Received 17 May 2019
Revised 31 July 2019
Accepted 8 August 2019

Estimating Poisson pseudo-maximum-likelihood rather than log-linear model of a log-transformed dependent variable

Victor Motta[a,*]
[a]*Fundacao Getulio Vargas, Sao Paulo, Brazil*

Abstract

Purpose – The purpose of this study is to account for a recent non-mainstream econometric approach using microdata and how it can inform research in business administration. More specifically, the paper draws from the applied microeconometric literature stances in favor of fitting Poisson regression with robust standard errors rather than the OLS linear regression of a log-transformed dependent variable. In addition, the authors point to the appropriate Stata coding and take into account the possibility of failing to check for the existence of the estimates – convergency issues – as well as being sensitive to numerical problems.

Design/methodology/approach – The author details the main issues with the log-linear model, drawing from the applied econometric literature in favor of estimating multiplicative models for non-count data. Then, he provides the Stata commands and illustrates the differences in the coefficient and standard errors between both OLS and Poisson models using the health expenditure dataset from the RAND Health Insurance Experiment (RHIE).

Findings – The results indicate that the use of Poisson pseudo maximum likelihood estimators yield better results that the log-linear model, as well as other alternative models, such as Tobit and two-part models.

Originality/value – The originality of this study lies in demonstrating an alternative microeconometric technique to deal with positive skewness of dependent variables.

Keywords Health economics, Applied microeconometrics,
Poisson pseudo maximum likelihood estimator

Paper type Research paper

1. Introduction

Researchers in the different fields within business administration often estimate models with a log-transformed dependent variable. The main reasons for log transforming the outcome variable include dealing with a positively skewed variable as well as interpreting a covariate as either elasticity or having a multiplicative response (Manning, 1998). An unfortunate consequence of this approach, however, is that the estimated coefficients are relevant to the distribution of the log-transformed dependent variable rather than to the distribution of the dependent variable in their natural units. As a result, coefficients from the

Victor Motta was the only contributor to this paper.

RAUSP Manag. J.
Vol. 54 No. 4, 2019
pp. 508-518
Emerald Publishing Limited
2531-0488
DOI 10.1108/RAUSP-05-2019-0110

log-transformed ordinary least squares (OLS) model are often retransformed back to unlogged terms to make inferences in their natural units.

The retransformed estimate of either the conditional mean or the impact of an independent variable on the dependent variable – the slope – needs to adjust for both heteroskedasticity and the distribution of the residual (Mullahy, 1998). Failure to account for both may lead to biased estimates of the conditional mean and the slope on its original scale. The presence of heteroskedasticity can generate different estimates in log-linear models rather than estimated in levels. This suggests that inferences drawn on log-linear regressions may produce misleading conclusions.

Although suggestions have been offered in favor of estimating log-linear models to inform about the conditional mean of the distribution of the dependent variable, they rely on strong underlying assumptions that may not hold. Among the several models used to correct the issues of coefficient biasedness and heteroskedasticity in log-linear models, the Poisson pseudo-maximum-likelihood estimator is a robust substitute for the standard log-linear model (Silva & Tenreyro, 2006).

The purpose of this paper is to account for a recent non-mainstream econometric approach using microdata and how it can inform research in business administration. More specifically, the paper draws from the applied microeconometric literature stances in favor of fitting Poisson regression with robust standard errors rather than the OLS linear regression of a log-transformed dependent variable. In addition, we point to the appropriate Stata coding and take into account the possibility of failing to check for the existence of the estimates – convergency issues – as well as being sensitive to numerical problems.

The remainder of the paper proceeds as follow. Section 2 details the main issues with the log-linear model, while Section 3 draws from the applied econometric literature in favor of estimating multiplicative models for non-count data. Section 4 provides the Stata commands, while Section 5 illustrates the differences in the coefficient and standard errors between both OLS and Poisson models using the health expenditure dataset from the RAND Health Insurance Experiment (RHIE). Section 6 concludes the paper.

