Data Visualization and Health Econometrics

Andrew M. Jones
Department of Economics and Related Studies
University of York, Heslington, York
United Kingdom
andrew.jones@york.ac.uk

the essence of knowledge
Boston — Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. M. Jones. *Data Visualization and Health Econometrics*. Foundations and Trends® in Econometrics, vol. 9, no. 1, pp. 1–78, 2017.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-318-8

Foundations and Trends® in Econometrics

Volume 9, Issues 1, 2017

Editorial Scope

Topics

Foundations and Trends® in Econometrics publishes survey and tutorial articles in the following topics:

- Econometric models
- Simultaneous equation models
- Estimation frameworks
- Biased estimation
- Computational problems
- Microeconometrics
- Treatment modeling
- Discrete choice modeling
- Models for count data
- Duration models
- Limited dependent variables
- Panel data
- Time series analysis
- Latent variable models
- Qualitative response models
- Hypothesis testing
- Econometric theory
- Financial econometrics
- Measurement error in survey data
- Productivity measurement and analysis
- Semiparametric and nonparametric estimation
- Bootstrap methods
- Nonstationary time series
- Robust estimation

Information for Librarians

Foundations and Trends® in Econometrics, 2017, Volume 9, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Foundations and Trends® in Econometrics
Vol. 9, No. 1 (2017) 1–78

DOI: 10.1561/0800000033

Data Visualization and Health Econometrics

Andrew M. Jones
Department of Economics and Related Studies
University of York, Heslington, York
United Kingdom
andrew.jones@york.ac.uk

Contents

Abstract

This article reviews econometric methods for health outcomes and health care costs that are used for prediction and forecasting, risk adjustment, resource allocation, technology assessment, and policy evaluation. It focuses on the principles and practical application of data visualization and statistical graphics and how these can enhance applied econometric analysis. Particular attention is devoted to methods for skewed and heavy-tailed distributions. Practical examples show how these methods can be applied to data on individual healthcare costs and health outcomes. Topics include: an introduction to data visualization; data description and regression; generalized linear models; flexible parametric models; semiparametric models; and an application to biomarkers.

A. M. Jones. *Data Visualization and Health Econometrics*. Foundations and Trends® in Econometrics, vol. 9, no. 1, pp. 1–78, 2017.

DOI: 10.1561/0800000033.

1

Introduction

Econometric models for health outcomes and health care costs are used for prediction and forecasting in health care planning, risk adjustment by insurers and public providers of health care, geographic resource allocation, health technology assessment, and health policy impact evaluations. Methods for risk adjustment focus on predicting the treatment costs for particular types of patient, often with very large survey or administrative data sets.

Microdata for individual medical expenditures and costs of treatment are typically non-normal. Survey data often feature a spike at zero, if there are non-users in the data. Both survey and administrative data, such as registers and discharge records, typically have a heavily skewed distribution and heavy tails. The spike at zero is often modeled by a two-part specification, with a binary choice model for the probability of any costs, and a conditional regression model for the positive costs [Jones, 2000]. Due to the skewness and excess kurtosis of the data and the importance of influential observations, regression models applied directly to the raw data on the level of costs can perform poorly. Traditionally the positive observations have been transformed prior to fitting a regression model, most often by taking a logarithmic or, sometimes, a square root transformation. Once these models have been fitted then predictions have to be retransformed back to the original — raw cost — scale. This is not straightforward to do in a robust way, especially if there is heteroskedasticity in the data on the transformed scale [Manning, 1998, Manning and Mullahy, 2001, Mullahy, 1998].

In the recent literature, attention has shifted away from linear regression models to semiparametric and flexible parametric estimators. A popular semiparametric approach is to use generalized linear models (GLMs) [e.g., Buntin and Zaslavsky, 2004, Manning and Mullahy, 2001, Manning et al., 2005, Manning, 2006]. GLMs are built around a *link function* that specifies the relationship between the conditional mean and a linear function of the covariates and a *distributional family* that specifies the form of the conditional variance as a function of the conditional mean. GLM models are estimated using a quasi-likelihood approach derived from the quasi-score or "estimating equations."

In a conventional GLM the choice of link and distribution has to be specified a priori. In practice the most frequently used GLM specification for medical costs has been the log-link with a gamma variance [Blough et al., 1999, Manning and Mullahy, 2001, Manning et al., 2005]. Basu and Rathouz [2005] have developed a flexible semiparametric approach to the problem of selecting the appropriate link and variance functions. Their extended estimating equations estimator (EEE) approach uses a Box–Cox transformation for the link function and either a power variance or quadratic variance function for the distribution. The particular form of the link and distribution are thereby estimated from the data at hand.

