A contamination model for multivariate zero-inflated data

Davide Di Cecco, Ugo Guarnera, Danila Filipponi

Zero-inflated continuous data

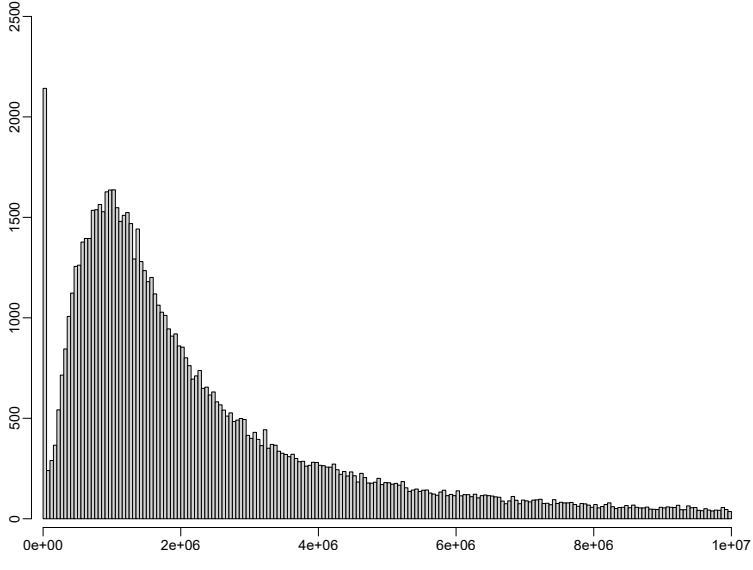
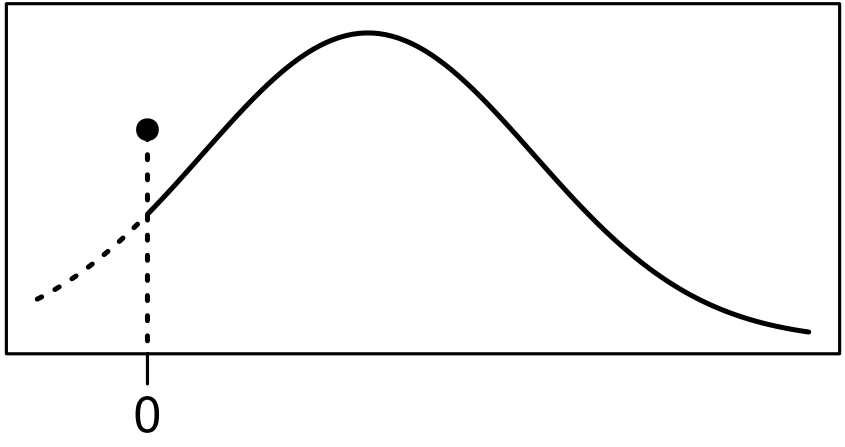
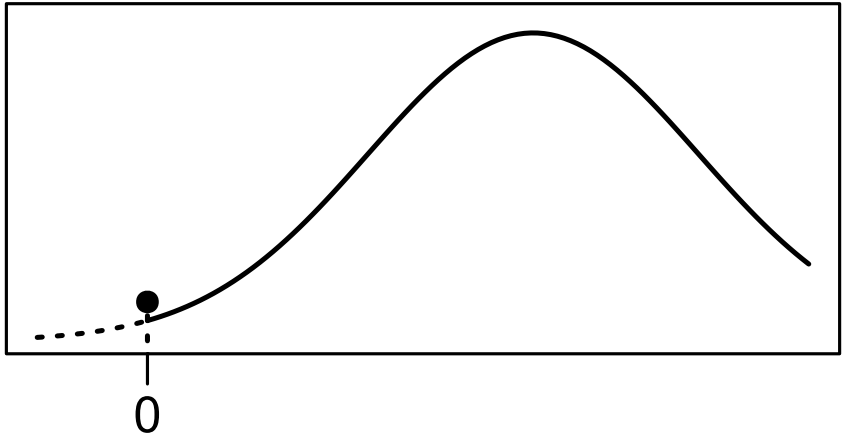


Figure: Distribution of the added value of Italian enterprises with more than 19 workers.