2. Main issues with log-linearized model

Jensen's inequality implies that $E(\text{In}\, y) \neq \text{In}\, E(y)$, that is, the expected value of the logarithm of a random variable is different from the logarithm of its expected value. An important implication of Jensen's inequality is that interpreting the parameters of log-linear models estimated by OLS as elasticities may be misleading in the presence of heteroskedastic. The use of the log-transformed dependent variable creates a potential bias when computing estimates of $E[y\,|\,x]$ on the original scale provided the residual term does not have a normal distribution or is heteroskedastic. As Silva and Tenreyro (2006) posit, estimating the log-linear model $\ln y_i = x_i^{'}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$, where x_i is a $K \times 1$ vector of regressors, $\boldsymbol{\beta}_i$ is a $K \times 1$ vector of coefficients and $\boldsymbol{\varepsilon}_i$ is a vector of residuals of each observation i, by OLS is inappropriate for several reasons.

First, log-linearization is not feasible if $y_i = 0$ since $\text{In}\ 0 = -\infty$. In addition, even if all observations of $y_i > 0$, the expected value of the log-linear residual will depend on the vector of covariates. Therefore, estimating by OLS will yield in inconsistent estimators. For instance, consider a model:

$$y_i = e^{x_i^{'}\beta_i}\,\boldsymbol{\eta}_i$$

where $\boldsymbol{\eta}_i = 1 + \frac{\varepsilon_i}{e^{x_i^{'}\beta_i}}$ and $E[\boldsymbol{\eta}_i|x] = 1$. Assuming $y_i > 0$, the model can be made linear in the parameters by taking logarithms of both sides of the equations. As a result, this yields to:

$$\ln y_i = x_i^{\prime}\beta_i + \ln \eta_i$$

To obtain a consistent estimator of the slope parameters of y_i estimating the log-linear equation above by OLS, it is necessary that $E\,[\ln\ \eta_i|x]$ does not depend on x_i. In addition, consistent estimation of the intercept also requires that $E[\ln\ \eta_i|x] = 0$ Since $\eta_i = 1 + \frac{\varepsilon_i}{e^{x_i^{\prime}\beta_i}}$, the aforementioned condition is only met if $\varepsilon_i = e^{x_i^{\prime}\beta_i}\upsilon_i$, where υ_i is a random variable statistically independent of x_i. In such case, $\eta_i = 1 + \upsilon_i$ implies that $Ea\,[\ln\ \eta_i|x]$ is constant and statistically independent of x_i. As a result, the log-linear model representation is useful to estimate the parameters of interest only under specific conditions on the error term.

Since $y_i > 0$, the probability of y_i approaches zero when $E(y_i|x_i)$ approaches zero. This implies that the conditional variance of y_i, $Var(y_i|x_i)$ tends to disappear as $E(y_i|x_i)$ approaches zero. However, it may be possible to observe large deviations from the conditional mean – thus leading to greater dispersion – when the expected value of y_i is far away from its lower bound. The residual term ε_i is likely heteroskedastic and its variance will depend on $e^{x_i^{\prime}\beta_i}$. As a result, regressing $\ln\ y_i$ on x_i by OLS will lead to inconsistent estimates of β. The main reason for heteroskedasticity affecting the consistency of an estimator is that the nonlinear transformation of the dependent variable changes the properties of the residual term. Unless strong assumptions are imposed on the distribution form, recovering information about the expectation of y_i from the conditional mean of $\ln\ y_i$ may not be possible since the logarithm of the residual term is correlated with the regressors. In general, even if all observations on $\underline{y}_i$ are positive, estimating β from the log-linear model by OLS will yield inconsistent estimators and heteroskedasticity across the regressors

3. Using the Poisson pseudo-maximum-likelihood estimator

A possible way of obtaining a more efficient estimator without resorting to non-parametric regression is to estimate the parameters of interest using a pseudo-maximum-likelihood estimator based on some assumption of the functional form of $Var(y_i\ |\ x_i)$ (Manning & Mullahy, 2001; Papke & Wooldridge, 1996). Among possible specifications, under the assumption that the conditional variance is proportional to the conditional mean, $E[y_i|x_i] = e^{x_i^{\prime}\beta_i} \propto Var(y_i|x_i)$ and β can be estimated by solving the following set of first-order conditions:

$$\sum_{i=1}^{n}\left[y_i - e^{x_i^{\prime}\tilde{\beta}}\right]x_i = 0$$

The estimator defined below is numerically equal to the Poisson pseudo-maximum-likelihood (PPML), often used for count data. The form of the equation implies that the correct specification of the conditional mean, $E[y_i|x_i] = e^{x_i^{\prime}\beta_i}$. Therefore, the data do not have to have a Poisson distribution (count data) and y_i does not have to be an integer in order for the estimator based on the Poisson likelihood function to be consistent (Gourieroux, Monfort, & Trognon, 1984).

