STAT 215B Assignment 4

Xiaowei Zeng

March 12, 2024

Abstract

The main job in this assignment is to study the performance of OLS and IVLS coefficient estimators, as well as two different estimators of the error variance using simulation. The first section introduces the model we assume. The second section depicts the simulation setting and results.

1 Model

Consider the model

$$\begin{aligned} Y_i &= X_i + \epsilon_i, \\ X_i &= U_i + 2V_i + \delta_i. \end{aligned} \tag{1}$$

Everything in sight is scalar. The vectors

$$(U_i, V_i, \epsilon_i, \delta_i), i = 1, ..., n,$$

are i.i.d. across i. Each vector is normal with mean zero.

For each i we take the three random objects (1) U_i, (2) V_i, and (3) (ϵ_i, δ_i) to be mutually independent; furthermore,

$$\begin{aligned} Var(U_i) &= Var(V_i) = 1 \\ Var(\epsilon_i) &= Var(\delta_i) = \sigma^2, \text{ and} \\ Cov(\epsilon_i, \delta_i) &= \rho. \end{aligned}$$

1.1 Endogeneity of X_i

By definition, a regressor X is call endogenous when $\mathbb{E}(\epsilon X) \neq 0$.

In (1), we observe that X_i depends on three random variables: U_i, V_i, and δ_i. Among these variables, U_i and V_i are independent of ϵ_i, but δ_i and ϵ_i is correlated, which may be attributed to the presence of unmeasured confounding. Thus, X_i and ϵ_i are correlated with $\mathbb{E}(\epsilon X) \neq 0$, indicating that X_i is endogenous.

1.2 Instrumental Variables

By definition, a **good** instrumental variable Z must satisfy the following conditions (Ding, 2023):

- Z is independent of the unmeasured confounding.
- Z is correlated to X, which ensures that Z is useful for estimation.
- Z should affect the outcome Y only indirectly through X but not directly.

In this context, we know:

- U and V are independent of ϵ_i, where ϵ_i includes the unmeasured confounding.
- U and V are correlated to X because $X_i = U_i + 2V_i + \delta_i$ in (1).
- We don't really know whether indirectly or directly U and V affect Y since there's no background knowledge about the model, and it's meaningless to discuss the causal mechanism without background information. So we can simply assume that in this model they only affect Y indirectly by X.

Therefore, U_i and V_i are instrumental variables.

2 Simulation

2.1 Setting

Suppose $\beta = 3, \sigma^2 = 1$, and $\rho = 3/4$. For each of 1,000 simulation runs, generate $n = 100$ independent realizations of the vector $(U_i, V_i, \epsilon_i, \delta_i, X_i, Y_i)$ according to (1). Use OLS to obtain the estimate $\hat{\beta}_{OLS}$, and IVLS to obtain $\hat{\beta}_{IVLS}$.

For each simulation, estimate the error variance σ^2 in two ways: first using the residuals obtained from plugging $\hat{\beta}_{IVLS}$ into (1), then using the residuals from the transformed equation (Freedman, 2009)

$$(Z^T Z)^{-1/2} Z^T Y = (Z^T Z)^{-1/2} Z^T X\beta + \eta, \tag{2}$$

where Z is the $n \times 2$ matrix of instruments, X is the $n \times 1$ design matrix, Y is the $n \times 1$ vector of responses, and $\eta = (Z^T Z)^{-1/2} Z^T \epsilon$.

2.2 Estimation Formula

Original least squares (OLS) estimation:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

```
beta_ols <- solve(t(X) %*% X) %*% t(X) %*% Y.
```

Instrumental variables least squares (IVLS) estimation:

$$\hat{\beta}_{2SLS} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y,$$

where $\hat{X}$ is the projection of X onto Z, or

$$\hat{\beta}_{IVLS} = (X^T P_Z X)^{-1} X^T P_Z Y,$$

where $P_Z = Z(Z^T Z)^{-1} Z^T$ is the projection matrix of Z. The two-stage least squares method and the generalized least squares estimator for IV give exactly the same results.

```r
# Projection Matrix of Z
P_Z <- Z %*% solve(t(Z) %*% Z) %*% t(Z)
# Two Stage Least Squares
X_hat <- P_Z %*% X
beta_2sls <- solve(t(X_hat) %*% X_hat) %*% t(X_hat) %*% Y
# Generalized Least Squares for IV
beta_ivls <- solve(t(X) %*% P_Z %*% X) %*% t(X) %*% P_Z %*% Y
# Verify
beta_2sls == beta_ivls
```

