Poisson-based expectile regression for non-negative data with a mass-point at zero

Jeff H. Bergstrand
University of Notre Dame

Matthew W. Clance
University of Pretoria

J.M.C. Santos Silva
University of Surrey

31st UK Stata Conference

12 September 2025

OLS and quantiles

- Consider a linear model of the form

$$y_i = x_i'\beta + \varepsilon_i,$$

  where the error ε_i is not independent of x_i.

- The standard way to learn about the effect of x_i on y_i is to assume that $\mathrm{E}\left[y_i|x_i\right]$ is linear and estimate the parameters by **least squares**

$$\hat{\beta} = \arg\min_b \frac{1}{n} \sum \left(y_i - x_i'b\right)^2 .$$

- We can also estimate **conditional quantiles**, $\mathrm{Q}_{y_i}\left[\alpha|x_i\right]$, using the method introduced by Koenker and Bassett (1978)

$$\tilde{\beta}\left(\alpha\right) = \arg\min_b \frac{1}{n} \left\{\sum\nolimits_{y_i \geq x_i'b} \alpha \left|y_i - x_i'b\right| + \sum\nolimits_{y_i < x_i'b} \left(1-\alpha\right) \left|y_i - x_i'b\right|\right\} .$$

- **Median regression** is a special case.
- Quantiles are **local** measures of location that depend on only on the properties of the distribution around the relevant quantile.

Expectiles: Introduction

- Newey and Powell (1987) introduced **expectile regressions**, whose parameters can be estimated by solving

$$\hat{\beta}(\tau) = \arg\min_{b} \frac{1}{n} \left\{ \sum_{y_i \geq x_i' b} \tau \left(y_i - x_i' b\right)^2 + \sum_{y_i < x_i' b} (1-\tau) \left(y_i - x_i' b\right)^2 \right\}.$$

- **Mean regression** (OLS) is a special case when $\tau = 0.5$.
- For any $\tau \in (0, 1)$, the expectile τ of x, denoted $\mathrm{E}_x(\tau)$, **can be interpreted** as the expectation of x in a population where values of x above the expectile occur $\tau / (1-\tau)$ times as often as they do in the population of interest.
- An **analogous** results holds for quantiles.
- Unlike most estimators, here the **estimator defines** the object being estimated.

Expectiles: Properties

- In the **unconditional** case, each expectile corresponds to a quantile, and vice-versa.
- However, except in special cases, there is **no correspondence** between conditional expectiles and **conditional** quantiles.
- **Like** quantiles, expectiles provide information on the **location of different regions** of the distribution of a variable.
- In **contrast** to quantiles, expectiles are **global** measures of location that depend on global properties of the distribution.
- Admittedly, the **interpretation** of expectiles is not as intuitive as that of quantiles.
- In general, expectiles have **no advantage** over quantiles and Roger Koenker's (2013) view is that "*Expectiles belong in the spittoon*."

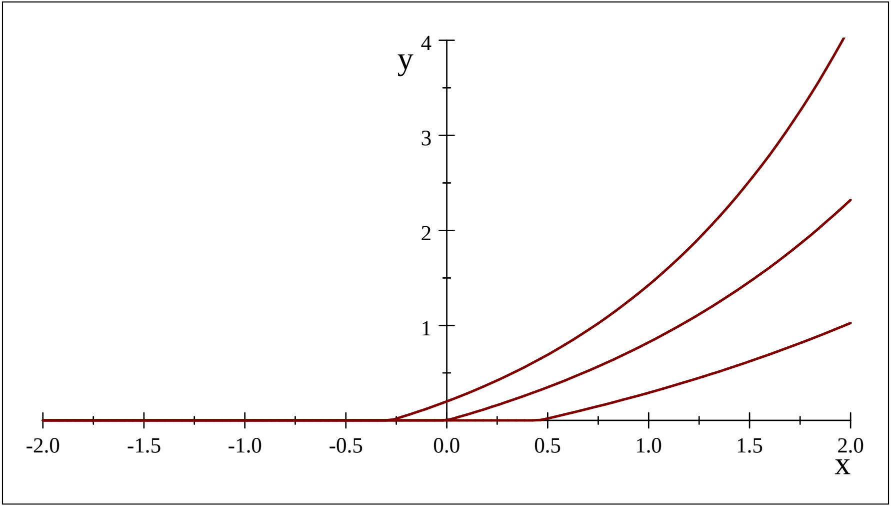
Quantiles vs. expectiles with non-negative data

- In many applications, the variable of interest takes only **non-negative** values and there is a mass-point at zero.
- We will look at the labour supply of married women (average hours per week) from the 1987 wave of PSID; this sample was used by Lee (1995).
- The table below displays some quantiles and expectiles for these data.