The implementation of the pseudo-maximum-likelihood is estimated via Poisson regression even when the dependent variable is not an integer. However, because the assumption $Var(y_i|x_i) \propto E\{y_i|x_i\}$ is unlikely to hold, this estimator does not take full account of the heteroskedasticity in the model. As a result, the inference has to be based on an Eicker–White robust covariance estimator (Eicker, 1963; White, 1980).

The Poisson regression model is defined by:

$$\Pr(y_i = j|x_i) = \frac{e^{-\lambda}\lambda^j}{j!}, \ j = 0, 1, 2, \ldots$$

where λ is generally specified as $\lambda = e^{x_i'\beta} = e^{\beta_0+\beta_1 x_{1i}+\cdots}$. The vector of parameters of interest, β, can be estimated by maximizing the log-likelihood function given by:

$$\ln L(\beta) = \sum_{i=1}^{n}\left[-e^{x_i'\beta} + \left(x_i'\beta\right)y_i - \ln(y_i!)\right].$$

Poisson regression is not only the most widely used model for count data (Cameron & Trivedi, 1986), but it is also becoming increasingly popular to estimate multiplicative models for other kinds of data (Blackburn, 2007; Manning & Mullahy, 2001).

The reasons that make this estimator popular can be clearly understood by inspecting the corresponding score vector and Hessian matrix, given respectively below:

$$s(\beta) = \sum_{i=1}^{n}\left[y_i - e^{x_i'\beta}\right]x_i \,\text{and}\, H(\beta) = -\sum_{i=1}^{n} e^{x_i'\beta}x_i x_i'$$

The form of the score vector makes it possible that β will be consistently estimated as long as $E[y_i|x_i] = e^{x_i'\beta}$. For instance, the only condition required for consistency is the correct specification of the conditional mean. Since the estimator of the covariance matrix neither assumes equality between the mean value and the variance of the dependent variable, nor does it require constant variance, Poisson regression with the Huber-White-Sandwich linearized estimator of variance is a permissible alternative to log linear regression (Gourieroux *et al.*, 1984).

Running a Poisson regression with robust standard errors may be preferred to estimating a log-linear model by OLS. First, Poisson handles zero outcomes that arise in correspondence to the model. However, Poisson regression does not handle cases where some individuals participate, and others do not, and among the non-participating ones, they would likely product an outcome greater than 0 had they participated. For instance, Poisson does not handle zeros in a Mincerian income model (Mincer, 1958) since those that earned 0 did not participate in the labor force. Had they participated, their earnings might have been low, but they would be positive. More recent studies using the Poisson model with robust standard errors rather than log-linear regression have examined the impact of medical marijuana laws on addiction-related to pain killers (Powell, Pacula, & Jacobson, 2018), medical care spending and labor market outcomes (Powell & Seabury, 2018), innovation and production expenditure (Arkolakis *et al.*, 2018) and tourism and economic development (Faber & Gaubert, 2019), among many other studies.

4. Commands using Stata

This section briefly describes the Poisson commands in Stata, including some of its shortcomings.

OLS regressions of the algebraic form $\ln y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$ is usually coded using the following Stata command:

- generate lny = ln(y); and
- regress lny x1 x2 [. . .] xk.

Rather than estimating this log-linear model, we would instead fit a Poisson regression using the Huber-White-Sandwich linearized estimator of variance. In Stata this is done with the following command:

- poisson y x1 x2 . . . xk, vce(robust).

Note that there is no need to take the natural log of the dependent variable. The Poisson regression with robust standard errors specify that the variance-covariance matrix neither assumes $E(y_i) = Var(y_i)$, nor requires $Var(y_i)$ to be constant across all i. Therefore, the Poisson regression with robust standard errors (Huber-White-Sandwich linearized estimator of variance) is an alternative to log-linear regressions.

