Rounding Error and Ordinary Least Squares Regressions

Jos van Bommel

Luxembourg School of Finance, University of Luxembourg

Most scientific variables (length, time, weight, temperature, pressure) are **rounded** to the nearest tickmark.

Rounding error is not independent noise, noise that is independent from the latent variable.

Quite the opposite: **rounding error is deterministic**. It's dependent on the latent variable and the measurement grid.

It is well known that *error-in-variable* leads to downward biased slope estimates in regression analysis.

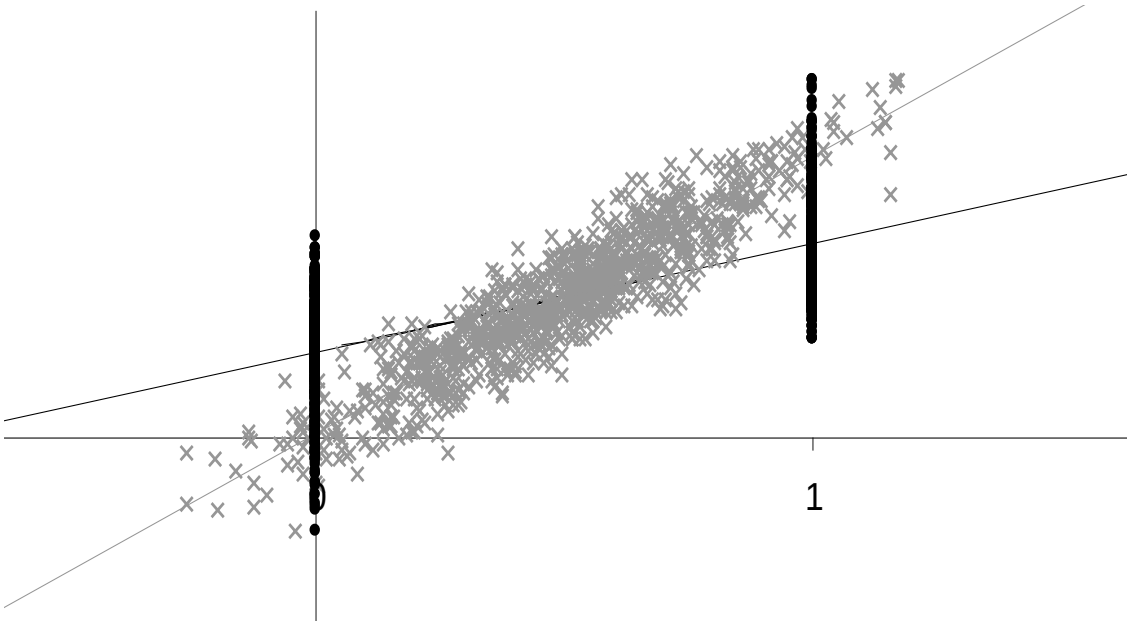
For rounding error this is not so simple. The incumbent literature only offers approximate corrections.

We compute the precise bias due to rounding for the Normal case,

and suggest a Maximum likelihood method to arrive at unbiased regression estimates.

Problem I

(X-variable is rounded)



The **grey markers** denote the true values. The **black markers** denote what we observe due to (severe) rounding. The grey line gives the true regression slope, the black line the measured OLS regression slope. Clearly our measured regression slope is ***biased***.

The incumbent literature (and statistical software), uses the *Sheppard's Correction:*

