**GROUP WORK PROJECT #** 1 M5 MScFE 610: FINANCIAL ECONOMETRICS
**GROUP NUMBER:** 9659

| FULL LEGAL NAME | LOCATION (COUNTRY) | EMAIL ADDRESS | MARK X FOR ANY NON-CONTRIBUTING MEMBER |
|---|---|---|---|
| Biswajit Palit | Germany | biswajitpalit23.08.01@gmail.com | |
| | | | |
| | | | |

**Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

| | |
|---|---|
| **Team member 1** | **Biswajit Palit** |
| **Team member 2** | |
| **Team member 3** | |

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
**Note:** You may be required to provide proof of your outreach to non-contributing members upon request.

N/A

**GROUP WORK PROJECT #** 1 M5 MScFE 610: FINANCIAL ECONOMETRICS
**GROUP NUMBER:** 9659

1. **On top left of your screen click on File →Download → Microsoft Word (.docx) to download this template**
2. **Upload the template in Google Drive and share it with your group members**
3. **Delete this page with the requirements before submitting your report. Leaving them will result in an increased similarity score on Turnitin.**

Keep in mind the following:

- Make sure you address all the questions in the GWP assignment document published in the Course Overview.
- Follow the "Submission requirements and format' instructions included in each Group Work Project Assignment, including report length.
- **Including in-text citations and related references is mandatory for all submissions.** You will receive a '0' grade for missing in-text citations and references, or penalties for partial completion. Use the **In-Text Citations and References Guide** to learn how to include them.
- Additional writing aids: Anti-Plagiarism Guide, Academic Writing Guide, Online Writing Resources.
- To avoid an increase in the Turnitin similarity score, **DO NOT copy the questions** from the GWP assignment document.
- Submission format tips:
    - Use the same font type and size and same format throughout your report. You can use Calibri 11, Arial 10, or Times 11.
    - Do NOT split charts, graphs, and tables between two separate pages.
    - Always include the axes labels and scales in your graphs as well as an explanation of how the data should be read.
- Use the LIRN Library for your research. It can be accessed via the left navigation pane inside the WQU learning platform.
- Carefully read **Academic Policy on the use of AI** explaining how the use of AI tools is restricted and regulated. Severe penalties apply for excessive and improper use of AI

***The PDF file with your report must be uploaded separately from the zipped folder that includes any other types of files. This allows Turnitin to generate a similarity report.***

**Problem 1**

In this section, we simulate the equations (1) and (2) given in the question set in order to understand the omitted variable bias.

Equation (1):

$$Y(i) = \alpha + \beta x_i + \gamma w_i + \delta z_i + \varepsilon_i$$

Equation (2):

$$Y(i) = \alpha + \beta x_i + \gamma w_i + \mu_i$$

**Part a.**

It is given that $\varepsilon_i$ follows the standard OLS Assumptions. We work with the following assumptions of the error distribution:

1. Error $\varepsilon_i$ follows a normal distribution, that is, $\varepsilon_i \sim N(0, \sigma^2)$
2. $\varepsilon_i$ is exogenous – Mean of $\varepsilon_i$ is independent of $X_i$. That is: $E[\varepsilon_i \mid X_i] = 0$
3. $\varepsilon_i$ has a constant variance: $Var(\varepsilon_i \mid X) = \sigma^2$

Now, given that Equation (2) omits variable $z_i$, we can say that the error $\mu_i$ can be represented as: $\mu_i = \delta z_i + \varepsilon_i$

Now, let's check whether the exogeneity assumption satisfies for $\mu_i$:

$$E[\mu_i \mid x_i, w_i] = E[\delta z_i + \varepsilon_i \mid x_i, w_i]$$

$$\Rightarrow \delta E[z_i \mid x_i, w_i] + E[\varepsilon_i \mid x_i, w_i]$$

We know that $\varepsilon_i$ is exogenous. Therefore, $E[\varepsilon_i \mid x_i, w_i] = 0$.

But, $E[z_i \mid x_i, w_i] = 0$ iff $\rho(z_i, (x_i, w_i)) = 0$.

That is, if $z_i$ is uncorrelated with both $x_i$ and $w_i$. Otherwise, $E[z_i \mid x_i, w_i] \neq 0$.

Thus, the exogeneity assumption breaks down in the error of Equation (2).

The variance of $\mu_i$ will be as follows:

$$Var(\mu_i) = Var(\delta z_i + \varepsilon_i)$$

$$\Rightarrow \delta^2 Var(z_i) + Var(\varepsilon_i) + 2Cov(z_i, \varepsilon_i)$$

We know that $Var(\varepsilon_i) = \sigma^2$.

