Some approaches in analyzing the data with excess of zeros

Tetiana Ianevych

Taras Shevchenko National University of Kyiv, Ukraine

BaNoCoSS, Helsinki, 2015

Main goal

- To develop methodology for quarter surveys of capital expenditure in Ukraine based on probability sampling

What we had before

- Annual surveys – censuses
- Quarter surveys – censored non-probabilistic sampling

Data to analyze

- Annual and quarter capital expenditure of the Ukrainian enterprises
- Annual surveys – censuses, 2009, 2010
- Quarter surveys – “sample”, 2010, 2011

Population features

- the majority of enterprises are small ones

- The main contribution into the total capital expenditure is made by big and middle-size enterprises

Big and middle-size enterprises are always surveyed, small ones are sampled

Population features

- the sampling design of the small enterprises is suppose to be stratifying according to the type of economical activity

- BUT we also need to obtain the estimates for different regions, types of capital expenditure, etc.

Objective

* to incorporate auxiliary information in order to improve the estimates for different domains leaving the sampling design simple

How?

Utilizing GREG estimator

$$\hat{Y} = \left(\sum_{i \in U} \hat{y}_i + \sum_{i \in S} w_i \left(y_i - \hat{y}_i \right) \right)$$

$\hat{y}_i$ are predictive values from linear model

Linear Regression Model

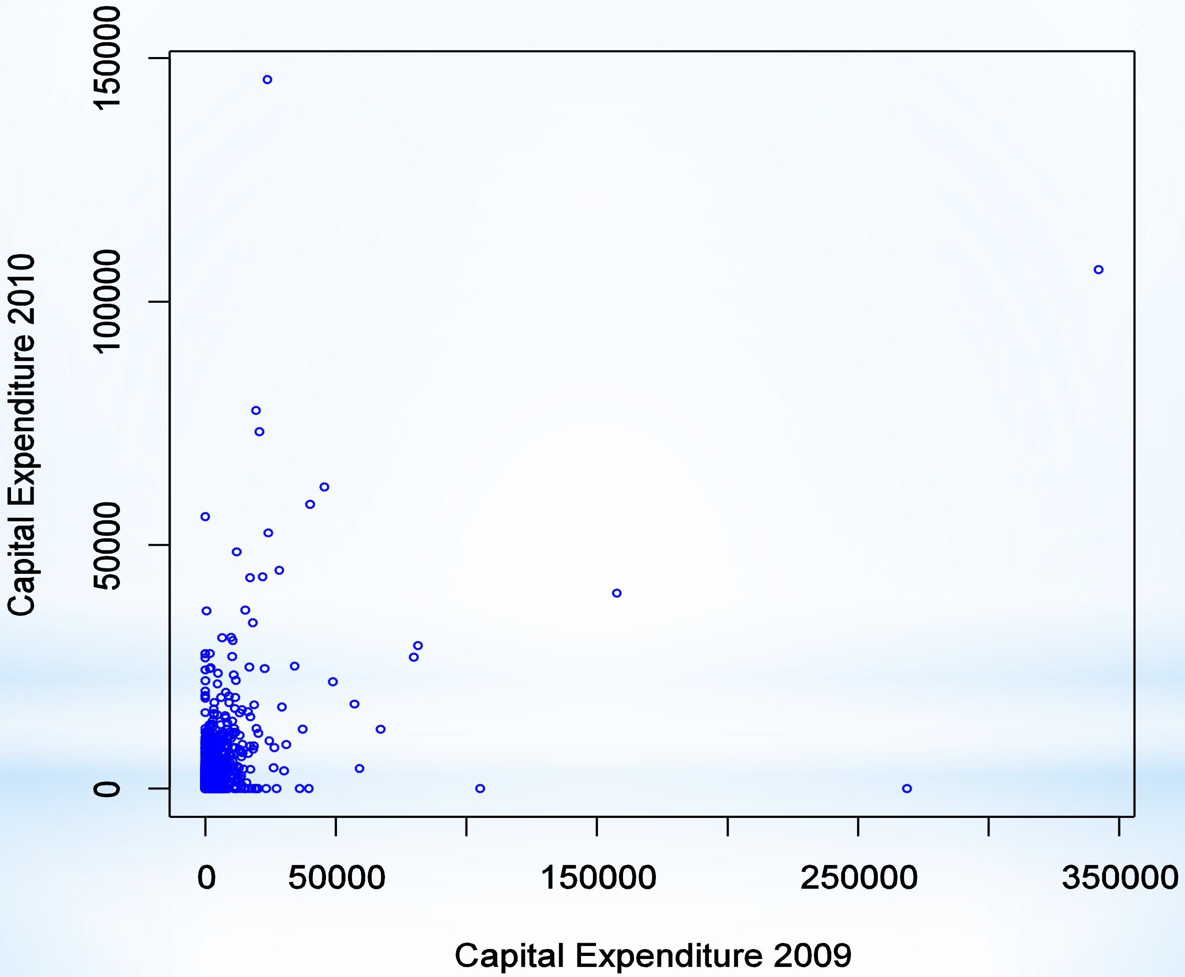
$$y_i = x_i^{'}\beta + u_i\,,\ \ i = 1,2,\ldots,n$$