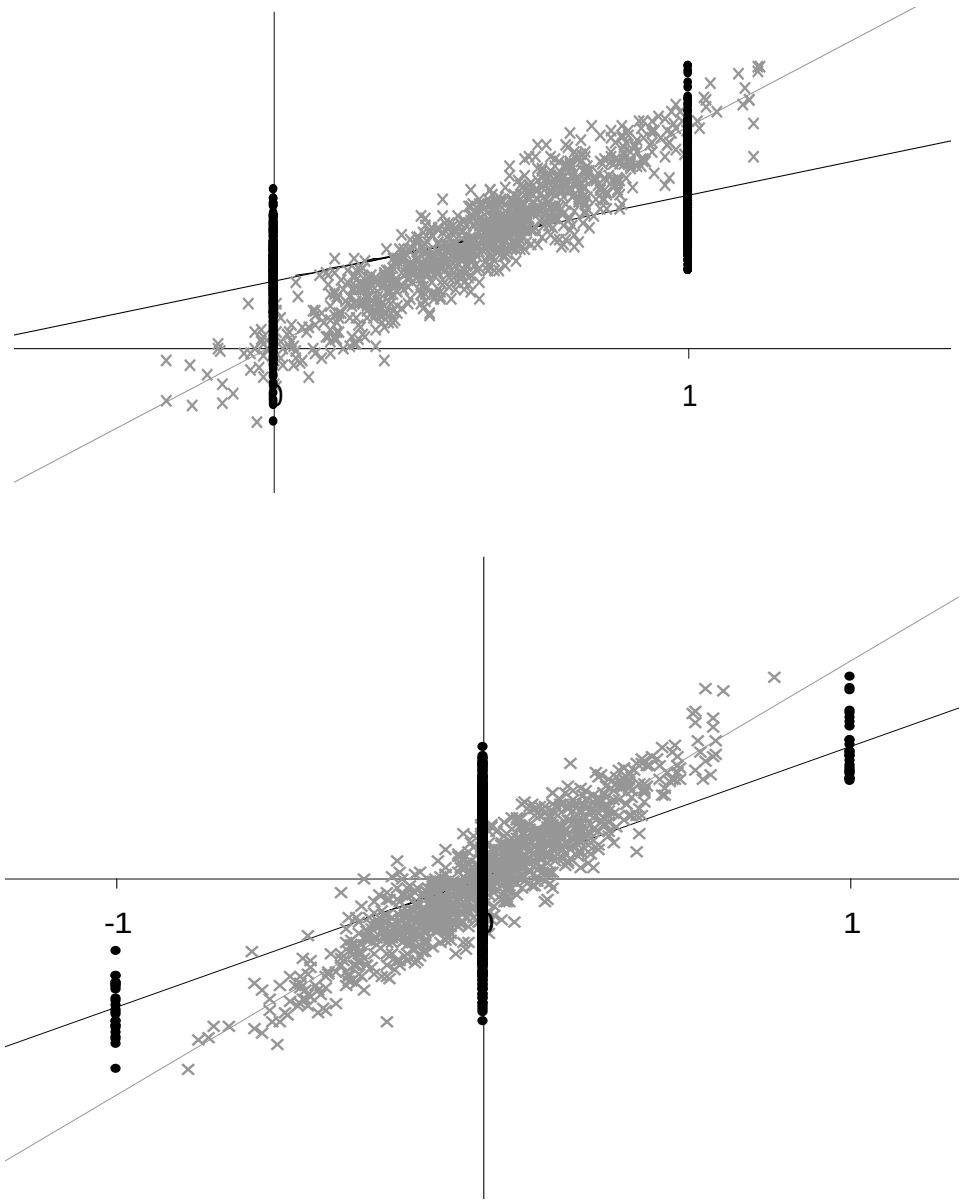
$$\hat{\hat{\beta}} = \hat{\beta}_{OLS} \frac{\hat{\sigma}_x^2 + \frac{1}{12}\lambda_x^2}{\hat{\sigma}_x^2}$$

Where $\hat{\sigma}_x^2$ is the estimated variance of the independent variable.

This is a decent correction, but not good enough!

Analysis

(X-variable is rounded)