The estimator is also well-behaved since the Hessian is negative definite for all x and β. This facilitates the estimation and ensures the uniqueness of a maximum, conditional on its existence. As a result, estimation of β converges in a few iterations. However, the parameters in β are not identified by PPML for certain data configurations because they do not exist. The non-existence of PPML estimates are more likely when the data have a large number of zeros, such as the number of crimes committed, volume of trade between pairs of countries, among others (Silva & Tenreyro, 2011a). Since this type of identification problem has not been widely recognized as a major issue in count data models, Stata's Poisson command does not check for its presence.

In such cases, checking whether or not the results obtained actually correspond to a maximum of the log-likelihood function is recommended. We can check for this through the overfitting of the observation with $y_i = 0$ by computing descriptive statistics for the fitted values of y for the relevant sub sample. Silva and Tenreyro (2011b) identify and illustrate some shortcomings of the Poisson command in Stata. More specifically, they point out that the command fails to check for the existence of estimates and show that it is sensitive to numerical problems. The Poisson command does not check for the existence of the estimates and therefore, it is unable to identify whether convergence is not achieved or spurious.

In addition, even if maximum likelihood estimates of the Poisson regression exist, Stata may not correctly identify them due to its sensitivity to numerical problems of the algorithms available in the Poisson command in three situations: when the dependent variable has some very large values, when regressors are highly collinear and have different magnitudes, and when the covariates are highly (although no perfectly) collinear. A potential solution to explore when the maximum likelihood estimates exist but convergence is not achieved is to use different optimization methods offered in the Poisson command, such as the NR, BHHH, DFP and BFGS. One can also relax the convergence criteria and ensure convergence, by the algorithm may not deliver the desired maximum likelihood estimates.

A simple way to deal with the shortcomings of Stata's Poisson command is to use the *glm* command for the generalized linear model with the options family (Poisson) link(log) IRLS. The iterated reweighted least squares (IRLS) algorithm provided by the GLM command seems to be more stable than the algorithms in Poisson command and give the correct results, overcoming the command's limitations. To facilitate the estimation of Poisson regressions, the existence of the pseudo maximum likelihood estimates can be checked through the PPML command, offering methods to drop regressors that may cause the non-existence of the estimates. The command also warns if the variables have large values likely to create numerical problems. Estimation can be then implemented using the generalized linear model (GLM) method.

5. RAND health insurance experiment (RHIE health expenditure dataset

To illustrate the use of Poisson pseudo maximum likelihood rather than log-linear models, use data from the RAND Health Insurance Experiment (RHIE). The experiment, conducted by the RAND corporation from 1974 to 1982, has been the longest running and largest controlled social experiment in medical care research. The main goal of the experiment to assess how the patient's use of health services is affected by types of randomly assigned health insurance, including both fee-for-service and health maintenance organizations (HMOs). In the experiment, the data were collected from about 8,000 enrollees in 2,823 families, from six sites across the USA. Each family was enrolled in one of 14 different health insurance plans for either three or five years. The plans ranged from free care to 95 per cent coinsurance below a maximum dollar expenditure (MDE), and also included an assignment in a prepaid group practice. RHIE dataset consists of utilization, expenditures, demographic characteristics, health status and insurance status variables. The final sample consists of 20190 observations; each observation represents data for an experimental subject in a given year.

Several of the RHIE studies on health expenditures relies on regression models with logged dependent variables. With standard deviations two to four times the mean, the log transformation was essential to finding estimates of the response of health care expenditures that were robust to the skewness in the data (Duan, 1983). In several analyses, the residual errors indicated the presence of heteroskedasticity by insurance plan, the main covariate of interest.

The central point here is that we do not face the problem of endogenous treatment effect – the central causal parameter of interest in the study – since insurance plans are randomly assigned, not freely chosen by the participant. Data were collected from the enrollee's use of medical care services and health status throughout the randomly assigned term of enrollment for either three or five years. For additional details of the data, see Manning et al. (1987) and Deb and Trivedi (2002). The sample used in this study consists of second-year data for individuals in the fee-for-service plans only.