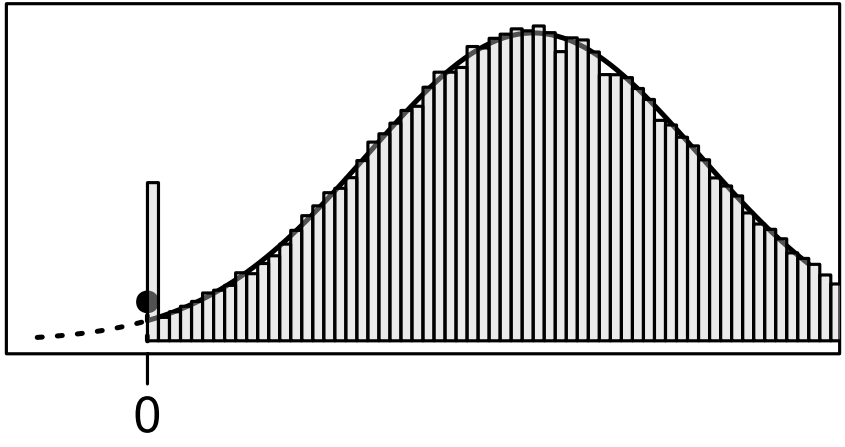
Tobit Model



Tobit



Tobit

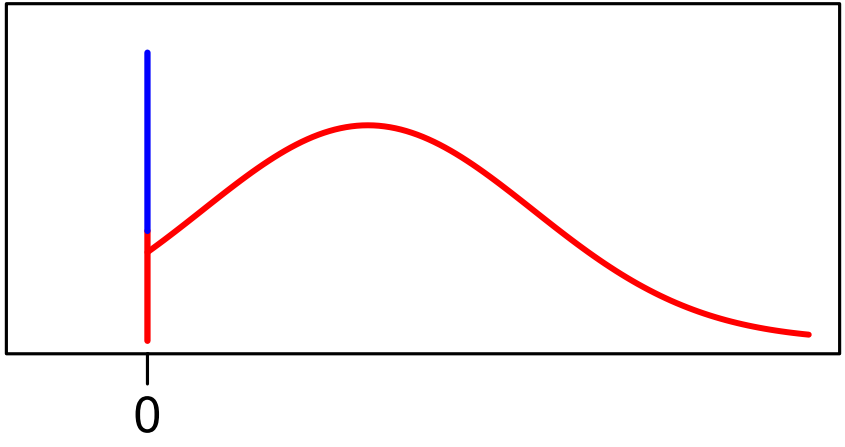


Zero-inflated Tobit

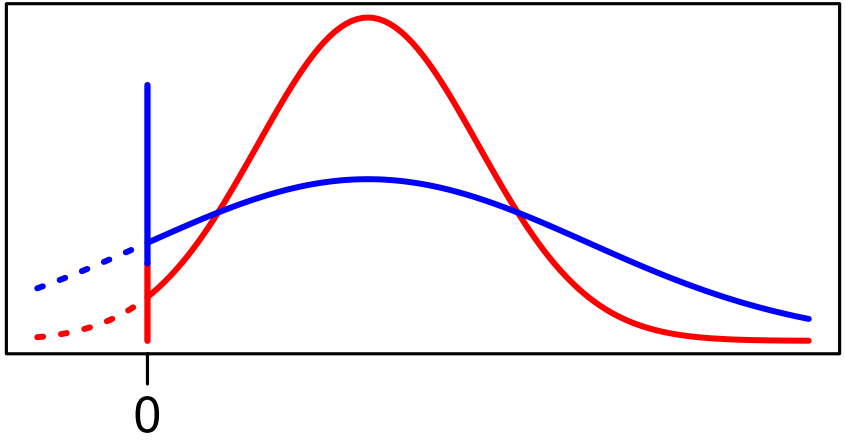
$$f(y \mid \mu, \sigma, \psi) = \begin{cases} (1-\psi)\Phi(0; \mu, \sigma) + \psi & \text{if } y = 0; \\ (1-\psi)\phi(y; \mu, \sigma) & \text{if } y > 1. \end{cases}$$

