Where is the Bias in OLS Estimates?

Model: $y_i = \alpha + \beta x_i + \epsilon_i$

Recall that the estimate of β is defined as

$$(1) \qquad \hat{\beta} = \sum_i^n (x_i - \bar{x})(y_i - \bar{y}) \,/\, \sum_i^n (x_i - \bar{x})^2$$

We can reduce the algebra slightly

$$\textit{substitute}: \quad a = \sum_i^n (x_i - \bar{x})^2$$

$$\sum_i^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_i^n (x_i - \bar{x})\, y_i$$

Substitute these values into (1)

$$(2) \qquad \hat{\beta} = (1/a) \sum_i^n (x_i - \bar{x})\, y_i$$

now substitute $\alpha + \beta x_i + \epsilon_i$ for y_i and expand

$$(3) \quad \hat{\beta} = (1/a) \sum_i^n (x_i - \bar{x})\, (\alpha + \beta\, x_i + \epsilon_i) =$$

$$(1/a) \sum_i^n (x_i - \bar{x})\, \alpha \; + \; (1/a) \sum_i^n \beta(x_i - \bar{x})\, x_i \; + \; (1/a) \sum_i^n (x_i - \bar{x}) + \epsilon_i$$

Some simple algebra tricks

$$Recall: \quad \sum_i^n (x_i - \bar{x}) = 0 \qquad so \qquad (1/a) \sum_i^n (x_i - \bar{x})\alpha = 0$$

also

$$\sum_i^n (x_i - \bar{x})x_i = \sum_i^n (x_i - \bar{x})(x_i - \bar{x}) = \sum_i^n (x_i - \bar{x})^2$$

so

$$(1/a) \sum_i^n \beta(x_i - \bar{x})^2 = \beta \, [\sum_i^n (x_i - \bar{x})^2]/[\sum_i^n (x_i - \bar{x})^2] = \beta$$

(4) $$\hat{\beta} = \beta + (1/a) \sum_i^n (x_i - \bar{x})\epsilon_i$$

so

$$\hat{\beta} = \beta + (1/a) \sum_i^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})$$

Recall the definition of covariance

$$\hat{\sigma}_{x,y} = C\hat{o}v(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Divide top and bottom of right-hand-side terms in previous equation by (n-1)

$$\hat{\beta} = \beta + \frac{[1/(n-1)]\sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{[1/(n-1)]\sum_{i=1}^n (x_i - \bar{x})^2}$$

Notice that:

$$\hat{\sigma}_{xe} = [1/(n-1)]\sum_{i=1}^{n}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})$$

$$\hat{\sigma}_x^2 = [1/(n-1)]\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Substitute these values into previous equation for β:

$$\hat{\beta} = \beta + \frac{\hat{\sigma}_{xe}}{\hat{\sigma}_x^2}$$

Notice that if the correlation between x and ϵ_i is zero, the estimate of β equals the true value. In contrast, if x and ϵ_i are correlated, then the estimate will systematically differ from the true value (hence it is biased)