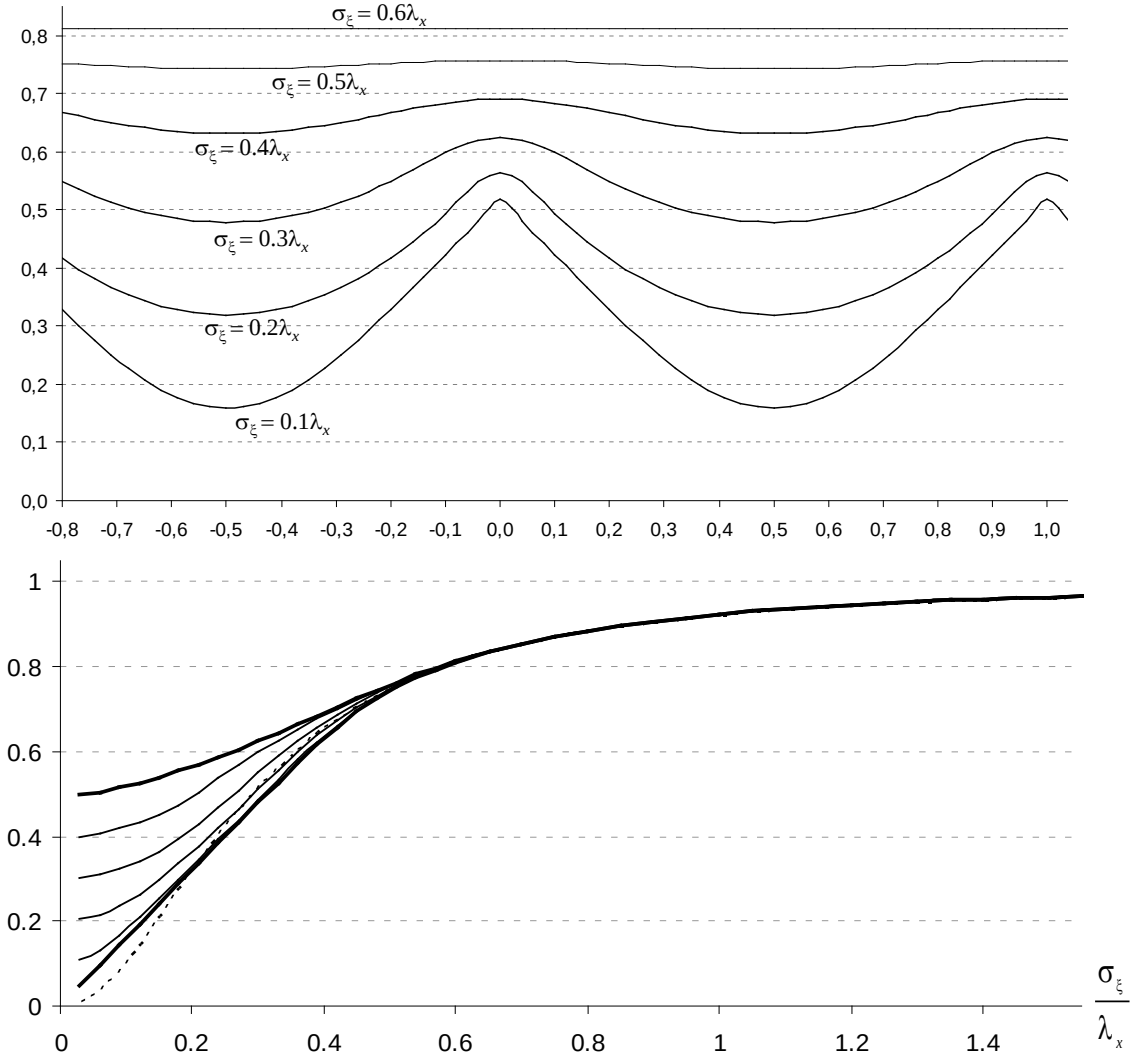


The bias does not only depend on σ_{ξ}^{2} (sigma of the latent variable) and λ, but also on μ_{ξ} (the location)

Finding
(X-variable is rounded)

