$	0.199	0.052	0.200	0.062	-0.299	0.032	0.200	0.034
$\gamma = 75\%$	0.199	0.048	0.199	0.055	-0.300	0.027	-0.050	0.033
$\gamma = 50\%$	0.199	0.047	0.198	0.058	-0.300	0.025	-0.300	0.035
$\gamma = 25\%$	0.199	0.048	0.199	0.069	-0.300	0.025	-0.550	0.038
$\gamma = 0\%$	0.199	0.052	0.197	0.084	-0.301	0.029	-0.800	0.042

Table 3: Simulated OLS, Poisson, and Rate regression rejection rates

This table presents results from a series of simulations which compare the rejection rates of OLS, Poisson, and Rate regressions. We simulate a set of observations (x_1, x_2, y), where $E[y|x] = e^{\beta_1 x_1 + x_2}$. 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 $x_{1,it} = .5\mu_i + .5\nu_{it}$. both μ_i and ν_{it} are drawn from a truncate standard normal distribution. 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 -.3, -.1, .1 and .3.

Panels A, B, and C simulate discrete outcomes by using a negative binomial data generating process. Each group has an overdispersion parameter that is determined by the μ_i, where $\alpha_{i,NB} \in [.5 * \alpha_{sim,NB}, 1.5 * \alpha_{sim,NB}]$. $\alpha_{sim,NB}$, is equal to .001, 3, and 8, respectively.

Panels D, E, and F simulate continuous outcomes by using mixture model. The number of mixture components is determined by conditional mean multiplied by a lognormal random variable with a mean of 1 and variance of $E[y|x] + b * E[y|x]^2$, where b is .001, 3 and 8 in their respective panels. Each mixture component is a $\chi^2_{(1)}$ random variable. This implies that $V(y|x) = 3 * E[y|x] + b * E[y|x]^2$.

Panel A: $\alpha_{sim,NB} = 0.001$

	β =-.3	β =-.1	β =.1	β =.3
OLS	0.503	0.0864	0.0895	0.497
Poisson	1.000	1.000	1.000	1.000
Rate OLS	1.000	0.694	0.675	1.000

Panel B: $\alpha_{sim,NB} = 3$

	β =-.3	β =-.1	β =.1	β =.3
OLS	0.215	0.0529	0.0501	0.205
Poisson	0.729	0.247	0.241	0.700
Rate OLS	0.882	0.199	0.197	0.873

Panel C: $\alpha_{sim,NB} = 8$

	β =-.3	β =-.1	β =.1	β =.3
OLS	0.128	0.0374	0.0394	0.119
Poisson	0.467	0.187	0.198	0.448
Rate OLS	0.730	0.139	0.131	0.697

Panel D: $Var(y|x) = 3 * E[y|x] + 0.001 * E[y|x]^2$

	β =-.3	β =-.1	β =.1	β =.3
OLS	0.490	0.0915	0.0882	0.496
Poisson	1.000	0.959	0.952	1.000
Rate OLS	0.795	0.165	0.177	0.790

Panel E: $Var(y|x) = 3 * E[y|x] + 3 * E[y|x]^2$

	β =-.3	β =-.1	β =.1	β =.3
OLS	0.367	0.0718	0.0683	0.371
Poisson	0.882	0.605	0.604	0.883
Rate OLS	0.753	0.151	0.161	0.754

Panel F: $Var(y|x) = 3 * E[y|x] + 8 * E[y|x]^2$

	β =-.3	β =-.1	β =.1	β =.3
OLS	0.303	0.0613	0.0590	0.304
Poisson	0.776	0.473	0.475	0.777
Rate OLS	0.706	0.140	0.147	0.707

Table 4: Replication: Hirshleifer, Low, and Teoh (2012)

This table presents a series of regressions based on the regression specification in Table V column (1) 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-linear regression. The final three column present results from log1plus-linear, log-linear, and Poisson regressions, based on our attempt to replicate the original data set. 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.