θ	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
$Q_x(\theta)$	0	0	0	0	25.1	37.4	40.1	40.1	53.9
$E_x(\theta)$	0.8	3.6	6.6	13.3	21.8	29.6	35.3	35.3	45.8

- Expectiles **smooth out** the mass point at zero.
- Because they are global measures of location, expectiles are **always positive**.

The trouble with quantiles with zeros



Setup and notation

- We consider a standard exponential model, typically used for this kind of data

$$y_i = \exp\left(x_i'\beta\right)\eta_i,$$

  where

  - y_i denotes the outcome of interest,
  - x_i is a vector of explanatory variables,
  - β is a conformable vector of parameters,
  - and η_i is a non-negative error term such that $\mathrm{E}\left(\eta_i|x_i\right) = 1$.

- Therefore $\mathrm{E}\left[y_i|x_i\right] = \exp\left(x_i'\beta\right)$.
- $\mathrm{E}\left(\eta_i|x_i\right) = 1$ but other features of its distribution may depend on x_i.
- In particular, η_i is generally **heteroskedastic**.

Expectiles

- As in Bergstrand, Clance and Santos Silva (2025), we **assume** that the τ-th conditional expectile of η_i has the form

$$\mathrm{E}_{\eta}\left(\tau|y_i\right) = \exp\left(x_i'\delta\left(\tau\right)\right).$$

  - The **exponential** function is used because all expectiles of η_i are positive.
  - The parameters are indexed by τ because they **vary** across expectiles.

- This setup implies that the **conditional expectiles** of y_i have the form

$$\mathrm{E}_{y}\left(\tau|x_i\right) = \exp\left(x_i'\beta\left(\tau\right)\right),$$

  with $\beta\left(\tau\right) = \beta + \delta\left(\tau\right)$.

  - x_i affects both the **mean** and the **dispersion** of y.
  - If η_i is **independent** of x_i, only the intercept changes with τ and all expectiles are proportional to each other.

The APPML estimator

- To estimate exponential expectiles we can use Efron's (1992) **asymmetric Poisson maximum likelihood estimator (APPML).**
- The estimator was intended for **count data** but can be used for other kinds of non-negative data.
- The APPML estimator of $\beta(\tau)$ based on a sample $\{(y_i, x_i)\}$ is the solution to moment conditions of the form:

$$\sum_{i=1}^{n} \omega_i \left(y_i - \exp\left(x_i'\hat{\beta}(\tau)\right)\right) x_i = 0,$$

  with

$$\omega_i = \left|\tau - \mathbf{1}\left(y_i < \exp\left(x_i'\hat{\beta}(\tau)\right)\right)\right|.$$

- This a Poisson regression that gives different **weights** to observations above or below the estimated expectile.
- The `appmlhdfe` command (Clance and Santos Silva, 2025) implements this estimator.

appmlhdfe

- `appmlhdfe` is based on the powerful `ppmlhdfe` command by Correia et al. (2019) and shares many of its functionalities.

Syntax

```
appmlhdfe depvar [indepvars] [if] [in] [, options]
```

`expectile(#)`: estimates # expectile; default is expectile(.5), which corresponds to Poisson regression.

`absorb`(varlist): categorical variables to be absorbed (fixed effects).

`residual`(varname): saves the residuals as varname.

`start`(varname): vector of residuals to be used as starting values.

Illustration

- Data on **labour supply** of married women (average hours per week) from the 1987 wave of PSID as used by Lee (1995).
- The independent variables are:
  - **education** in years (educ),
  - **age**, in years
  - number of **children** by age group (pkid, skid, hkid),
  - **race** (0 if white, 1 otherwise),
  - **home** (1 if owner, 0 otherwise),
  - **mort** (1 if mortgage on home, 0 otherwise),
  - husband's **occupation** dummies (manager, clerical, farmer),
  - local **unemployment** rate in percentage points (ur).
- We will ignore the **upper bound** on the number of hours per week, and estimate exponential models.