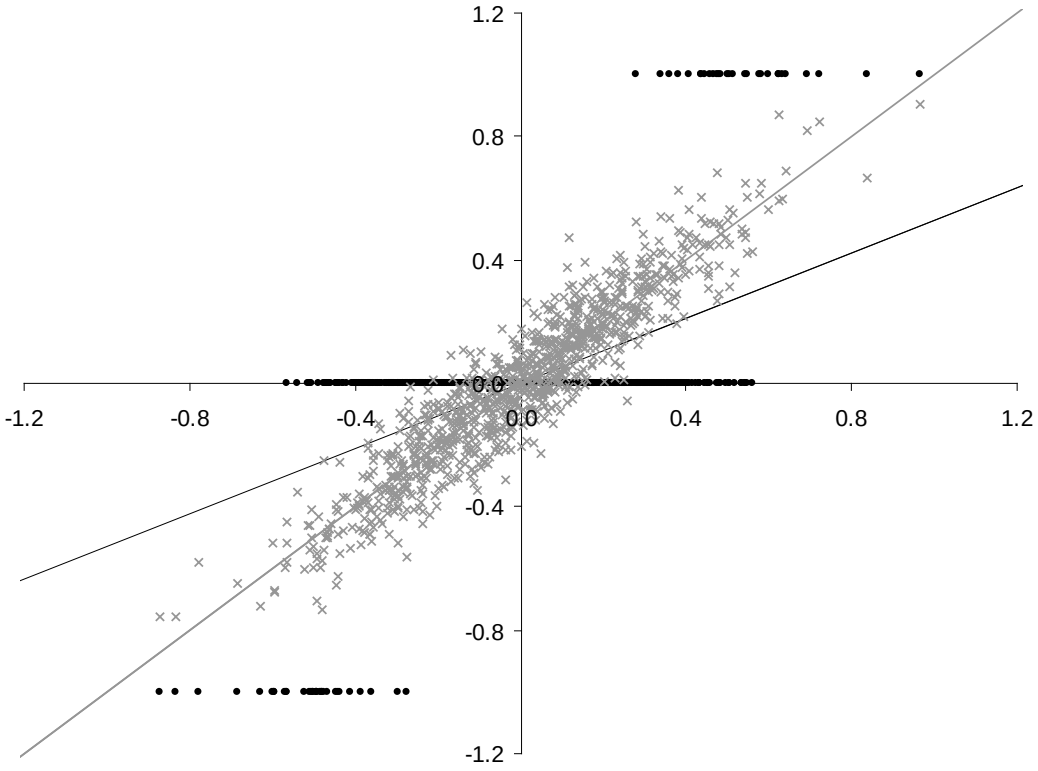
Figure 1: beta bias as a function of the distribution of the independent variable

This figure gives *plim* $\frac{\hat{\beta}}{\beta}$ as a function of $\frac{\sigma_\xi}{\lambda_x}$ and $\frac{\mu_\xi}{\lambda_x}$ where μ_ξ and σ_ξ are the mean and standard deviation of the unrounded variable and λ_x is the width of the measurement grid. In the bottom graph, the top bold line gives *plim* $\frac{\hat{\beta}}{\beta}$ if the independent variable is centred around a multiple of λ_x. The bottom line gives the attenuation factor if $\tilde{\xi}$ is centred around $(½+i)\lambda_x$ for any integer i. The thin lines have μ_ξ at a distance of $0.1\lambda_x$, $0.2\lambda_x$, $0.3\lambda_x$, and $0.4\lambda_x$ from a tickmark. The dashed line is the approximation $\frac{\sigma_\xi^2}{\sigma_\xi^2 + \frac{1}{12}\lambda_x^2}$.



Problem II

(Y-variable is rounded)