Even if the $Cov(z_i, \varepsilon_i) = 0$, the $Var(\mu_i)$ increases by $\delta^2 Var(z_i)$

**Parts b and c.**

Effect of exogeneity failure on $\alpha$, $\beta$, and $\delta$:

We know that OLS provides the Best Linear Unbiased Estimators under the Gauss-Markov Assumptions. We have considered the Gauss-Markov assumptions for Equation 1. Hence, we can call the estimated parameters $\phi = (\alpha, \beta, \gamma)$ BLUE. Being BLUE means that the parameter estimated is unbiased and has the lowest variance (best).

Further, for an estimator to be unbiased, the necessary condition is that the exogeneity assumption is satisfied. From part a, we saw that the exogeneity assumption is satisfied only under certain conditions, where the observed variables are uncorrelated.

**Effect on $\alpha$:**
Since the constant term is chosen to be 1 in our case, it is uncorrelated with the other observed values of $x_i$ and $w_i$. So we can safely say that the $\alpha$ parameter estimated from equation 1 and equation 2 will be unbiased. However, due to an inflated variance of $\mu_i$ in equation 2, we see that the variance of the estimated $\alpha$ in equation 2 is higher than in equation 1.

**Effect on $\beta$:**
$\beta$ will be unbiased iff $E[z_i \mid x_i] = 0$, that is, $x_i$ and $z_i$ are uncorrelated. Regardless, the estimated $\beta$ parameter from equation 2 will have a higher variance. We construct a Monte Carlo simulation with correlated $x_i$, $w_i$, and $z_i$ to capture the effect of omitted variable on the estimated parameter bias.

**Effect on $\delta$:**
Equation 1 estimates BLUE estimates of $\delta$, since $\varepsilon_i$ follows Gauss-Markov assumptions, and equation 2 does not estimate it. The effect of this parameter is captured in the error $\mu_i$.

The following Table 1 summarises the results of the estimated coefficients for 1000 simulations. Here we have generated $z_i$ to be correlated with $x_i$ and $w_i$ to capture the effect of the bias.

In Table 1 we can see that after 1000 simulations, the estimated $\alpha$ in both cases is 1. The t-test with the null hypothesis that the estimated parameter is equal to the true value fails to reject the null hypothesis for both the $\alpha$ s. Thus, we can infer that the omitted variable is not inducing bias if the observed variables are uncorrelated. On the other hand, the $\beta$ and $\gamma$ coefficients for the "Correct" model give unbiased estimates, whereas the misspecified model significantly induces bias at the 5% level. This is because our construction provides a positive correlation between $x_i$, $w_i$ and $z_i$. Unsurprisingly, for each coefficient in the "Omitted" model, the estimates have a higher variance than the "Correct" model.

| Coefficient | Mean | STD | True Val | T-Stat |
|---|---|---|---|---|
| Correct Alpha | 1.000121 | 0.032122 | 1 | 0.003774 |
| Correct Beta | 3.500528 | 0.036530 | 3.5 | 0.014467 |
| Correct Gamma | 5.000122 | 0.031877 | 5 | 0.003833 |
| Omitted Alpha | 0.997468 | 0.086451 | 1 | -0.029284 |
| Omitted Beta | 4.802827 | 0.088259 | 3.5 | 14.761425 |
| Omitted Gamma | 5.518642 | 0.089230 | 5 | 5.812426 |

Table 1: Model Coefficient Comparison: Correct vs Omitted

**Part d.**

The simulation exercise is performed in the zip file. The findings are noted here as follows:

A. Generating Independent Random Variables

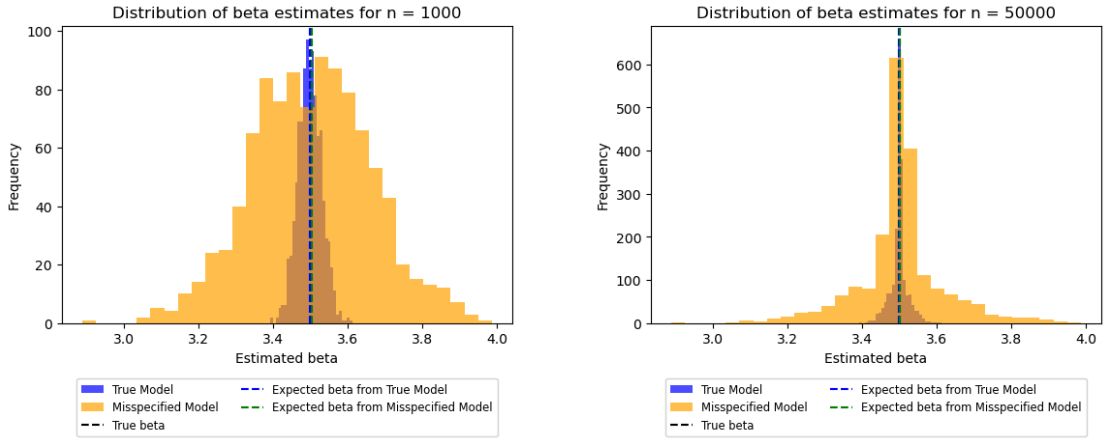


Figure 1: Distribution of estimated parameters for uncorrelated random variables

In Figure 1, we note that when the observed values of the variables are uncorrelated, the estimator $\beta$ is unbiased. But an interesting trend is noted when increasing the sample size from 1000 to 50000. When we increase the sample size, the estimated coefficients tend to get closer to the true value. This property is the consistency of the OLS estimator. Omitted variable does not tamper with the consistency of the OLS estimator, simply biases it, and increases its variance with the same sample size. We observe that when the sample size increases, the frequency of estimated parameters being close to the true value rises, and the tails become flatter. We can say that $\hat{\beta}$ converges to $\beta$ as n tends to $\infty$ .

B. Generating Correlated Random Variables

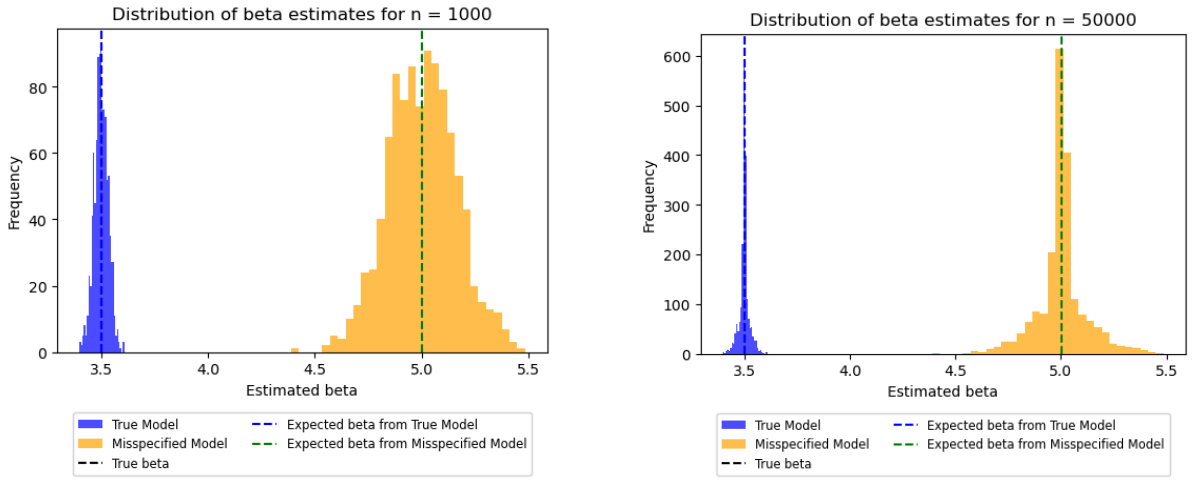


Figure 2: Distribution of estimated parameters for correlated random variables

In Figure 2, we induce positive bias in the estimated coefficients by creating positively correlated X and W, along with a positive true value coefficient for W. Even in this case, we see, raising the sample size makes the estimated β converge to the mean (which is now biased).

**Problem 6**

The given equation is as follows:

$$(A) \quad Y(t) = \alpha + \beta X(t) + \varepsilon(t) \qquad t=1, 2, .., 20$$

However, there is a structural break in the parameter β at t = 10.

To test this in a single equation, we use the dummy variable approach. We assign the dummy 0 if t <= 10, and the dummy 1 if t > 10. Thus, the revised equation to estimate β is as follows:

$$(B) \quad Y(t) = \alpha + \beta 1 X(t) + \delta(Dt \cdot Xt) + \varepsilon(t)$$

Thus, after the structural break, t > 10, the coefficient becomes β2 = β1 + δ. And before the structural break, t <= 10, β1 is the coefficient of interest.