To estimate the error variance σ^2, we have two ways:

1. Plug $\hat{\beta}_{IVLS}$ into (1) and obtain the residuals

   $$\hat{\epsilon} = Y - X\hat{\beta}_{IVLS}.$$

   σ^2 is then unbiasedly estimated as (Ding, 2024)

   $$\hat{\sigma}^2 = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n-1}.$$

2. Use the transformed equation in (2) and obtain the residuals

   $$\hat{\eta} = (Z^TZ)^{-1/2}Z^T(Y - X\hat{\beta}_{IVLS}) = P_Z^{1/2}\hat{\epsilon},$$

   where $P_Z = P_Z^{1/2^T} P_Z^{1/2}$.
   We know that

   $$\mathrm{Cov}(Y \mid Z) = \sigma^2 I_n.$$

   Therefore,

   $$\begin{aligned}
   \hat{\eta} &= P_Z^{1/2}(Y - X\hat{\beta}_{IVLS}) \\
   &= P_Z^{1/2}(Y - X(X^TP_ZX)^{-1}X^TP_ZY) \\
   &= P_Z^{1/2}(I_n - X(X^TP_ZX)^{-1}X^TP_Z)Y
   \end{aligned}$$

   $$\begin{aligned}
   \mathbb{E}(\hat{\eta}\hat{\eta}^T \mid Z) &= \mathrm{Cov}(\hat{\eta} \mid Z) \\
   &= P_Z^{1/2}(I_n - X(X^TP_ZX)^{-1}X^TP_Z)\mathrm{Cov}(Y \mid Z)(I_n - P_ZX(X^TP_ZX)^{-1}X^T)P_Z^{1/2^T} \\
   &= P_Z^{1/2}(I_n - X(X^TP_ZX)^{-1}X^TP_Z)\sigma^2 I_n(I_n - P_ZX(X^TP_ZX)^{-1}X^T)P_Z^{1/2^T} \\
   &= \sigma^2 P_Z^{1/2}(I_n - X(X^TP_ZX)^{-1}X^TP_Z)(I_n - P_ZX(X^TP_ZX)^{-1}X^T)P_Z^{1/2^T} \\
   \mathbb{E}(\hat{\eta}^T\hat{\eta} \mid Z) &= \mathbb{E}\{\mathrm{trace}(\hat{\eta}\hat{\eta}^T) \mid Z\} \\
   &= \sigma^2\mathrm{trace}\{P_Z^{1/2}(I_n - X(X^TP_ZX)^{-1}X^TP_Z)(I_n - P_ZX(X^TP_ZX)^{-1}X^T)P_Z^{1/2^T}\} \\
   &= \sigma^2\mathrm{trace}\{(I_n - P_ZX(X^TP_ZX)^{-1}X^T - X(X^TP_ZX)^{-1}X^TP_Z \\
   &\qquad\qquad + X(X^TP_ZX)^{-1}X^T)P_Z\} \\
   &= \sigma^2\mathrm{trace}\{P_Z - P_ZX(X^TP_ZX)^{-1}X^TP_Z\} \\
   &= \sigma^2\mathrm{trace}(P_Z) - \sigma^2\mathrm{trace}\{(X^TP_ZX)^{-1}X^TP_ZX\} \\
   &= \sigma^2(2-1) = \sigma^2
   \end{aligned}$$

   σ^2 is then estimated as

   $$\hat{\sigma}^2 = \hat{\eta}^T\hat{\eta}.$$

2.3 Results

The distributions for the two estimators of β are shown in Figure 1. It is obvious that the OLS estimator is biased and inconsistent but with lower standard deviation, while the IVLS estimator is consistent with a bit higher standard deviation (Table 1). The root mean square error (RMSE) of IVLS estimator is much smaller than that of OLS estimator, verifying the effectiveness of IV model.