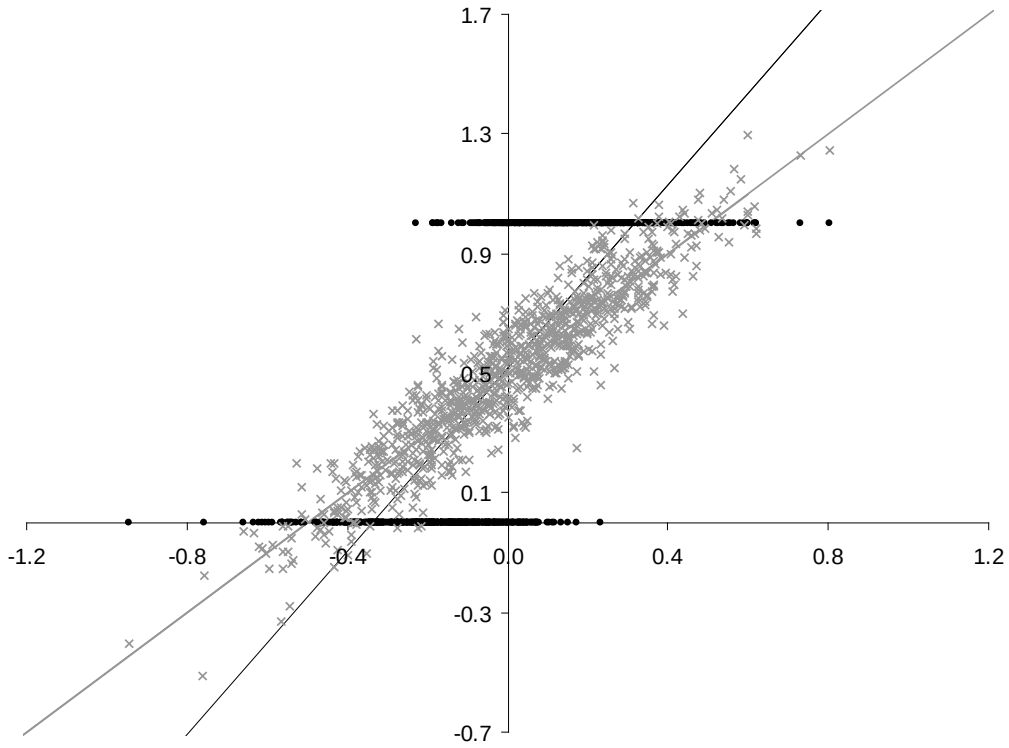


***True values in grey*, observed values in black**

Due to rounding of the dependent variable, the OLS regression slope is ***downward biased***..

Problem II

(Y-variable is rounded)

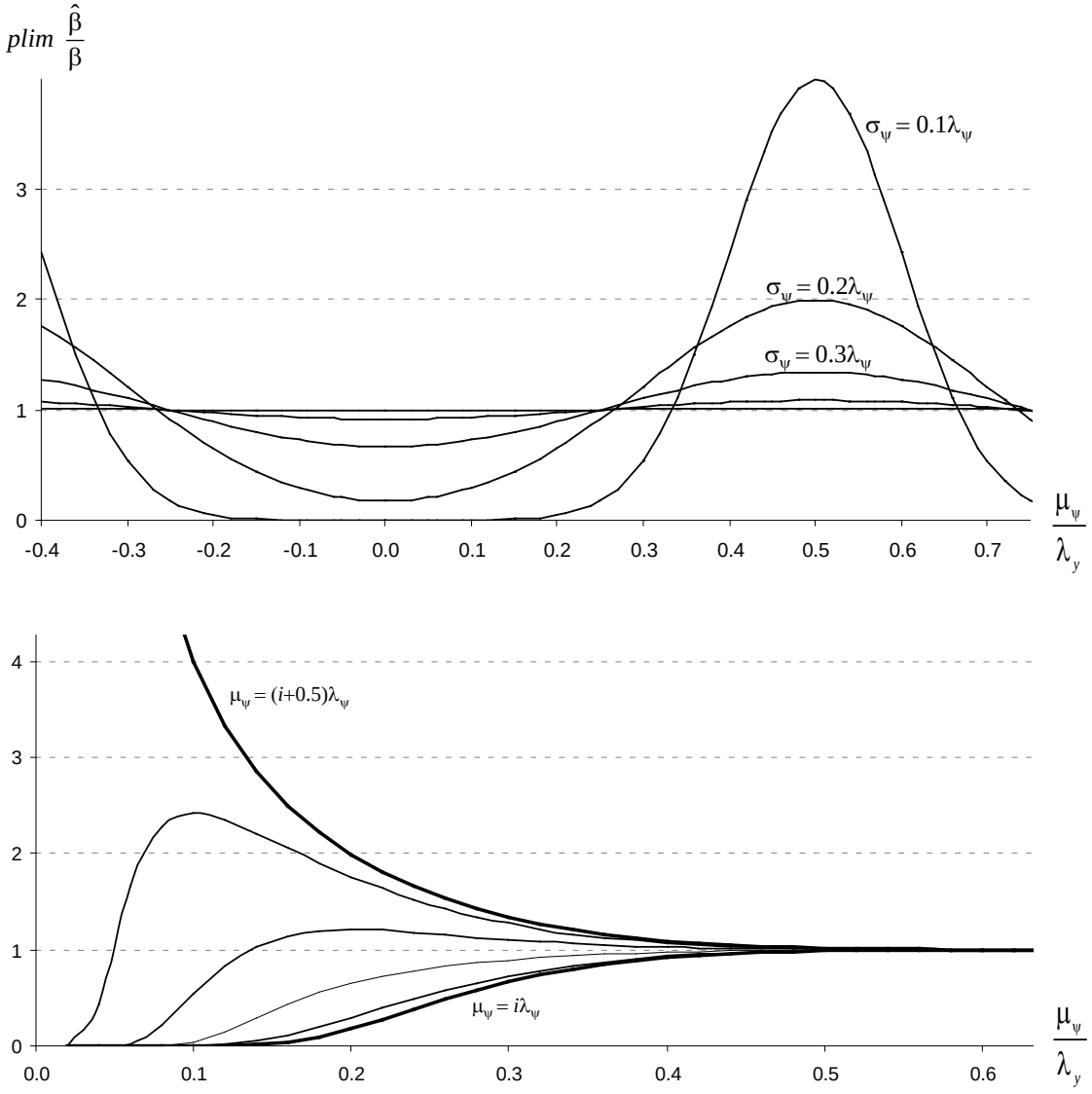


With the mean of the latent Y-variable lies higher, the OLS slope is ***upward biased***!!