To illustrate the main issues, Table I reports the first four moment generating functions, mean, variance, skewness and kurtosis, as well as the percentiles. Medical expenditure is heavily skewed to the right and kurtotic. The standard deviation is four times the mean. In addition, the mean of $169.70 is much larger than the median of $32.38. As a result, using a natural logarithmic transformation of the dependent variable, medical expenditure, to perform a log-linear model has become the standard in both business and applied microeconomic work. Once the estimates from such a model are obtained, the usual practice is to interpret the response to a particular covariate as being the exponential of the

		Medical exp excl outpatient men		
(%)	Percentiles	Smallest		
1	0	0		
5	0	0		
10	0	0	Obs	5,575
25	3.849658	0	Sum of wgt.	5,575
50	32.37693		Mean	169.7003
		Largest	SD	802.7604
75	101.2285	12044.11		
90	330.9775	17465.98	Variance	644424.2
95	732.6303	18641.98	Skewness	27.03142
99	2232.54	39182.02	Kurtosis	1113.741

Table I. Descriptive statistics of medical expenditure

coefficient of that variable in the model. As regressors, we include health insurance variables, socioeconomic characteristics and heath-status variables. Table II contains the list of all regressors in our model.

Table III displays the descriptive statistics of the log-transformed medical expenses. The logarithmic transformation eliminates this skewness, with a mean of 4.07 close to the median of 3.96, and the skewness statistics falls from 27.03 to 0.35. The kurtosis is 3.29, close to the normal value of 3. Table IV displays the estimation outcomes resulting from various techniques. The first column reports OLS estimates using the logarithm of medical expenses as the dependent variable. As noted before, this regression leaves out individuals with no medical expenditure (about 23 per cent of the observations). The second column reports the OLS estimates using the logarithm transformation of 1 plus medical expenses, ln $(1 + \mathit{meddol})$, as the dependent variable to deal with the zeros. The fourth column uses reports Poisson estimates using only the subsample of positive medical expenditure while

Table II.
List of explanatory variables

Explanatory variable	Definition
logc	$\ln(\mathit{coinsurance} + 1)$, $0 \leq \mathit{coinsurance} \leq 100$
idp	1 if individual deductible plan, 0 otherwise
lpi	ln(max(1, *annual participation incentive payment*))
fmde	$0 \text{ if idp} = 1, \ln\left(\max\left(1, \frac{MDE}{0.01\ \mathit{coinsurance}}\right)\right)$ *otherwise*
linc	ln(*family income*)
lfam	ln (*family size*)
female	1 if person is a woman
child	1 if age is less than 18
fchild	Female*child
black	1 if race of household head is black
educdec	Education of the household head in years
physlim	1 if the person has a physical limitation
disea	Number of chronic diseases
hlthg	1 if self-rated health is good
hlthf	1 if self-rated health is fair
hlthp	1 if self-rated health is poor
	Omitted category is excellent self-rated health

Table III.
Descriptive statistics of the log-transformed medical expenditure

		Lnmeddol		
(%)	Percentiles	Smallest		
1	0.746548	−0.5343859		
5	1.749707	−0.4108706		
10	2.238203	−0.3899609	Obs	4,282
25	3.059381	−0.3899609	Sum of Wgt.	4,282
50	3.963396		Mean	4.069336
		Largest	SD	1.499219
75	4.915971	9.396331		
90	6.11767	9.76801	Variance	2.247659
95	6.807192	9.833171	Skewness	0.347961
99	7.888451	10.57597	Kurtosis	3.28978