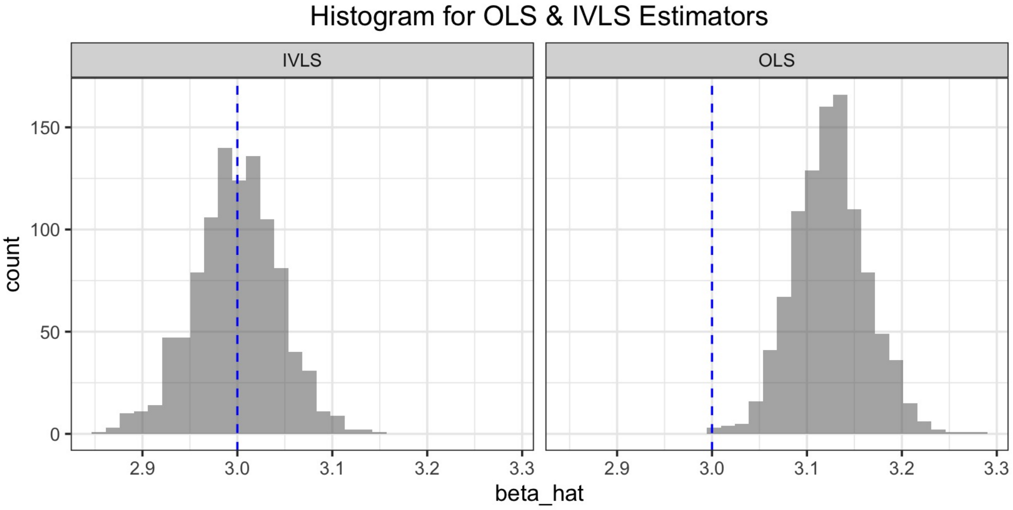

Figure 1: Histogram for the two estimators of $\hat{\beta}$

Table 1: Summary for $\hat{\beta}$

	Mean	SD	RMSE
OLS	3.124	0.039	0.130
IVLS	2.999	0.044	0.044

1. Proof of biasness of $\hat{\beta}_{OLS}$:

$$\mathbb{E}[\hat{\beta}_{OLS} \mid X] = \beta + (X^T X)^{-1} X^T \mathbb{E}[\delta \mid X] \neq \beta,$$

where $(X^T X)^{-1} X^T \mathbb{E}[\delta \mid X]$ is called the ”endogeneity bias”.

2. Proof of inconsistency of $\hat{\beta}_{OLS}$:

$$\begin{aligned}\hat{\beta}_{OLS} &= \beta + (X^T X)^{-1} X^T \delta \\ &= \beta + \left(\frac{1}{n} X^T X\right)^{-1} \left(\frac{1}{n} X^T \delta\right) \\ &\to \beta + \mathbb{E}[x_n x_n^T]^{-1} \mathbb{E}[x_n \delta_n] \neq \beta, \quad N \to \infty.\end{aligned}$$

3. Proof of biasness of $\hat{\beta}_{IVLS}$ when $p = q$:

$$\mathbb{E}[\hat{\beta}_{IVLS} \mid Z] = \beta + \mathbb{E}[(Z^T X)^{-1} Z^T \delta \mid Z] \neq \beta,$$

where X is still correlated with δ and thus $\mathbb{E}[(Z^T X)^{-1} Z^T \delta \mid Z] \neq 0$.

4. Proof of consistency of $\hat{\beta}_{IVLS}$ when $p = q$:

$$\begin{aligned}\hat{\beta}_{IVLS} &= \beta + (Z^T X)^{-1} Z^T \delta \\ &= \beta + \left(\frac{1}{n} Z^T X\right)^{-1} \left(\frac{1}{n} Z^T \delta\right) \\ &\to \beta + \mathbb{E}[z_n x_n^T]^{-1} \mathbb{E}[z_n \delta_n] = \beta, \quad N \to \infty,\end{aligned}$$

indicating that $\hat{\beta}_{IVLS}$ is unbiased but consistent.

