Moon Oulatta, PhD

0.1 The Simple Linear Regression Model: Selected Issues

A company wishes to examine the effect of **consumer interactions** on the **quality of employee performance** during in-store transactions as means to make better managerial decisions to improve sales. The lead economist at the company naively proposes the following population regression model

$$quality = \gamma_0 + \gamma_1 time + \varepsilon \tag{1}$$

where *quality* (y) denotes a continuous index of performance quality for the sales representative and **time** (z) denotes the total number of minutes that the consumer spends interacting with the sales representative during the in-store transaction. γ_0 and γ_1 are the unknown population parameters to be estimated and ε denotes the error term, which captures any variables that affect performance quality, which are not captured in equation (1). The company has multiple stores in the country and it is impossible to collect transaction data for the entire **population of stores** on a timely basis. Therefore, the lead economist proposes to collect a sample of transactions from the entire population of stores: let us assume that the lead economist relies on a random sampling method (for example, **cluster sampling**, which is the method that I covered in class) to ensure that the data are not subject to a **sample selection bias**. Secondly, we make the assumption that there is **no missing data** and that the **sample size** is relatively large, which is a necessary condition to ensure that the **sampling distribution** of the ordinary least squares (OLS) estimators is **approximately normal**. Lastly, using the **cross sectional data** set, the lead economist estimates the naive **sample regression model** as follows

$$\widehat{quality}_i = \hat{\gamma}_0 + \hat{\gamma}_1 time_i \tag{2}$$

where i denotes a subscript for in-store transaction and $\widehat{quality}_i$ denotes the predicted performance quality scores. $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the OLS estimators of the unknown population parameters. Here, the estimated regression model specified in equation (1) is problematic for multiple reasons:

- The relationship between transaction time and performance quality does not appear to be **unidirectional**. In this case, **simultaneity causation** is prevalent. The more time the sales representative spends interacting with the costumer, the more he can learn about the consumer's preferences and then be in a better position to assist him effectively: therefore, interaction time is expected to improve quality performance. Concomitantly, a stronger sales performance also requires more interaction time with the customer, which means that an increase in quality performance simultaneously causes an increase in interaction time (that's a reasonable theoretical assumption to make). Hence simultaneity causation implies the following

  $$E(\varepsilon|z) \neq 0$$

  which indicates that interaction time is not **mean independent** of the error term (review the concept of independence from class notes) due to simultaneity causation. This is a clear violation of the **Gauss Markov Assumptions** covered in class, because interaction time is now correlated with the error term: $\text{cov}(z,\varepsilon) \neq 0$. Consequently $\hat{\gamma}_1$ is a biased and inconsistent estimator of γ_1: this is shown as follows

  $$E(\hat{\gamma}_1) = \gamma_1 + \frac{cov(z,\varepsilon)}{var(z)}$$

- Another issue that is worth considering here is the **omitted variable bias** problem. It is possible to assume that there are other relevant variables that affect performance quality, which are not

included in equation (1), but are also likely to be correlated to interaction time: for instance, **price discounts** for different items in the store is a potent omitted variable in equation (1). Firstly, because the more different types of price discounts the store offers, the more interaction time would be needed to inform the customer about these available benefits. Secondly, the more benefits the costumer receives, the more she will appreciate the transaction and purchase more items, which implies that the price discounts variable is also an important determinant of performance quality. Given these two conditions, it becomes clear that equation (1) is misspecified. Price discounts should be included in equation (1), but it is not: this implies that it lies in the error term and the correlation between interaction time and the error term is no longer zero, because price discounts are also correlated (but not perfectly) to interaction time. Therefore, it is entirely possible to assume that equation (1) suffers from an omitted variable bias. Which implies that the OLS estimators obtained from equation (2) are biased and inconsistent. Nonetheless, the regression model could be improved by controlling for other important factors (for example, price discounts, etc.). However, addressing the issue simultaneity causation would require a different empirical model (for example, an **instrumental variable approach** would be useful here, but this beyond the scope of this specific lecture).

- Another potential issue is **functional-form misspecification**: if the true relationship between interaction time and performance quality is nonlinear; in other words, let us assume that there is an **optimal interaction time** that maximizes performance quality and past this threshold, any increases in interaction time contributes to a decrease in performance quality (note that this is a reasonable practical assumption). This implication is that the true population regression function should be specified as follows

$$E(quality|time) = \gamma_0 + \gamma_1 time - \gamma_2 time^2$$

  which is different than the estimated regression model specified in equation (2), which does not account for **non-linearity in interaction time**: hence the issue. Functional-form misspecification arises when the sample regression function differs from the population regression function. This issue clearly leads to biased and inconsistent OLS estimators.

- **Observational data** are often subject to imperfections (for example, **macroeconomic data** are extremely difficult to measure and prone to large residual errors and revisions). **Survey data** are often affected by **measurement errors** induced by inaccurate reporting by the survey's participants. The idea is that if you are working with survey data or macroeconomic data: you are more likely to experience data issues emanating from measurement errors. Therefore, it is important to understand the different ways in which these measurement errors in the data could influence the validity of the empirical results derived from a linear regression model. For the simple linear regression model, a **classical measurement error** in the data leads to inefficient or biased and inconsistent OLS estimators. When the classical measurement error emanates from the dependent variable, the OLS estimators remain unbiased, but they are no longer **BLUE**, due to the inflated variance of the estimators caused by the measurement error. Consequently, from class notes, we know that the standard errors of the estimators will be larger than normal when the variance of the OLS estimators is inflated. This lack of efficiency implies that the t-statistics are always going to be lower in light of a classical measurement error in the dependent variable: hence a principal limitation of the simple linear regression model. More importantly, the main point here is that we are more likely to make a **type-2 error** in the inference process when the dependent variable is measured with a single classical measurement error: in other words, we are more likely to make the mistake of accepting the null-hypothesis, when it should actually be rejected.

