**DISTURBANCE TERMS**

In a field of research design, we often have the question about whether there is a relationship between an observed variable (say, \( y \)) and the other observed variables (say, \( x \)). To answer the question, we may construct the model in which \( y \) depends on \( x \). Because \( y \) is not necessarily explained only by \( x \) from some reasons discussed below, however, there always exists the discrepancy between the observed value of \( y \) and the predicted value of \( y \) obtained from the model. The discrepancy is taken as a disturbance term or an error term.

Suppose that \( n \) sets of data, \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), are observed, where \( y_i \) is a scalar and \( x_i \) is a vector (say, \( 1 \times k \) vector). We assume that there is a relationship between \( x \) and \( y \), which is represented as the model: \( y = f(x) \), where \( f(x) \) is a function of \( x \). We say that \( y \) is explained by \( x \) or \( y \) is regressed on \( x \). \( y \) is called the dependent or explained variable and \( x \) is a vector of the independent or explanatory variables. Suppose that a vector of the unknown parameter (say, \( \beta \), which is a \( 1 \times k \) vector) is included in \( f(x) \). Using the \( n \) sets of data, we consider estimating \( \beta \) in \( f(x) \). Adding a disturbance term (say, \( u \), which is also called an error term), the relationship between \( y \) and \( x \) is given by: \( y = f(x) + u \). The disturbance term \( u \) indicates the term that cannot be explained by \( x \) in \( y \). Usually, \( x \) is assumed to be nonstochastic. Note that \( x \) is said to be nonstochastic when it takes a fixed value. \( f(x) \) is deterministic, while \( u \) is stochastic. \( f(x) \) has to be specified by a researcher. Representatively, \( f(x) \) is often specified as the linear function: \( f(x) = x\beta \).

The reasons why we add the disturbance term \( u \) are as follows: (i) there are some unpredictable elements of randomness in human responses, (ii) an effect of a large number of omitted variables is contained in \( x \), (iii) there is a measurement error in \( y \), (iv) a functional form of \( f(x) \) is not known in general. More details are as follows. For (i), as an example, gross domestic product (GDP) data is observed as a result of human behavior, which is usually unpredictable and is thought of a source of randomness. For (ii), we cannot know all the explanatory variables that depend on \( y \). Most of the variables are omitted, and only the important variables needed for analysis are included in \( x \). The influence of the omitted variables is thought of a source of \( u \). For (iii), some kinds of errors are included in almost all the data, either because of data collection difficulties or because the explained variable is inherently unmeasurable and
a proxy variable has to be used in its stead. For (iv), conventionally we specify $f(x)$ as: $f(x) = x\beta$. However, there is no reason to specify the linear function. Exceptionally, we have the case where the functional form of $f(x)$ comes from the underlying theoretical aspect. Even in this case, however, $f(x)$ is derived from a very limited theoretical aspect, not every theoretical aspect.

For simplicity, hereafter, consider the linear regression model: $y_i = x_i\beta + u_i$, $i = 1, 2, ..., n$. When $u_1, u_2, ..., u_n$ are assumed to be mutually independent and identically distributed with mean zero and variance $\sigma^2$, the sum of squared residuals, $\sum_{i=1}^{n}(y_i - x_i\beta)^2$, is minimized with respect to $\beta$. Then, the estimator of $\beta$ (say, $\hat{\beta}$) is: $\hat{\beta} = (\sum_{i=1}^{n}x_i'x_i)^{-1}\sum_{i=1}^{n}x_i'y_i$, which is called the ordinary least squares (OLS) estimator. $\hat{\beta}$ is known as the best linear unbiased estimator (BLUE). It is distributed as: $N(\beta, \sigma^2(\sum_{i=1}^{n}x_i'x_i)^{-1})$ under normality assumption on $u_i$, because $\hat{\beta}$ is rewritten as: $\hat{\beta} = \beta + (\sum_{i=1}^{n}x_i'x_i)^{-1}\sum_{i=1}^{n}x_i'u_i$. Note from the central limit theorem that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean zero and variance $\sigma^2M_x^{-1}$ even when the disturbance term $u_i$ is not normal, where we have to assume $(1/n)\sum_{i=1}^{n}x_i'x_i \to M_x$ as $n$ goes to infinity, i.e., $n \to \infty$ ($a \to b$ indicates that $a$ approaches $b$).

On the disturbance term $u_i$, we have assumed as follows: (i) $V(u_i) = \sigma^2$ for all $i$, (ii) $\text{Cov}(u_i, u_j) = 0$ for all $i \neq j$, and (iii) $\text{Cov}(u_i, x_j) = 0$ for all $i$ and $j$. In the following, we examine $\hat{\beta}$ in the case where each assumption is violated.

(i) Violation of the Assumption: $V(u_i) = \sigma^2$ for all $i$

When the assumption on variance of $u_i$ is changed to $V(u_i) = \sigma_i^2$, i.e., heteroscedastic disturbance term, the OLS estimator $\hat{\beta}$ is no longer BLUE. The
variance of $\hat{\beta}$ is given by: 
$$V(\hat{\beta}) = (\sum_{i=1}^{n} x_i' x_i)^{-1} (\sum_{i=1}^{n} \sigma_i^2 x_i' x_i) (\sum_{i=1}^{n} x_i' x_i)^{-1}.$$ 
Let $b$ be a solution of minimization of $\sum_{i=1}^{n} (y_i - x_i \beta)^2 / \sigma_i^2$ with respect to $\beta$. Then, 
$$b = (\sum_{i=1}^{n} x_i' x_i / \sigma_i^2)^{-1} \sum_{i=1}^{n} x_i' y_i / \sigma_i^2$$ 
and $b \sim N(\beta, (\sum_{i=1}^{n} x_i' x_i / \sigma_i^2)^{-1})$ are derived under normality assumption on $u_i$. We have the result that $\hat{\beta}$ is not BLUE because of $V(b) \leq V(\hat{\beta})$. The equality holds only when $\sigma_i^2 = \sigma^2$ for all $i$. For estimation, $\sigma_i^2$ has to be specified, e.g., $\sigma_i = |z_i \gamma|$, where $z_i$ represents a vector of the other exogenous variables.