Variable list	OLS	OLS2	Poisson y > 0	Poisson
logc	−0.0190 (0.0313)	−0.144*** (0.0371)	0.00791 (0.0563)	−0.0205 (0.0562)
idp	−0.0777 (0.0618)	−0.200*** (0.0721)	−0.0200 (0.141)	−0.0704 (0.132)
lpi	0.00433 (0.00970)	0.0344*** (0.0118)	0.0289 (0.0176)	0.0382** (0.0177)
fmde	−0.0297 (0.0181)	−0.0118 (0.0219)	−0.0290 (0.0339)	−0.0276 (0.0348)
linc	0.101*** (0.0216)	0.125*** (0.0238)	0.133 (0.0847)	0.168* (0.0928)
lfam	−0.159*** (0.0456)	−0.146*** (0.0554)	−0.213 (0.169)	−0.223 (0.170)
female	0.334*** (0.0570)	0.732*** (0.0708)	−0.0660 (0.178)	0.0652 (0.179)
child	−0.416*** (0.0676)	−0.186** (0.0813)	−0.767*** (0.182)	−0.731*** (0.183)
fchild	−0.340*** (0.0896)	−0.738*** (0.108)	0.153 (0.236)	0.0202 (0.240)
black	−0.194*** (0.0677)	−0.853*** (0.0758)	−0.100 (0.141)	−0.284** (0.144)
educdec	−0.00265 (0.00820)	0.0353*** (0.0101)	0.0284 (0.0275)	0.0376 (0.0281)
disea	0.0215*** (0.00339)	0.0395*** (0.00430)	0.0122** (0.00592)	0.0172*** (0.00612)
physlm	0.276*** (0.0685)	0.461*** (0.0886)	0.477*** (0.141)	0.513*** (0.140)
hlthg	0.151*** (0.0483)	0.160*** (0.0588)	0.198* (0.119)	0.224* (0.120)
hlthf	0.383*** (0.0878)	0.497*** (0.108)	0.522** (0.224)	0.588*** (0.227)
hlthp	0.817*** (0.170)	1.221*** (0.223)	1.579*** (0.529)	1.711*** (0.541)
_cons	3.242*** (0.211)	1.496*** (0.243)	3.863*** (0.858)	3.123*** (0.926)
N	4281	5574	4281	5574

Notes: ***Significance at 0.01; **significance at 0.05 and *significance at 0.1

Table IV. Estimation outcomes from various techniques

the last column shows the Poisson results for the whole sample (including observations with zero medical expenditure).

The main point to notice is that the Poisson estimated coefficients are similar using the entire sample and using the positive expenditure sample only. However, most coefficients differ from those obtained using a log-linear model. This suggests that in this case, heteroskedasticity may be responsible for the differences in the results between Poisson with robust standard errors and those of OLS (Wooldridge, 2010). Further evidence using the Breusch–Pagan/Cook–Weisberg test for heteroscedasticity, rejects the hull hypothesis of homoskedasticity ($\chi^2 = 17.81$ p-value = 0.0000).

In addition, we have included the distribution of residuals for all four models. Figures 1 and 2 display the quantiles of residuals against the quantiles of the normal distribution. For both Poisson models, we used deviance residuals since they have the best properties for examining the goodness of fit of Generalized Linear Models, such as a Poisson family. The results indicate that Poisson provides a better fit.

5.1 Alternative models: Tobit and two-part models

Alternatively, other models could be considered, such as the Tobit, and two-part or hurdle models. The Tobit model could be considered since medical expenditures are left-censored at zero. For instance, 23 per cent of the observations had no medical expenditure for year 2. A potential approach would be to put a small number a for every zero (smaller than the smallest observed positive y), take the log and then specify ln a as the left-censoring point (Cameron & Trivedi, 2005). However, the choice of a is arbitrary and affects the estimation. For instance, choosing $a = 0.01$ results in $\ln y = -4.6$ and choosing $a = 0.000001$ results in $\ln y = -13.8$, and there is not any clear reason to prefer one over the other when the smallest positive y is 1. In addition, the Tobit model has strong assumptions of normality and homoscedasticity. If these assumptions fail, then the Tobit maximum likelihood estimator is not robust. Tobit also assumes that a single mechanism drives the two dimensions of the expenditure data.

