Measurement Error

H.M. James Hung

Consider a typical simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- x is a random variable with mean μ_x and variance σ_x^2 , the values of both parameters are unknown.
- ε is a random error with mean 0 and variance σ_ε^2
- x and ε are statistically independent.
- β_0 and β_1 are parameters whose values are unknown.
- In fact, $\beta_1 = \sigma_{xy}/\sigma_x^2$, where σ_{xy} is the covariance between y an x.

Often in practice, the measurement on x is subjective to a measurement error. That is, instead of obtaining the value of x, we can only obtain the measurement X, where $X = x + u$, where u is a measurement error.

Assume

- u is a random variable with mean 0 and variance σ_u^2 , the value of which is unknown.
- u and x are statistically independent.
- u and ε are statistically independent.

Let the data from n independent items be $(y_1, X_1), \ldots, (y_n, X_n)$, noting that x_i $(i = 1, \ldots, n)$ are unavailable. For the slope parameter β_1 , regressing y on 1 and X, instead of x, will yield the OLS estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \ ,$$

This is a "consistent" estimator of $\frac{\sigma_{Xy}}{\sigma_X^2}$.

$$\sigma_{Xy} = Cov(x + u, y) = Cov(x, y) + Cov(u, y)$$
$$Cov(u, y) = Cov(u, \beta_0 + \beta_1 x + \varepsilon) = 0$$

Thus, $\sigma_{Xy} = \sigma_{xy}$.

$$\sigma_X^2 = Cov(X, X) = Var(x) + Var(u)$$

The OLS estimator in regressing y on 1 and X is a "consistent" estimator of $\frac{\sigma_{Xy}}{\sigma_X^2} = \frac{\sigma_{xy}}{\sigma_x^2+\sigma_u^2} = \boldsymbol{\beta_1 \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}}$

The OLS estimator in regressing regressing y on (1) and X is a "consistent" estimator of $\boldsymbol{\beta_1 \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}}$

Conclusion: Ignoring the measurement error in a simple linear regression can underestimate the magnitude of nonzero slope in a simple linear regression analysis.

Exercise

Create a hypothetical numerical data to support the Conclusion on slide #6

[Hint: generate data, researching on GOOGLE and assuming that all the variables, x, u, ε, are independent normal distributions and setting values to $\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\varepsilon^2$, respectively]

Reference

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons. ISBN 978-0-471-86187-4.