(ii) Violation of the Assumption: $\text{Cov}(u_i, u_j) = 0$ for all $i \neq j$

The correlation between $u_i$ and $u_j$ is called the spatial correlation in the case of cross-section data and the autocorrelation or serial correlation in time series data. Let $\rho_{ij}$ be the correlation coefficient between $u_i$ and $u_j$, where $\rho_{ii} = 1$ for all $i = j$ and $\rho_{ij} = \rho_{ji}$ for all $i \neq j$. That is, we have $\text{Cov}(u_i, u_j) = \sigma^2 \rho_{ij}$. The matrix that the $(i,j)$ th-element is $\rho_{ij}$ should be positive definite. In this situation, the variance of $\hat{\beta}$ is: 
$$V(\hat{\beta}) = \sigma^2 (\sum_{i=1}^{n} x_i' x_i)^{-1} (\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} x_i' x_j) (\sum_{i=1}^{n} x_i' x_i)^{-1}.$$ 
Let $b^*$ be a solution of the minimization problem of 
$$\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} (y_i - x_i \beta)' (y_j - x_j \beta)$$ 
with respect to $\beta$, where $\rho_{ij}$ denotes the $(i,j)$th-element of the inverse matrix of the matrix that the $(i,j)$ th-element is $\rho_{ij}$. Then, under normality assumption on $u_i$, we obtain: 
$$b^* = (\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} x_i' x_j)^{-1} (\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} x_i' y_j)$$ 
and $b^* \sim N(\beta, \sigma^2 (\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} x_i' x_j)^{-1})$.

It can be verified that we obtain the following: $V(b^*) \leq V(\hat{\beta})$. The equality holds only when $\rho_{ij} = 1$ for all $i = j$ and $\rho_{ij} = 0$ for all $i \neq j$. For estimation, we need to
specify $\rho_j$. For an example, we may take the following specification: $\rho_j = \rho^{j-\delta}$, which corresponds to the first-order autocorrelation case (i.e., $u_i = \rho u_{i-1} + \varepsilon_i$, where $\varepsilon_i$ is the independently distributed error term) in time series data. For another example, in the spatial correlation model we may take the form: $\rho_j = 1$ when $i$ is in the neighborhood of $j$ and $\rho_j = 0$ otherwise.

(iii) Violation of the Assumption: $\text{Cov}(u_i, x_j) = 0$ for all $i$ and $j$

If $u_i$ is correlated with $x_j$ for some $i$ and $j$, it is known that $\hat{\beta}$ is not an unbiased estimator of $\beta$, i.e., $E(\hat{\beta}) \neq \beta$, because of $E((\sum_{i=1}^{n} x_i' x_i)^{-1} \sum_{i=1}^{n} x_i' u_i) \neq 0$. In order to obtain a consistent estimator of $\beta$, we need the condition: $(1/n) \sum_{i=1}^{n} x_i' u_i \to 0$ as $n \to \infty$. However, we have the fact: $(1/n) \sum_{i=1}^{n} x_i' u_i \to 0$ as $n \to \infty$ in the case of $\text{Cov}(u_i, x_j) \neq 0$. Therefore, $\hat{\beta}$ is not a consistent estimator of $\beta$, i.e., $\hat{\beta} \not\to \beta$ as $n \to \infty$. To improve this inconsistency problem, we utilize the instrumental variable (say, $z_j$), which satisfies the properties: $(1/n) \sum_{i=1}^{n} z_i' u_i \to 0$ and $(1/n) \sum_{i=1}^{n} z_i' x_i \to 0$ as $n \to \infty$. Then, it is known that $b_{iv} = (\sum_{i=1}^{n} z_i' x_i)^{-1} \sum_{i=1}^{n} z_i' y_i$ is a consistent estimator of $\beta$, i.e., $b_{iv} \to \beta$ as $n \to \infty$. $b_{iv}$ is called the instrumental variable (IV) estimator. It can be also shown that $\sqrt{n}(b_{iv} - \beta)$ is asymptotically normally distributed with mean zero and variance $\sigma^2 M_{xx}^{-1} M_{xz} M_{zz}^{-1}$, where

$(1/n) \sum_{i=1}^{n} z_i' x_i \to M_{zx}$, $(1/n) \sum_{i=1}^{n} z_i' z_i \to M_{zz}$ and $(1/n) \sum_{i=1}^{n} x_i' z_i \to M_{zx} = M_{zx}'$ as $n \to \infty$. As an example of $z_j$, we may choose $z_i = \hat{x}_i$, where $\hat{x}_i$ indicates the predicted value of $x_i$ when $x_i$ is regressed upon the other exogenous variables associated with $x_i$, using OLS.

Hisashi Tanizaki
See also Autocorrelation, Central Limit Theorem, Regression, Serial Correlation, Unbiased Estimator.

FURTHER READINGS