We run a simulation study to verify whether this approach can correctly capture the effect of the structural break and provide a better model fit. We run 1000 simulations. We generate t = 20 timestamps. And for each simulation, we generate a random variable X. We then add the time dummy variable and create the treatment variable Dt · Xt. For each simulation, we create the random variable Y

using equation A. However, we differentially multiply beta to X based on the timestamp. An observation at timestamp 5 will be multiplied by one value of beta, exogenously provided, and a value at timestamp 15 will be multiplied by another beta, taking into account the structural break.

Now, we run regressions A and B separately for each simulation and record the distribution of the estimated coefficients. We also record the mean RMSE of the models across simulations. Our results are as follows:

1. Average RMSE of Equation A across 1000 sims is: 1.65
2. Average RMSE of Equation B across 1000 sims is: 0.91

Thus, regression B provides a better fit to the model compared to equation A. We also plot the distributions of the estimated β1, β2 and δ coefficients from regression B. We find that the mean of the estimates confirms with our construction and corresponds to the set true values.
We had set the true values of the coefficients as follows: β1 = 2, β2 = 5. Thus, the structural break should be β2 - β1 = 3. The following figure shows us the results we achieved.

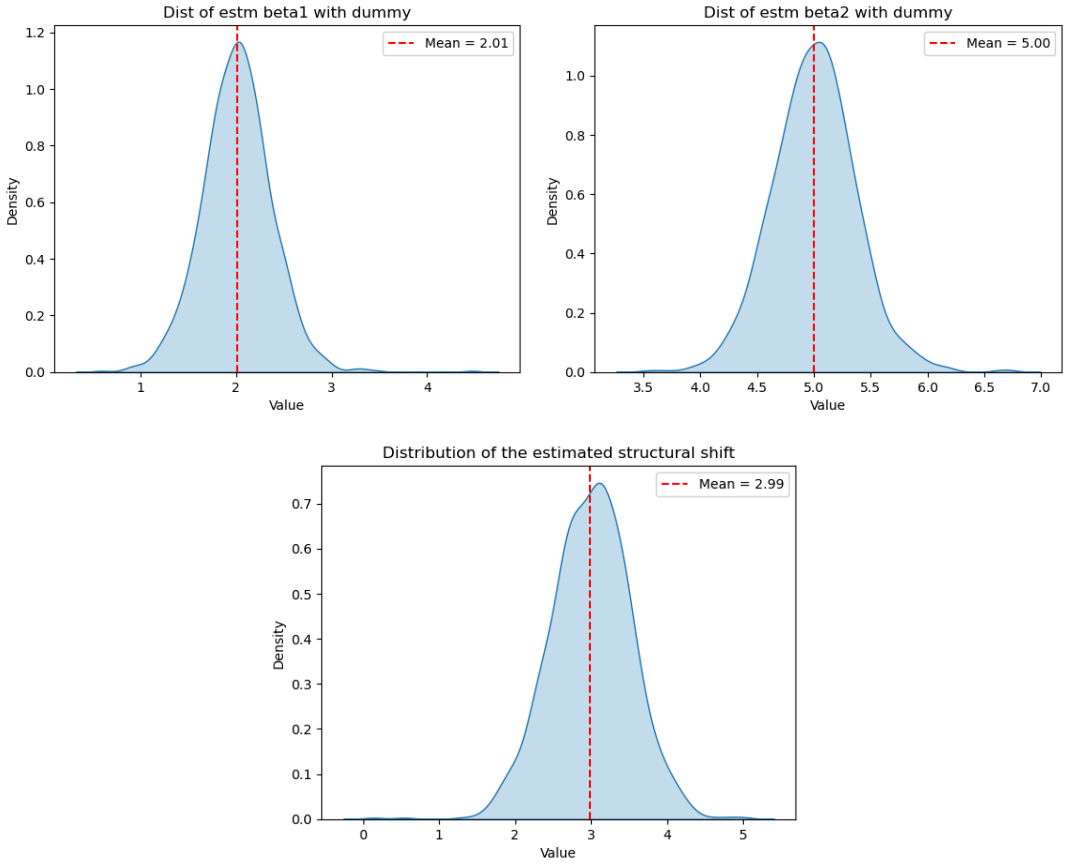


Figure 3: Distribution of the estimated coefficients from Dummy Variable Regression B

We can see that the mean of the estimated coefficients converges to the true value, thus giving a better picture in regression B than in regression A. The codes for the simulation are provided in the zipped file.