Finding
(Y-variable is rounded)

Figure 3: beta bias as a function of the distribution of the dependent variable

This figure gives *plim* $\frac{\hat{\beta}}{\beta}$ as a function of $\frac{\sigma_\psi}{\lambda_y}$ and $\frac{\mu_\psi}{\lambda_y}$ where μ_ψ and σ_ψ are the mean and standard deviation of the unrounded variable and λ_y is the width of the measurement grid. In the bottom graph, the bold lines give *plim* $\frac{\hat{\beta}}{\beta}$ if the independent variable is centred around a multiple of λ_y or around $(½+i)\lambda_y$, for any integer *i*. The thin lines have μ_ψ at a distance of $0.1\lambda_y$, $0.2\lambda_y$, $0.3\lambda_y$, and $0.4\lambda_y$ from a tickmark.



Propositions

Proposition 1 *(dependent variable rounded): If, in a univariate regression, the latent independent variable follows a normal distribution, and is precisely rounded, a consistent estimator for the slope coefficient is given by:*

$$\hat{\beta}_{xr} = \hat{\beta}_{OLS} \frac{\hat{\sigma}_{\xi}^2 + \hat{\sigma}^2(\tilde{\nu}_{\xi}) + 2\hat{\sigma}(\tilde{\xi}, \tilde{\nu}_{\xi})}{\hat{\sigma}_{\xi}^2 + \hat{\sigma}(\tilde{\xi}, \tilde{\nu}_{\xi})}$$

Where $\hat{\mu}_{\xi}$ *and* $\hat{\sigma}_{\xi}$ *are consistent estimates of the mean and standard deviation of* $\tilde{\xi}$, *and* $\hat{\sigma}^2(\tilde{\nu}_{\xi})$ *and* $\hat{\sigma}(\tilde{\xi}, \tilde{\nu}_{\xi})$, *are computed from*:

$$\text{var}(\tilde{\nu}_{\xi}) = \frac{1}{\sigma_{\xi}\sqrt{2\pi}} \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})\lambda_x}^{(i+\frac{1}{2})\lambda_x} \left(i\lambda_x - \xi - E\left[\tilde{\nu}_{\xi}\right]\right)^2 e^{\frac{-(\xi-\mu_{\xi})^2}{2\sigma_{\xi}^2}} d\xi$$

and

$$\text{cov}(\tilde{\xi}, \tilde{\nu}_{\xi}) = \frac{1}{\sigma_{\xi}\sqrt{2\pi}} \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})\lambda_x}^{(i+\frac{1}{2})\lambda_x} (\xi - \mu_{\xi})\left(i\lambda_x - \xi - E\left[\tilde{\nu}_{\xi}\right]\right) e^{\frac{-(\xi-\mu_{\xi})^2}{2\sigma_{\xi}^2}} d\xi,$$

where

$$E\left[\tilde{\nu}_{\xi}\right] = \frac{1}{\sigma_{\xi}\sqrt{2\pi}} \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})\lambda_x}^{(i+\frac{1}{2})\lambda_x} \left(i\lambda_x - \xi\right) e^{\frac{-(\xi-\mu_{\xi})^2}{2\sigma_{\xi}^2}} d\xi$$

Proposition 2 *(independent variable rounded): If, in a univariate regression, the latent dependent variable follows a normal distribution, and is precisely rounded, a consistent estimator for the slope coefficient is given by:*

$$\hat{\beta}_{yr} = \hat{\beta}_{OLS} \frac{\hat{\sigma}_{\psi}^2}{\hat{\sigma}_{\psi}^2 + \hat{\sigma}(\tilde{\xi}, \tilde{\nu}_{\psi})},$$