Results I

```
. appmlhdfe hours edu, a(age pkid skid hkid black ownh mort manager ///
> clerical farmer ur)

 Number of obs = 3373
 Iterations = 1
 Tolerance = 1.000e-07
 Objective function = 0
 % of negative residuals = .482
 R-squared: .19880988
.5 expectile regression
```

hours	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
edu	.0476741	.0064172	7.43	0.000	.0350965	.0602516
_cons	2.532422	.085541	29.60	0.000	2.364765	2.700079

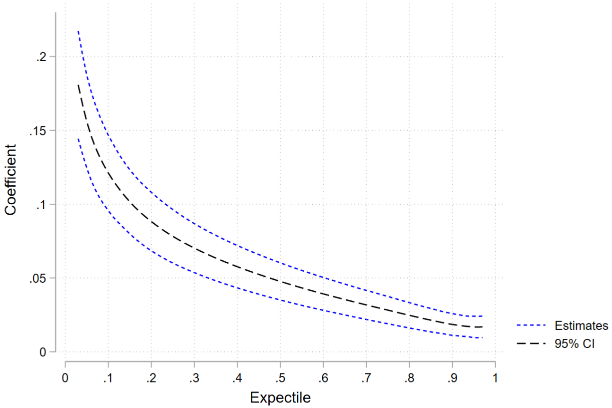
Results II

```
. appmlhdfe hours edu, a(age pkid skid hkid black ownh mort manager ///
> clerical farmer ur) e(.10)

Iteration 1: objective function = 8847.2249
Iteration 2: objective function = 29.708204
Iteration 3: objective function = .75390222
Iteration 4: objective function = .0000173
Iteration 5: objective function = 0

 Number of obs = 3373
 Iterations = 5
 Tolerance = 1.000e-07
 Objective function = 0
 % of negative residuals = .331
 R-squared: .17045263
.1 expectile regression
```

hours	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
edu	.1211607	.0129979	9.32	0.000	.0956854	.146636
_cons	.6872244	.1757405	3.91	0.000	.3427794	1.031669

Results III

Expectile	10th	25th	50th	75th	90th
Educ	0.121 (0.013)	0.078 (0.009)	0.048 (0.006)	0.028 (0.005)	0.019 (0.004)



- Educ **increases** the mean and **reduces** the dispersion of labour supply

Summary

- In most situations, expectiles are not particularly interesting.
- There are, however, cases where expectiles can be very useful.
- Here we considered that case of non-negative data with a mass-point at zero.
- Quantile regressions are not very appealing in this context.
- Expectiles provide an alternative way to study how the regressors affect different regions of the conditional distribution.
- The estimator is very easy to implement and the parameters have a straightforward interpretation.

References

- Bergstrand, J., M. Clance and J.M.C. Santos Silva (2025). “The Tails of Gravity: Using expectiles to quantify the trade-margins effects of economic integration agreements,” *Journal of International Economics*, forthcoming.
- Clance, M. and J.M.C. Santos Silva (2025). “APPMLHDFE: Stata module to estimate asymmetric Poisson regression with high dimensional fixed effects,” Statistical Software Components S459414, Boston College Department of Economics.
- Correia, S., P. Guimarães and T. Zylkin (2020). “Fast Poisson Estimation with High-Dimensional Fixed Effects,” *The Stata Journal*, 20, 95-115.
- Efron, B. (1992). “Poisson Overdispersion Estimates Based on the Method of Asymmetric Maximum Likelihood,” *Journal of the American Statistical Association*, 87, 98-107.

- Koenker, R. (2013). “Discussion: Living beyond our means,” *Statistical Modelling*, 13, 323-333.
- Koenker, R. and G. Bassett (1978). “Regression Quantiles,” *Econometrica*, 46, 33-50.
- Lee, M.-j. (1995). “Semiparametric Estimation of Simultaneous Equations with Limited Dependent Variables: A case study of female labor supply,” *Journal of Applied Econometrics*, 10, 187-200.
- Newey, W. and J. Powell (1987). “Asymmetric Least Squares Estimation and Testing,” *Econometrica*, 55, 819-847.