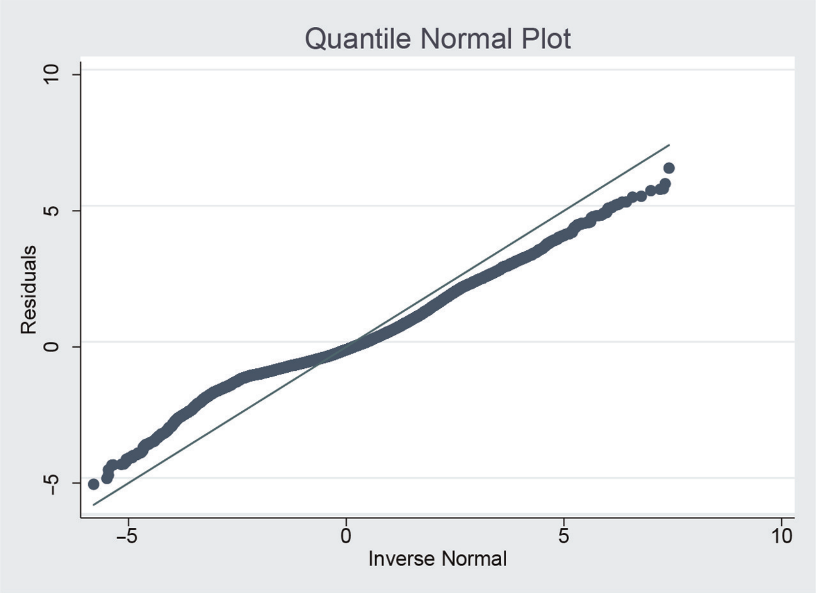

Figure 1.
Quantiles of the residuals plotted against the quantiles of the normal distribution for OLS regression

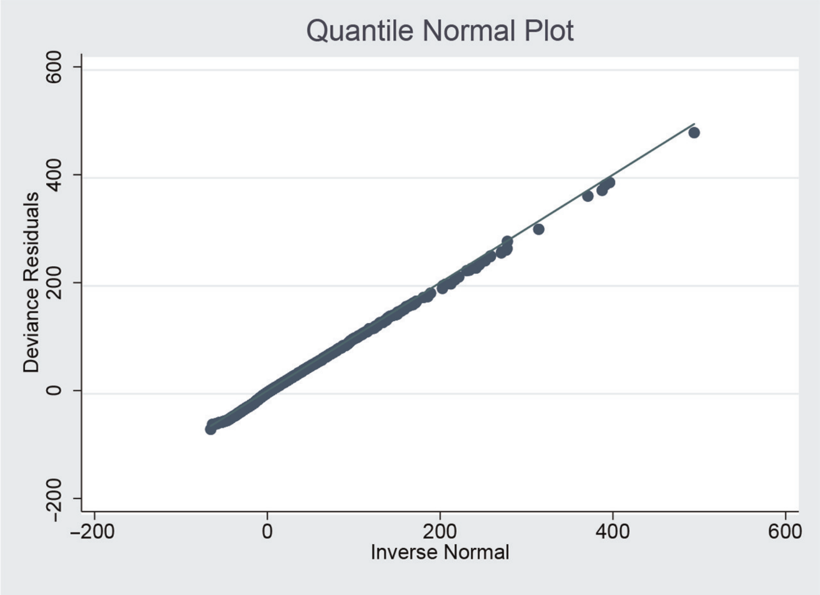

Figure 2.
Quantiles of the residuals plotted against the quantiles of the normal distribution for Poisson

To relax the latter assumption and to investigate if there is indeed a single mechanism we can use hurdle or two-part models, described by Mullahy (1986). The model involves estimating two separate regressions: the first models the probability that y is positive, while the second models the amount of y if y is positive. Using our RHIE dataset, for example, the idea is that a person decides whether to go to the doctor and then the doctor decides the expenditure conditional on $y > 0$. As a result, the first model can be fitted using a probit (or logit, complementary log log, etc.) using 1 ($y > 0$) as a dummy outcome, then run OLS regression ln y on the vector of regressors, or a truncated regression of y on the vector of regressors (Cragg, 1971; McDowell, 2003).

Unlike the Tobit model, the two-part model features two residuals: v, which impacts the decision to set $y > 0$ instead of $y = 0$, and u, which impacts y conditional on positive y. An important assumption underlying the two-part model is that v and u are independent. In other words, the unobservables which affect the decision to go to the doctor are independent of the unobservables that affect the decision of how much to spend. A potential drawback in using two-part models is that it may be difficult to include endogenous explanatory variables without strong maximum likelihood assumptions. In addition, a two-step assumption, in this case, may not be all too realistic since one may find herself getting medical care without any decision on her part, or one can also end her medical care provided she chose too. As a result, we would need more than two steps of the model to be correctly specified, or all the estimates would be inconsistent.