Other semiparametric methods that have appeared in the literature on modeling health care costs include the conditional density estimator and finite mixture models. The conditional density approach was advocated by Gilleskie and Mroz [2004] and divides the support of the distribution of the dependent variable into discrete intervals then applies discrete hazard models to these, implemented in practice as a series of sequential logit models. Finite mixture models use a discrete mixture of parametric models and, for example, have been applied to medical costs by Conway and Deb [2005]. Combining simple distributions such as the gamma or log-normal in a mixture of relatively few components may approximate complex empirical distributions effectively, especially for distributions that are multimodal.

In contrast to semiparametric methods, flexible parametric methods fully specify the distribution for health care costs. Building on standard distributions such as the log-normal and gamma distributions, they move to more flexible three and four-parameter distributions such as the generalized gamma and the generalized beta distributions of the second kind (GB2). This provides the additional flexibility to fit the high level of skewness and the heavy tails seen in cost data [Jones et al., 2014]. The downside of this flexibility is a risk of over-fitting and, in practice, these approaches may be best used as a guide to selecting one of the special or limiting cases that are nested within the general models. In this respect the flexible parametric models can play a similar role to using the EEE approach to select the link and distribution functions to be used in a GLM.

Earlier literature reviews have synthesized and compared the wide range of approaches to modeling health care costs [e.g., Hill and Miller, 2010, Jones, 2000, 2011, Jones et al., 2013, Mullahy, 2009]. In addition, studies using a quasi-Monte Carlo design, based on English administrative data for patient level costs of hospital care, have provided an assessment of the relative performance of these approaches [Jones et al., 2014, 2015, 2016]. To complement these earlier studies, this article focuses on the principles and practice of data visualization and statistical graphics and how these can enhance empirical analysis of health care costs and outcomes, especially for skewed and heavy-tailed distributions. The scope of this review is limited to non-normal but continuous outcomes such as health care costs and biomarkers. Many health economics applications deal with categorical and ordered outcomes, count data, or duration data. Methods for these are reviewed in Jones [2000] and Jones et al. [2013]. The methods and applications used here are limited to cross-sectional data. For discussions of methods for panel data see Jones [2009] and for the use of cohort data Von Hinke Kessler Scholder and Jones [2015].

Practical examples show how these graphical methods can be applied using the software package Stata, which is widely used in applied econometrics. Stata is not the obvious software of choice for specialist work in data visualization especially for users who wish to present their work online and to make use of animation or interactivity. Nevertheless, for many applied econometricians it is the workhorse for data management and econometric analysis. In this article Stata code, shown in the font `courier new`, is included to show how far it is possible to go within Stata so that graphical analysis can be integrated with statistical and econometric analysis within one piece of software and using one set of syntax.

The review of methods that have been developed for health care cost regressions is complemented by an empirical case study that focuses on objectively measured health outcomes, whose distributions share many of the features of cost data. The case study applies the econometric and graphical methods to blood-based biomarkers as the dependent variables. The data set is the UK Household Longitudinal Study (UKHLS), known as Understanding Society, which is a large nationally representative longitudinal study [Benzeval et al., 2016].

References

A. Basu and P. J. Rathouz. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*, 6:93–109, 2005.

M. Benzeval, A. Davillas, M. Kumari, and P. Lynn. *Understanding Society — UK Household Longitudinal Study: Biomarker User Guide and Glossary.* Colchester: University of Essex, 2014.

M. Benzeval, M. Kumari, and A. M. Jones. How do biomarkers and genetics contribute to understanding society? *Health Economics*, 25:1219–1222, 2016.

D. K. Blough, C. W. Madden, and M. C. Hornbrook. Modeling risk using generalized linear models. *Journal of Health Economics*, 18:153–171, 1999.

M. B. Buntin and A. M. Zaslavsky. Too much ado about two-part models and transformation?: comparing methods of modeling medicare expenditures. *Journal of Health Economics*, 23:525–542, 2004.

A. Cairo. *The Functional Art.* New Riders, 2012.

A. Cairo. *The Truthful Art.* New Riders, 2016.

J. Camoes. *Data at Work.* New Riders, 2016.

V. Chernozhukov, I. Fernandez-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81:2205–2268, 2013.

W. S. Cleveland. *The Elements of Graphing Data.* Wadsworth Advanced Books and Software, 1985.

W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984.

K. S. Conway and P. Deb. Is prenatal care really ineffective? or, is the 'devil' in the distribution? *Journal of Health Economics*, 24:489–513, 2005.

A. Davillas, A. M. Jones, and M. Benzeval. The income-health gradient: Evidence from self-reported health and biomarkers using longitudinal data on income. *HEDG Working Paper WP 17/04*, 2017.