- Moulton, L. H. and N. A. Halsey. 1995. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*: 1570–1578.
- Tang, B., H. A. Frye, J. A. Silander Jr, and A. E. Gelfand. 2025. Zero-inflated multivariate tobit regression modeling. *Journal of Statistical Planning and Inference*, Vol. 236.

Zero-inflated Tobit



Mixture of Tobits



Mixture of Tobits - Contaminated Tobit

$$Y_i = (Y_{i,1}, ..., Y_{i,q}), \qquad Y_i = \max(\mathbf{0}, Z_i)$$

$$f(Z_i = z_i) = (1 - \pi)\, \phi(z_i; \mu_i, \Sigma) + \pi\, \phi(z_i; \mu_i, \alpha\Sigma), \qquad \mu_i = x_i^T \beta$$

Intermittent Contaminated Tobit model:

$$Z_i = Z_i^* + H_i\, \epsilon$$

$Z_i^* \sim N(\mu_i, \Sigma)$ True value $\qquad \epsilon \sim N(0, \Sigma_\epsilon)$ Contamination error

$$H_i \sim Ber(\pi), \qquad \Sigma_\epsilon + \Sigma = \alpha\Sigma.$$

Editing and Imputation

$$\begin{aligned}\text{Predicted true value} &= \text{What we would expect without error} \\ &= \text{Expectation of true value } (Y_i^*) \text{ given the observed value } Y_i\end{aligned}$$

$$\begin{aligned}\hat{y}_i &= E[Y_i^* | Y_i = y_i] \\ &= P(H_i = 0)\, y_i + P(H_i = 1)\, E[Y_i^* | y_i, H_i = 1]\end{aligned}$$

- **Selective editing:** select units contributing more to the **Total expected error** $= \hat{E} = \sum_i (y_i - \hat{y}_i)$ and impute at random from $f(y_i^* | y_i)$
- use $\hat{y}_i$ for all units in the sample to get robust estimates of the total values of Y in any specific domain D, $\sum_{i \in D} \hat{y}_i$, accounting for the possible presence of errors.

Editing and Imputation

$$E[Y^*|Y=0] = E[Y^*|Z<0] = \frac{\sigma A + \mu B}{\Phi(G)},$$

where

$$G = -\frac{\mu}{\sigma\sqrt{\alpha}},$$

$$A = -\frac{\phi(G)}{\sqrt{\alpha}}\left(1 - \Phi\left(G\sqrt{\alpha-1}\right)\right) + \phi\left(-\frac{\mu}{\sigma}\right)/2,$$

$$B = \frac{1}{2}\left(\Phi(G) - \Phi\left(-\frac{\mu}{\sigma}\right)\right) + T\left(G, \sqrt{\alpha-1}\right),$$

and T is Owen's T function

Application

We consider the added value of the 224,482 Italian companies with more than 5 employees in 2021. 5,346 (2.38%) have a zero value (for various reasons).

Covariates: value of exports and imports, number of employees and self-employed workers, turnover, eventual membership to business groups, average years of education of employees

	Tobit	ZIT	ICT
AIC	3443220	3255560	3176849
BIC	3443323	3255664	3176973

employees	y	$\hat{y}$
06-09	4080242	4238161
10-19	6581319	6730261
20-49	17820935	18434103
50-99	47039861	46925275
100-250	110903943	116087624
250+	659777863	5939659315