- Here, I will demonstrate the biasedness proof for the case of the classical measurement error in the independent variable in the simple linear regression model. Firstly, let us assume that equation (1)

represents the true population model and that the model satisfies the **Gaus-Markov assumptions**. Then, in absence of measurement errors in the independent variable, we can show that the OLS estimator is unbiased as follows

$$E(\hat{\gamma}_1) = \frac{cov(\gamma_0, z) + \gamma_1 \cdot var(z) + cov(\varepsilon, z)}{var(z)} \equiv \gamma_1 \tag{3}$$

Now, let us assume that the lead economist cannot perfectly observe and measure interaction time. Therefore, he measures interaction time with a single error. This implies that measured interaction time can be defined as the sum of the true value of interaction time plus the measurement error as follows

$$z_m = z + \varepsilon_z.$$

solving for (z) in the above and reinserting it into equation (1) yields the following model

$$y = \gamma_0 + \gamma_1 z_m + v \tag{4}$$

where equation (4) is the model with the measurement error, which is to be estimated by the lead economist via OLS. Note that the residual term (v) now includes two components ($v = \varepsilon - \gamma_1 \varepsilon_z$), which is where the main problem lies. Here, the measurement error is considered classical as long as it is mean independent:

$$E(\varepsilon|z) = E(\varepsilon) = cov(\varepsilon_z, z) = 0$$

In the presence of a classical measurement error in interaction time, ($\hat{\gamma}_1$) is biased and inconsistent. To prove the latter, let us first assume that the economist estimates equation (4) with the measurement error and obtained the OLS estimator of the slope as follows

$$\hat{\gamma}_1 = \frac{cov(z_m, y)}{var(z_m)}$$

now inserting equation (4) into the above simplifies the solution as follows

$$\hat{\gamma}_1 = \frac{cov(z_m, \gamma_0 + \gamma_1 z_m + v)}{var(z_m)} \dashrightarrow \left\{ \gamma_1 \frac{cov(z_m, z_m)}{var(z_m)} + \frac{cov(z_m, \gamma_0)}{var(z_m)} + \frac{cov(z_m, v)}{var(z) + 2cov(z, \varepsilon_z) + var(\varepsilon_z)} \right\}$$

from class notes, using the covariance rules from the first week of class, the above can be simplified further as follows

$$\hat{\gamma}_1 = \gamma_1 + \frac{cov(z_m, v)}{var(z) + 2cov(z, \varepsilon_z) + var(\varepsilon_z)}$$

now inserting ($v = \varepsilon - \gamma_1 \varepsilon_z$) and ($z_m = z + \varepsilon_z$) into the above yields

$$\hat{\gamma}_1 = \gamma_1 + \frac{cov(z + \varepsilon_z, \varepsilon - \gamma_1 \varepsilon_z)}{var(z) + 2cov(z, \varepsilon_z) + var(\varepsilon_z)}$$

one needs to rely on the Gauss Markov assumptions and the classical error assumption to simplify the above. Effectively, the latter imply that both the error term and the measurement error are mean independent, which indirectly signifies that $cov(z, \varepsilon)$, $cov(\varepsilon_z, \varepsilon)$, and $cov(z, \varepsilon_z)$ are all equal to zero. Given these latter assumptions, the above can be simplified as follows

$$\hat{\gamma}_1 = \gamma_1 + \frac{cov(\varepsilon_z, -\gamma_1 \varepsilon_z)}{var(z) + var(\varepsilon_z)} \dashrightarrow \gamma_1 - \gamma_1 \frac{\sigma^2_{\varepsilon_z}}{(\sigma^2_z + \sigma^2_{\varepsilon_z})} \tag{5}$$

note that in the above, I used the sigma notation, instead of using the var notation, so as to be consistent with your book. By simplifying equation (5), it becomes relatively easy to derive equation (9.33) in your book as follows

$$\hat{\gamma}_1 = \frac{\gamma_1 \sigma_z^2}{(\sigma_z^2 + \sigma_{\varepsilon_z}^2)}$$

then by taking the expectation on both sides, one can show that estimating equation (4) via OLS leads to a biased estimator of the slope as follows

$$E(\hat{\gamma}_1) = \frac{\gamma_1 \sigma_z^2}{(\sigma_z^2 + \sigma_{\varepsilon_z}^2)} \neq \gamma_1 \tag{6}$$

here, mathematically and technically, equation (6) states that the variance of true interaction time (σ_z^2) is always going to be less than the variance of measured interaction time ($\sigma_z^2 + \sigma_{\varepsilon_z}^2$): as long as $\sigma_{\varepsilon_z}^2 > 0$, the ratio of the variances in equation (6) is always going to be less than 1. Which means that the OLS estimator of the slope is always going to be ***biased towards zero***: this is the so called **attenuation bias**. Using the proof, it becomes easy to visualize the bias that arises in a simple linear regression model with a classical error in the explanatory variable. The key here is that the magnitude of the bias depends on the size of the measurement error: holding everything else constant, the larger the measurement error, the more biased the estimate would be towards zero. (i) If $\gamma_1 > 0$, then the OLS method will always **underestimate** the true value of γ_1. Alternatively, if $\gamma_1 < 0$, then the OLS method will always overestimate the true value of γ_1.

- In theory, the classical measurement error assumption holds, but in practice it may not always be the case: for instance, a **non-classical measurement error** in the dependent variable can also lead to biasedness in the OLS estimators. Furthermore, for the **multiple regression model**, the case of a classical measurement error in one independent variable yields more ambiguous conclusions about the OLS estimators: this is discussed further in other chapters.