P. Deb and A. M. Holmes. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics*, 9:475–489, 2000.

P. Deb and P. K. Trivedi. Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12:313–336, 1997.

S. Few. *Now you see it. Simple visualization techniques for quantitative analysis*. Analytics Press, 2009.

S. Few. *Show Me the Numbers*. Analytics Press, 2nd edition, 2012.

S. Few. *Information Dashboard Design*. Analytics Press, 2nd edition, 2013.

S. Few. *Signal. Understanding What Matters in a World of Noise*. Analytics Press, 2015.

S. Firpo, N. M. Fortin, and T. Lemieux. Unconditional quantile regressions. *Econometrica*, 77:953–973, 2009.

S. Foresi and F. Peracchi. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90: 451–466, 1995.

D. B. Gilleskie and T. A. Mroz. A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics*, 23: 391–418, 2004.

A. Han and J. Hausman. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics*, 5:1–28, 1990.

R. L. Harris. *Information Graphics. A Comprehensive Illustrated Reference*. Oxford University Press, 1999.

S. C. Hill and G. E. Miller. Health expenditure estimation and functional form: Application of the generalized gamma and extended estimating equation models. *Health Economics*, 19:608–627, 2010.

A. Holly. *Modelling risk using fourth order pseudo maxiumum likelihood methods.* Institute of Health Economics and Management (IEMS), University of Lausanne: Switzerland, 2009.

K. Imai and D. A. Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99:854–866, 2004.

S. P. Jenkins. Distributionally-sensitive inequality indices and the gb2 income distribution. *The Review of Income and Wealth*, 55:392–398, 2009.

E. Johnson, F. Dominici, M. Griswold, and S. L. Zeger. Disease cases and their medical costs attributable to smoking: An analysis of the national medical expenditure survey. *Journal of Econometrics*, 112:135–151, 2003.

A. M. Jones. Health econometrics. In A. J. Culyer and J. P. Newhouse, editors, *Handbook of Health Economics.* Elsevier, 2000.

A. M. Jones. Panel data methods and applications to health economics. In T. C. Mills and K. Patterson, editors, *Palgrave Handbook of Econometrics*, volume 2. Palgrave MacMillan, 2009.

A. M. Jones. Models for health care. In D. Hendry and M. Clements, editors, *Oxford Handbook of Economic Forecasting.* Oxford University Press, 2011.

A. M. Jones, N. Rice, T. Bago d'Uva, and S. Balia. *Applied Health Economics.* Routledge, 2nd edition, 2013.

A. M. Jones, J. Lomas, and N. Rice. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, 29:649–670, 2014.

A. M. Jones, J. Lomas, and N. Rice. Healthcare cost regressions: going beyond the mean to estimate the full distribution. *Health Economics*, 24:1192–1212, 2015.

A. M. Jones, J. Lomas, P. Moore, and N. Rice. A quasi-monte carlo comparison of recent developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to health care costs. *Journal of the Royal Statistical Society Series A*, 179: 951–974, 2016.

G. Knies, editor. *Understanding Society — UK Household Longitudinal Study: Wave 1–5, 2009–2014, User Manual.* Colchester: University of Essex, 2015.

J. A. F. Machado and J. Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20:445–465, 2005.

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

W. Manning. Dealing with skewed data on costs and expenditure. In A. M. Jones, editor, *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar, 2006.

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

W. G. Manning, A. Basu, and J. Mullahy. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, 24:465–88, 2005.

J. B. McDonald, J. Sorensen, and P. A. Turley. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth*, 59: 360–374, 2011.

S. L. McFall, J. Petersen, O. Kaminska, and P. Lynn. *Understanding Society — UK Household Longitudinal Study: Waves 2 and 3 Nurse Health Assessment, 2010-2012, Guide to Nurse Health Assessment*. Colchester: University of Essex, 2014.

B. Melly. Decomposition of differences in distribution using quantile regression. *Labour Economics*, 12:577–590, 2005.

M. N. Mitchell. *A Visual Guide to Stata Graphics*. Stata Press, 2012.

J. Mullahy. Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17: 247–281, 1998.

J. Mullahy. Econometric modeling of health care costs and expenditures. a survey of analytical issues and related policy considerations. *Medical Care*, 47:S104–S108, 2009.

K. C. Nussbaumer. *Storytelling with Data. A Data Visualization Guide for Business Professionals*. Wiley, 2015.

N. B. Robbins. *Creating More Effective Graphs*. Wiley, 2005.

D. B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research*, 2:169–188, 2001.

J. A. Schwabish. An economist's guide to visualizing data. *Journal of Economic Perspectives*, 28:209–234, 2014.

E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1st edition, 1983.

E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.

E. R. Tufte. *Visual Explanations*. Graphics Press, 1997.

E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd edition, 2001.

E. R. Tufte. *Beautiful Evidence*. Graphics Press, 2006.

S. Von Hinke Kessler Scholder and A. M. Jones. Cohort data in health economics. In B. Baltagi, editor, *Oxford Handbook of Panel Data*. Oxford University Press, 2015.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2011.

N. Yau. *Visualize This*. Wiley, 2011.

N. Yau. *Data Points*. Wiley, 2013.