u_i are i.d.d. drawings from the Normal distribution $N(0, \sigma_u)$

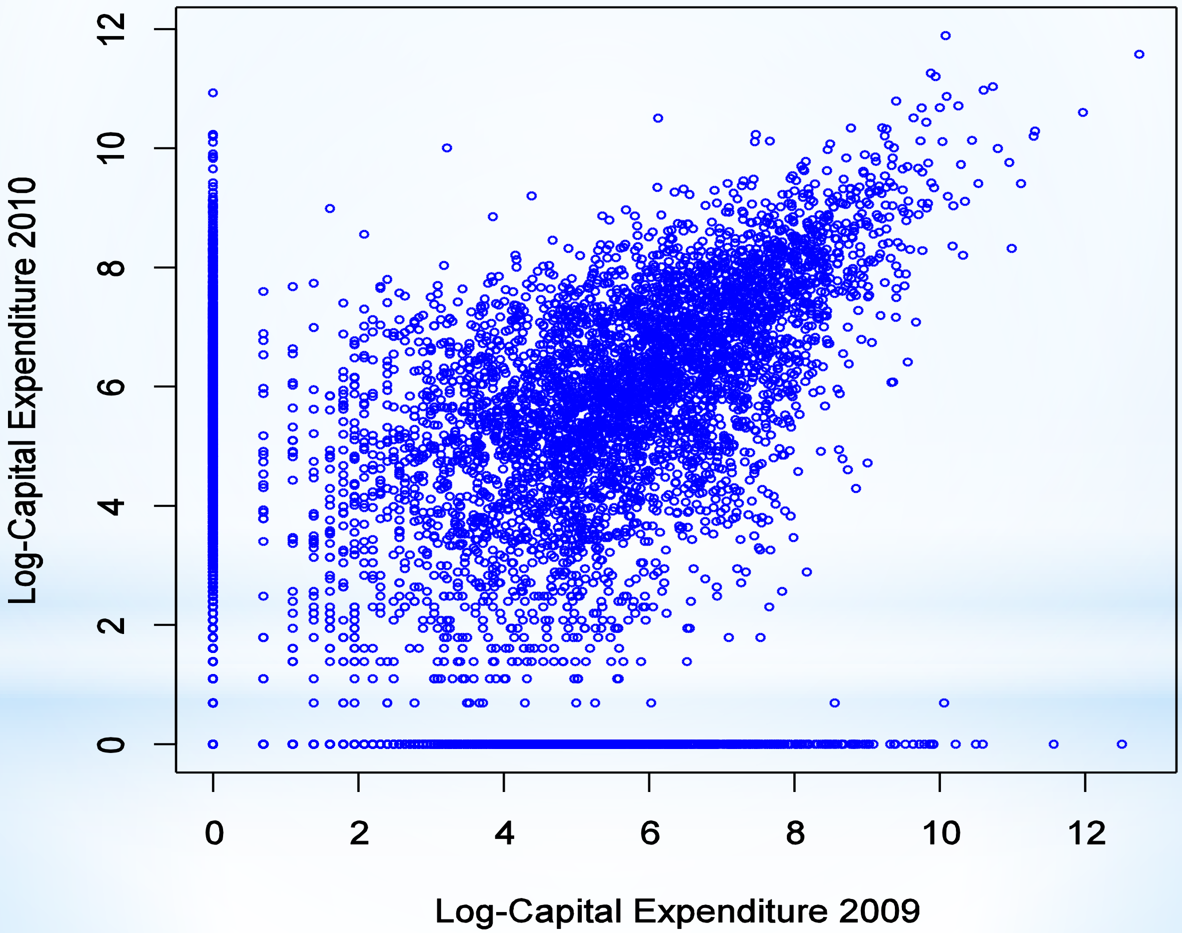
x_i could be:

- capital expenditure for the previous year;
- number of employee,
- revenue, etc

Capital Expenditure 2009 V 2010



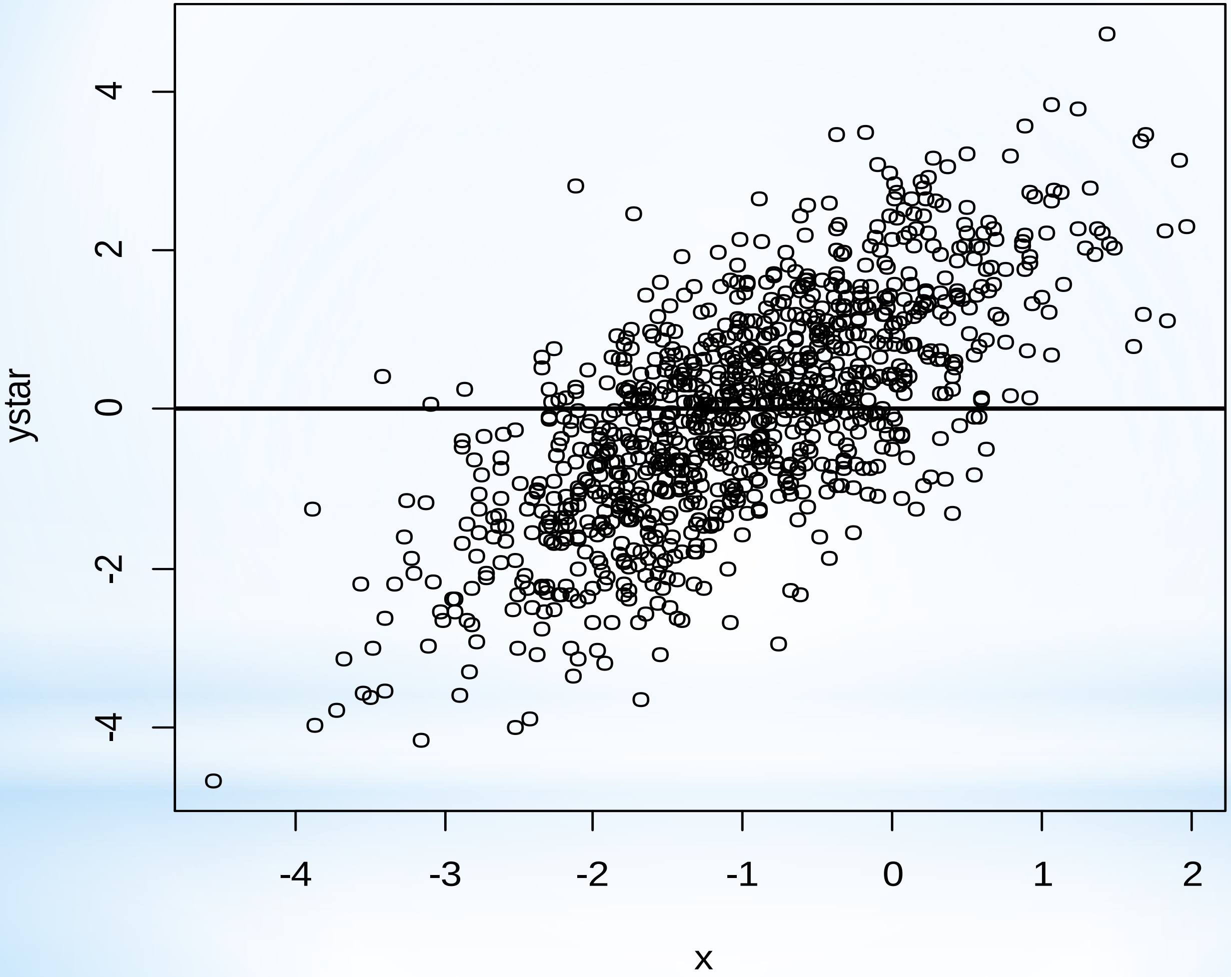
Log-Capital Expenditure 2009 V 2010



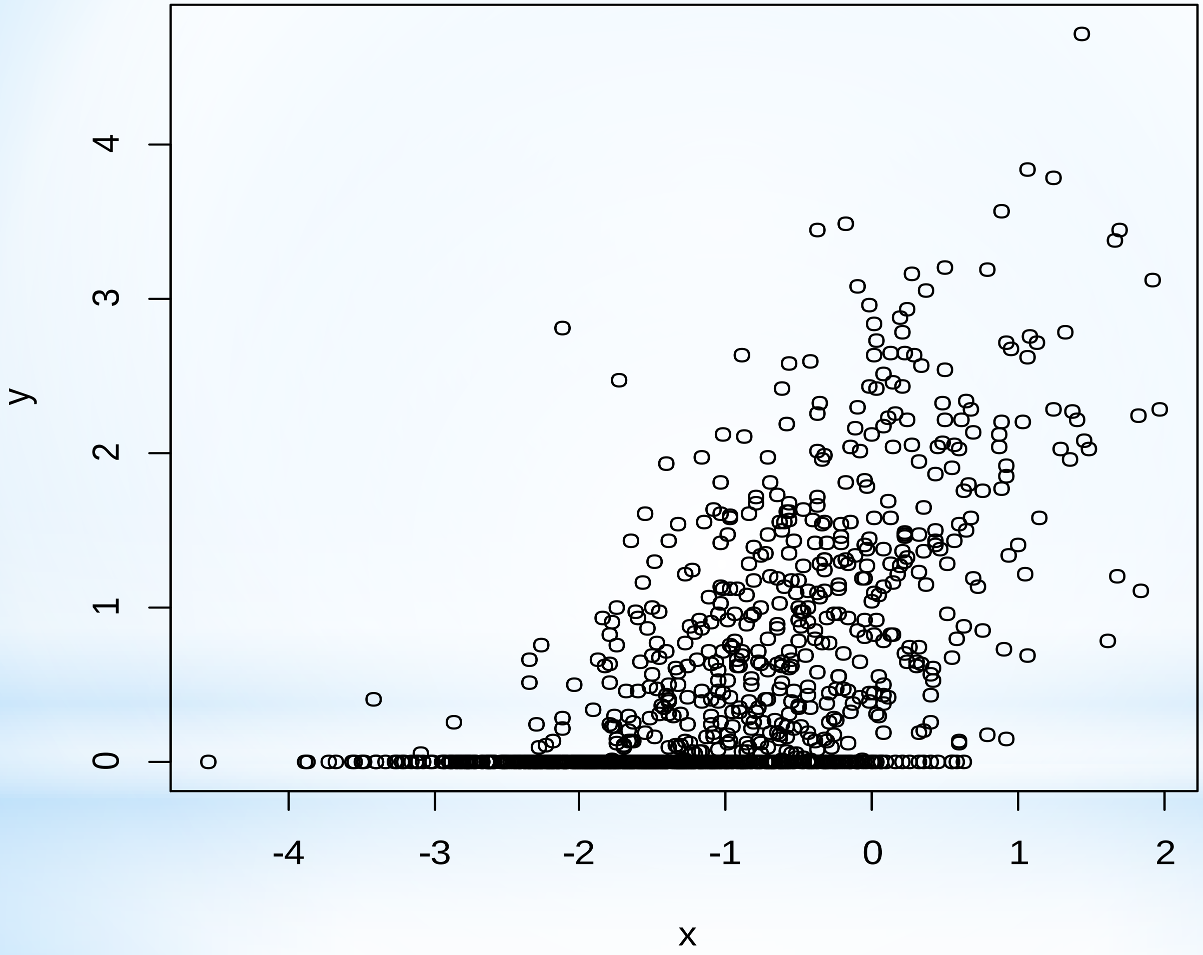
*Tobit Model

- introduced by Tobin (1958)
- also called as a censored regression model
- $$y_i^* = x_i^{'}\beta + u_i\,, \quad i = 1,2,\ldots,n$$
  $$y_i = \max(0,\, y_i^*)$$
- u_i are i.d.d. drawings from the Normal distribution $N(0,\sigma_u)$