6. Conclusion

Coefficients from the log-transformed ordinary least squares (OLS) model are often retransformed to unlogged terms to make inferences in their natural units. Failure to account for adjustments for heteroskedasticity and normality of residuals may lead to biased estimates of the conditional mean and the slope on its original scale. This suggests that inferences drawn on log-linear regressions may produce misleading conclusions. Among the several models used to correct the issues of coefficient biasedness and heteroskedasticity in log-linear models, the Poisson pseudo-maximum-likelihood. This study drew from the applied microeconometric literature in favor of fitting Poisson regression with robust standard errors rather than the OLS linear regression of a log-transformed dependent variable. We applied both models in a health expenditure dataset to show the main differences.

References

Arkolakis, C., Ramondo, N., Rodríguez-Clare, A., & Yeaple, S. (2018). Innovation and production in the global economy. *American Economic Review*, *108*(8), 2128–2173.

Blackburn, M. L. (2007). Estimating wage differentials without logarithms. *Labour Economics*, *14*, 73–98. https://doi.org/10.1016/j.labeco.2005.04.005

Cameron, A., & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests author(s). *Journal of Applied Econometrics*, *1*, 29–53. Retrieved from www.jstor.org/stable/2096536

Cameron, A., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*, Cambridge, MA: Cambridge University Press.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* , *39*, 829. https://doi.org/10.2307/1909582

Deb, P., & Trivedi, P. (2002). The structure of demand for health care: Latent class versus two-part models. *Journal of Health Economics*, *21*, 601–625. Retrieved from http://ac.els-cdn.com/S0167629602000085/1-s2.0-S0167629602000085-main.pdf?_tid=7bb7ac5e-8bbc-11e6-a110-00000aacb35f&acdnat=1475755412_d51c5cdad378a09490c51fcd0a4798ac

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, *78*, 605–610. https://doi.org/10.1080/01621459.1983.10478017

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, *34*, 447–456. https://doi.org/10.1214/aoms/1177704156

Faber, B., & Gaubert, C. (2019). Tourism and economic development: Evidence from Mexico's coastline. *American Economic Review*, *109*(6), 2245–2293.

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

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, *17*, 283–295. https://doi.org/10.1016/S0167-6296(98)00025-3

Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform?. *Journal of Health Economics*, *20*, 461–494. https://doi.org/10.1016/S0167-6296(01)00086-8

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Manning, B. W. G., Newhouse, J. P., . . . Marquis, M. S. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *Medical Care*, *77*, 251–277.

McDowell, A. (2003). From the help desk: Hurdle models. *The Stata Journal: Promoting Communications on Statistics and Stata*, *3*, 178–184. https://doi.org/10.1177/1536867x0300300207

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, *66*, 281–302. https://doi.org/10.1086/258055

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, *33*, 341–365. https://doi.org/10.1016/0304-4076(86)90002-3

Mullahy, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, *17*, 247–281. https://doi.org/10.1016/S0167-6296(98)00030-7

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional variables with an application to 401 (K) Plan participation rates. *Journal of Applied Econometrics*, *11*, 619–632.

Powell, D. & Seabury, S. (2018). Medical care spending and labor market outcomes: Evidence from workers' compensation reforms. *American Economic Review*, *108*(10), 2995–3027.

Powell, D., Pacula, R. L., & Jacobson, M. (2018). Do medical marijuana laws reduce addictions and deaths related to pain killers?. *Journal of Health Economics*, *58*, 29–42.

Silva, J. M. C. S., & Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics*, *88*, 641–658. https://doi.org/\url{10.1162/rest.88.4.641}

Silva, J. M. C. S., & Tenreyro, S. (2011a). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters*, *112*, 220–222. https://doi.org/10.1016/j.econlet.2011.05.008

Silva, J. M. S., & Tenreyro, S. (2011b). Poisson: Some convergence issues. *The Stata Journal*, *11*(2), 207–212.

White, H. (1980). A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817. https://doi.org/10.2307/1912934

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

***Corresponding author**
Victor Motta can be contacted at: victor.motta@fgv.br