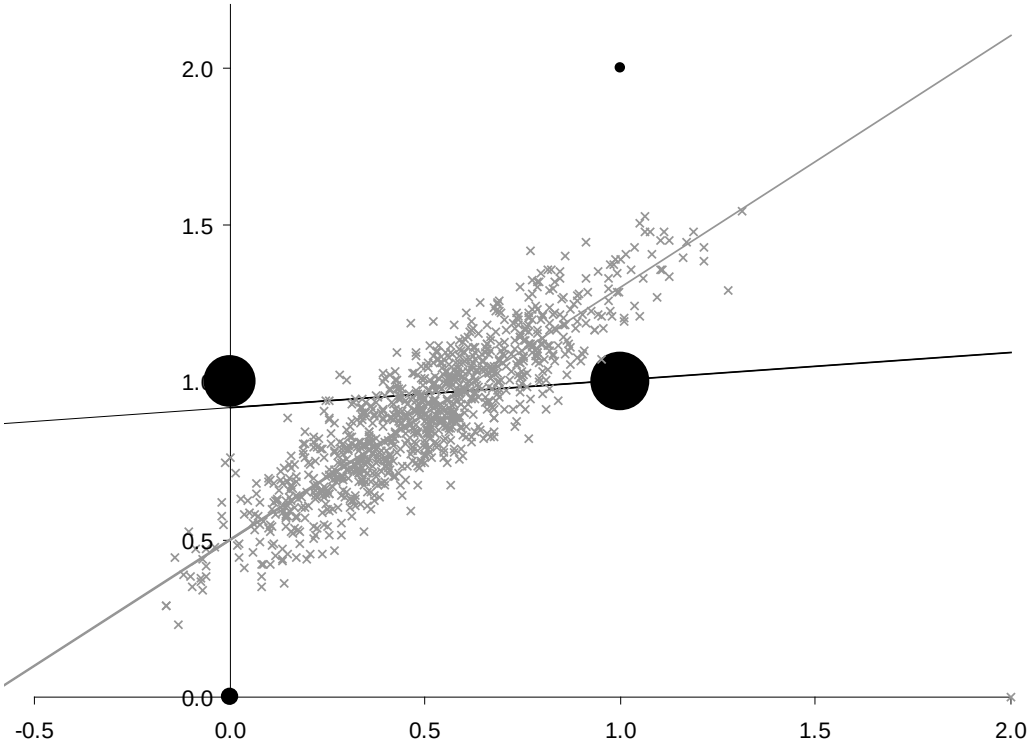
where $\hat{\mu}_{\psi}$ *and* $\hat{\sigma}_{\psi}$ *are consistent estimates of the mean and standard deviation of* $\tilde{\psi}$, *and* $\hat{\sigma}(\tilde{\psi}, \tilde{\nu}_{\psi})$ *is computed from*:

$$\text{cov}(\tilde{\psi}, \tilde{\nu}_{\psi}) = \frac{1}{\sigma_{\psi}\sqrt{2\pi}} \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})\lambda_y}^{(i+\frac{1}{2})\lambda_y} (\psi - \mu_{\psi})\left(i\lambda - \psi - E\left[\tilde{\nu}_{\psi}\right]\right) e^{\frac{-(\psi-\mu_{\psi})^2}{2\sigma_{\psi}^2}} d\psi$$

with $\hat{\mu}_{\psi}$ *and* $\hat{\sigma}_{\psi}$ *in lieu of* μ_{ψ} *and* σ_{ψ}.

Problem III

(X and Y-variable rounded)