	Actual	Replication	Replication	Replication	Replication
	Log1plus-Linear	Log1plus-Linear	Log-Linear	Poisson	Log1plus-Linear Poisson Sample
Overconfident CEO	0.093*	0.110**	0.196***	0.185*	0.112**
	(1.93)	(2.23)	(2.79)	(1.79)	(2.24)
Log(sales)	0.732***	0.446***	0.617***	0.921***	0.449***
	(16.23)	(16.88)	(19.20)	(11.79)	(16.89)
Log(PPE/Emp)	0.244***	0.169***	0.301***	0.390***	0.172***
	(4.76)	(4.74)	(4.59)	(2.92)	(4.73)
Observations	8,939	12,168	5,575	11,983	11,983
Adjusted R2	0.494	0.482	0.479		0.480

Table 5: 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-linear regression. The final three column present results from log1plus-linear, log-linear, and Poisson regressions, based on our attempt to replicate the original data set. 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.

	Actual	Replication	Replication	Replication	Replication
	Log1plus-Linear	Log1plus-Linear	Log-Linear	Poisson	Log1plus-Linear 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 6: 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-linear regression. The final three column present results from log1plus-linear, log-linear, and Poisson regressions, based on our attempt to replicate the original data set. 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. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Actual	Replication	Replication	Replication	Replication
	Log1plus-Linear	Log1plus-Linear	Log-Linear	Poisson	Log1plus-Linear 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 7: 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 final three column present results from log1plus-linear, log-linear, and Poisson regressions, based on our attempt to replicate the original data set. 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. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Actual	Replication	Replication	Replication	Replication
	Poisson	Log1plus-Linear	Log-Linear	Poisson	Log1plus-Linear 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 8: 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-linear regression. The final three column present results from log1plus-linear, log-linear, and Poisson regressions, based on our data. 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. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively, based on a two-tailed t-test.

	Actual Log1plus-Linear	Replication Log1plus-Linear	Replication Log-Linear	Replication Poisson	Replication Log1plus-Linear Poisson Sample
Bestfoods	0.047***	0.047***	0.119	-0.050	0.118**
	(3.24)	(3.24)	(1.53)	(-0.36)	(2.55)
Observations	501,259	501,259	61,510	182,454	182,454
Adjusted R2	0.541	0.541	0.570		0.448

Table 9: 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 (2020). 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 final three column present results from log1plus-linear, log-linear, and Poisson regressions, based on our attempt to replicate the original data set. 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.

	Actual	Replication	Replication	Replication	Replication
	Log-Linear	Log1plus-Linear	Log-Linear	Poisson	Log1plus-Linear 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

Table 10: Poisson and Log1plus-Linear Decomposition

This table decomposes the differences in the Poisson and log1plus-linear regression estimates reported in Tables 4 through 9. 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-linear 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-linear regression as the second column, but using the full sample. The fourth column reproduces the log1plus-linear 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-Linear ($\hat{y}$)	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear
Sample	Poisson	Poisson	Full	Full
Overconfident CEO	0.185	0.085	0.085	0.110
Observations	11,983	11,983	12,168	12,168
Controls, FEs	Yes	Yes	Yes	Yes

Panel B: He and Tian (2013)

Model	Poisson	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear
Sample	Poisson	Poisson	Full	Full
lnCoverage	0.026	-0.014	0.000	-0.026
Observations	15,857	15,857	15,857	27,064
Controls, FEs	Yes	Yes	Yes	Yes

Panel C: Fang, Tian, and Tice (2013)

Model	Poisson	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear
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-Linear ($\hat{y}$)	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear
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-Linear ($\hat{y}$)	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear
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-Linear ($\hat{y}$)	Log1plus-Linear ($\hat{y}$)	Log1plus-Linear
Sample	Poisson	Poisson	Full	Full
HM Debt	0.850**	0.835***	0.835***	0.255
	(0.351)	(0.006)	(0.006)	(0.175)
Observations	38,365	38,365	39,951	39,951
Controls, FEs	Yes	Yes	Yes	Yes