Underlying generating process



Censored data



*Heckit Model

*Model is called in honor of James Heckman

$$y_i^* = \beta' x_i + u_i$$

$$y_i = d \cdot y_i^*$$

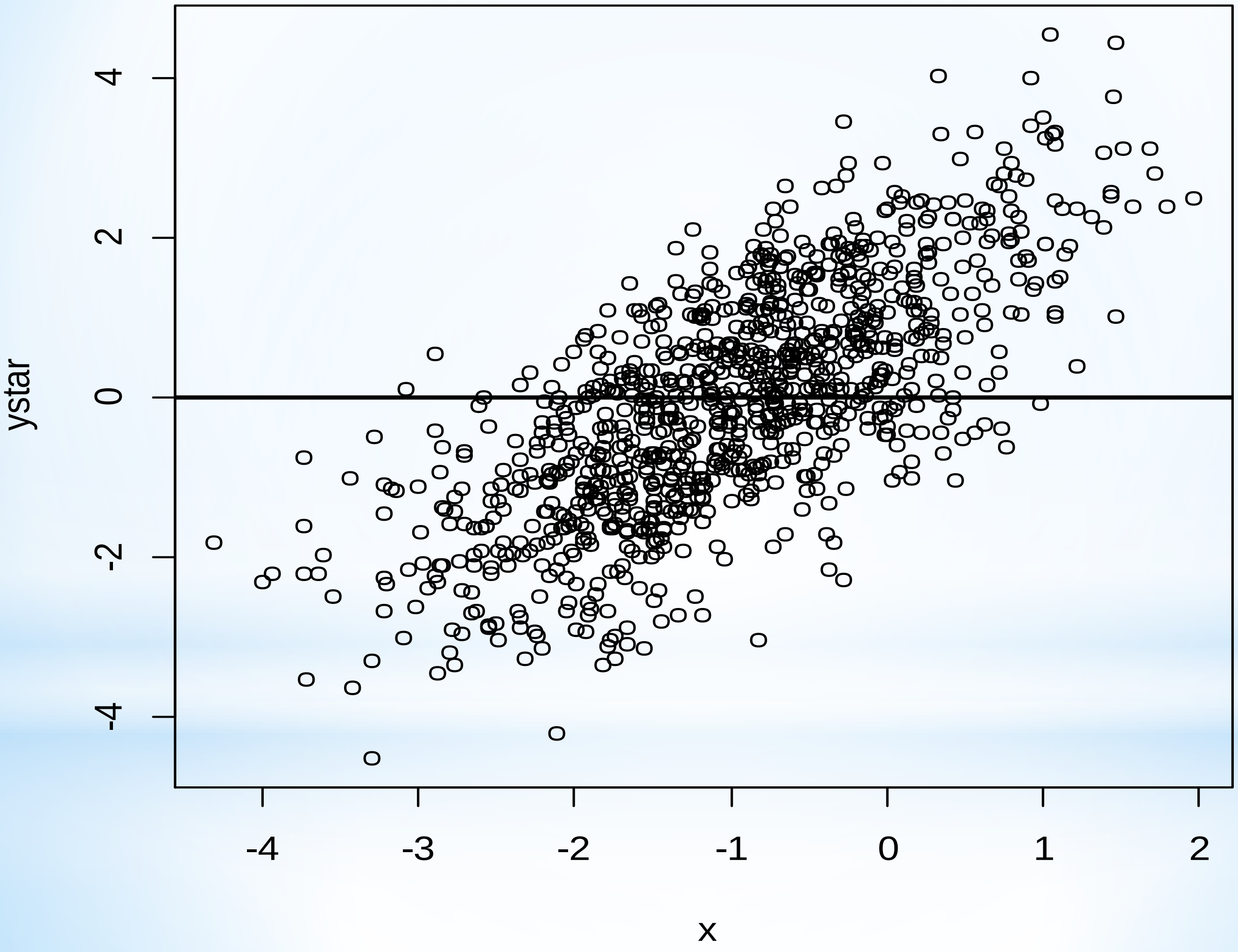
$$w = \alpha' z + v$$

$$d = 1 \ \ if \ \ \ w > 0$$

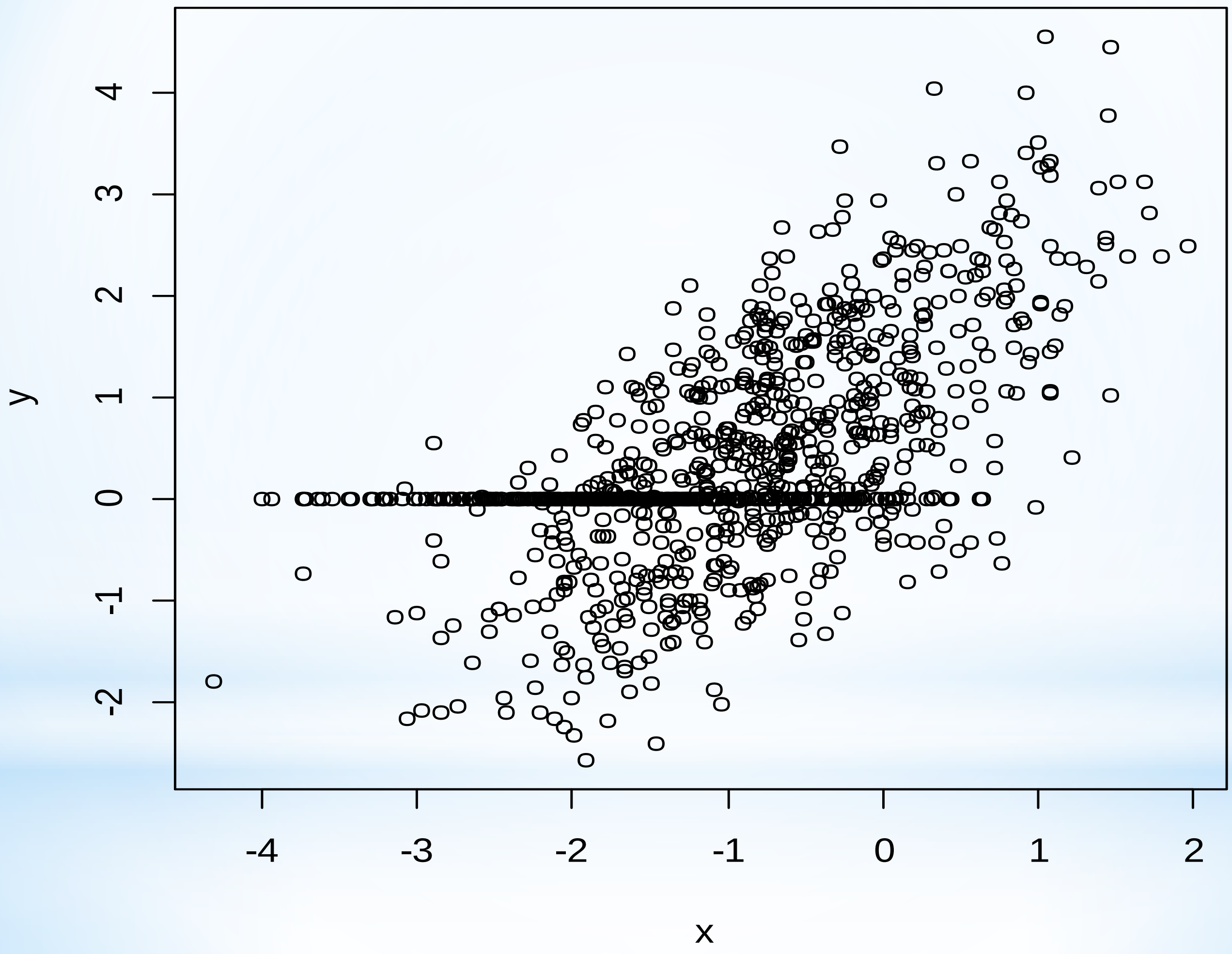
$$d = 0 \ otherwise$$

*(u,v) are jointly normally distributed

Underlying generating process



Zero-inflated data



Simulation study

- For simplicity we consider as a population only one strata – strata of agricultural enterprises

- N=40588
- Sample consists of n=4227 enterprises including 3485 big enterprises included into the sample with probability 1 and 742 small enterprises sampled by SRS with probability 0.002 (742/37103)

- Number of Monte-Carlo simulations – 10000

Bias and mean squared error

ARB absolute relative bias

$$ARB = \left| \frac{1}{K} \sum_{i=1}^{K} \hat{\bar{y}}_{GREG}(s_i) - \bar{Y} \right| / \bar{Y}$$

RRMSE relative root mean squared error

$$RRMSE \quad = \sqrt{\frac{1}{K} \sum_{i=1}^{K} \left(\hat{\bar{y}}_{GREG}(s_i) - \bar{Y} \right)^2} / \bar{Y}$$

*Comparison of HT with GREG based on LReg and Tobit models

%	HT	LREG	Log-LREG	Tobit	Log-Tobit
ARB	0,07	1,27	0,08	3,94	32,09
RRMSE	8,15	7,89	7,86	12,2	229,4

*Comparison of HT with GREG based on LReg and Heckit models

%	HT	LREG	Log-LREG	Heckit	Log-Heckit
ARB	0,07	1,27	0,08	0,13	0,05
RRMSE	8,15	7,89	7,86	7,89	7,74

Thank you for your attention!