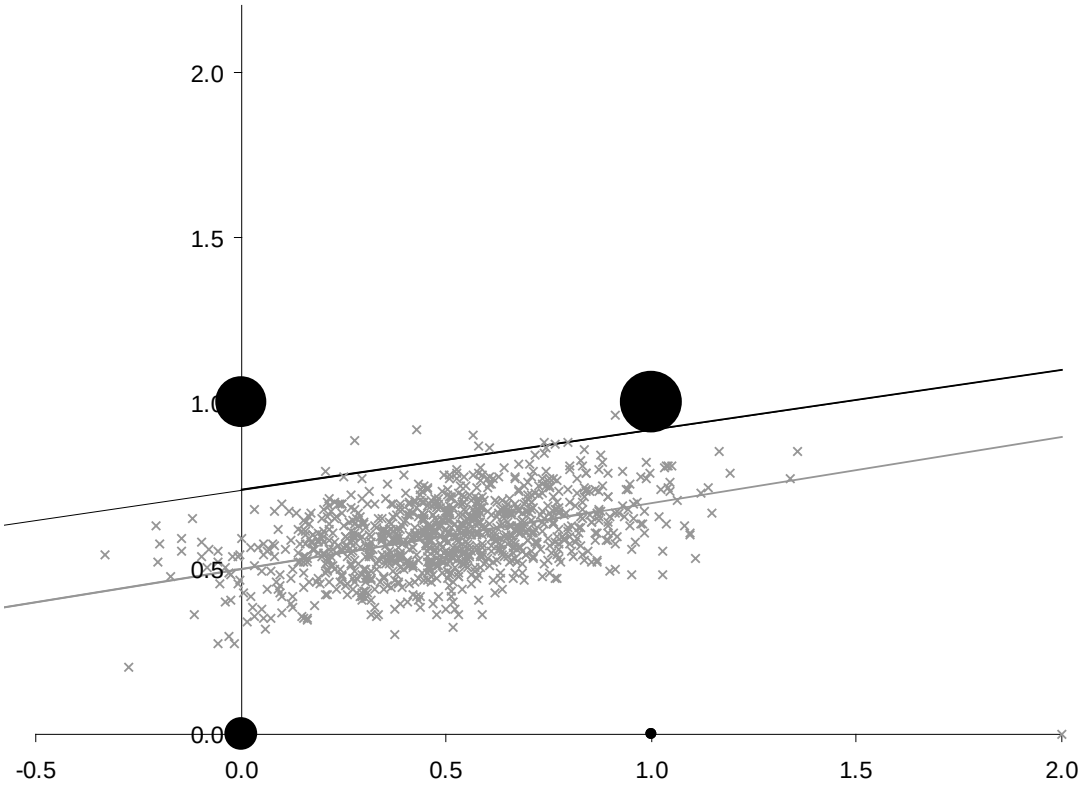


***True values in grey*, observed values in black**

The OLS regression slope is severely ***downward biased***..

Problem III

(X and Y-variable rounded)



We now ***decrease*** the true slope (intercept held constant), and the OLS measured slope ***increases***

→ Neglecting rounding error can cause very wrong inferences

Solution to rounding problems

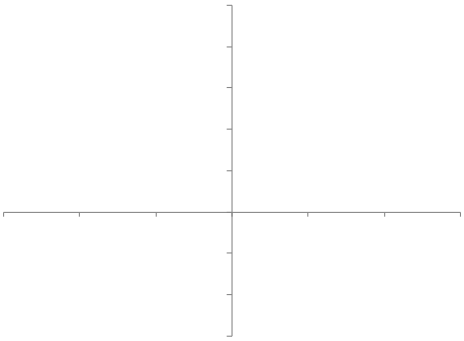
We specify a consistent estimator $\{\alpha, \beta\}$ in a OLS regression (see paper).

The key ingredients are the mean and standard deviations of the latent variables ξ and ψ.

We use a maximum likelihood method to estimate $\{\mu_\xi, \sigma_\xi\}$ from the observed X-values and $\{\mu_\psi, \sigma_\psi\}$ from the Y-observations, ***assuming that both are Normally distributed.***

Testing our Solution

We generate samples of 10,000 observations, $\tilde{\psi} = \alpha + \beta\tilde{\xi} + \tilde{\varepsilon}$, choosing α, β, μ_ξ, and σ_ξ, and inflict severe rounding:



		x			
		-1	0	1	total
y	2	-	-	2	2
	1	1	2,361	1,259	3,621
	0	30	5,044	1,283	6,357
	-1	-	20	-	20
	total	31	7,425	2,544	

Then we pretend we do not know α and β and estimate them, using OLS and the suggested Maximum Likelihood method and bias formulas.

For this particular example (true $\alpha = \beta = 0.3$), we find:

$\{\hat{\alpha}_{OLS}, \hat{\beta}_{OLS}\} = \{0.3145, 0.1832\}$

My method gives (after painful ML-maximization)

$\{\hat{\alpha}_{JVB}, \hat{\beta}_{JVB}\} = \{0.3023, 0.2897\}$

I am looking for a co-author with time and programming skills to finish the paper.

Interested? E-mail me: jos.vanbommel@uni.lu