The relative merits of OLS versus IVLS:

- According to Gauss Markov Theorem, $\hat{\beta}_{OLS}$ is best linear unbiased estimator (BLUE), namely the linear estimator with the smallest variance, among all unbiased linear estimators for β under Gauss Markov Assumptions which include the exogeneity of the regressors. When endogeneity exists, OLS estimator $\hat{\beta}_{OLS}$ is biased and not consistent.

- By contrast, the IVLS, though with larger variance, can give consistent estimate for β and thus address endogeneity by using instrumental variables that are correlated with the regressors but uncorrelated with the error term (confounding).

The distributions for the two estimators of σ^2 are shown in Figure 2. Figure 3 is the zoomed version of the histogram for $\hat{\epsilon}^T\hat{\epsilon}/(n-1)$, showing that the first estimator, $\hat{\epsilon}^T\hat{\epsilon}/(n-1)$, has a nearly symmetric distribution with a small variance.

On the contrary, in the right panel of Figure 2, we can observe that the second estimator, $\hat{\eta}^T\hat{\eta}$, has a highly right-skewed distribution with a large variance. The results in Table 2 lead to the same conclusion that $\hat{\epsilon}^T\hat{\epsilon}/(n-1)$ should be a better estimator of σ^2 because it has the mean closest to true σ^2, the lowest standard deviation and RMSE.

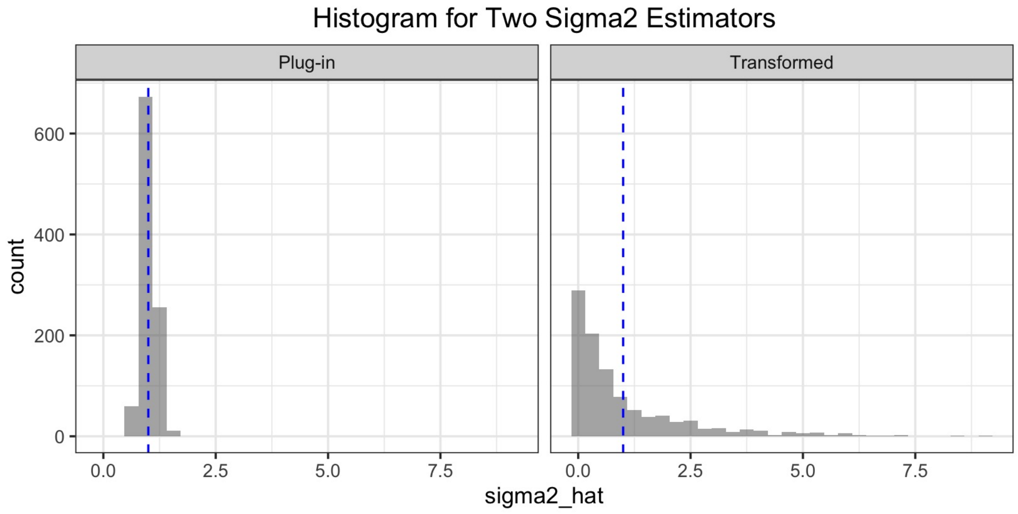

Figure 2: Histogram for the two estimators of σ^2

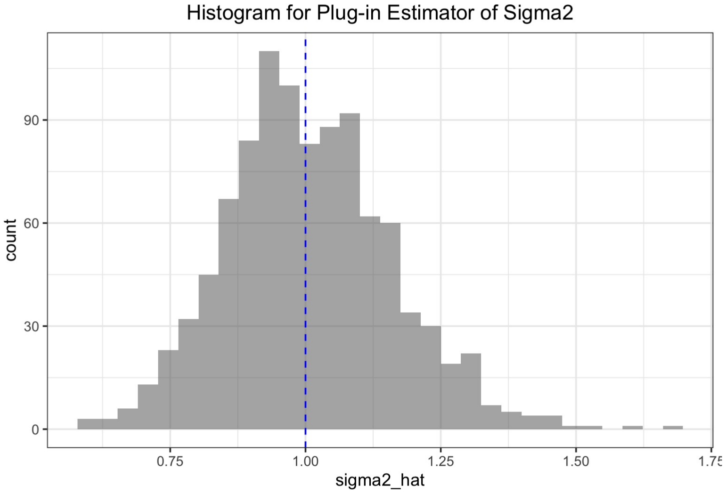

Figure 3: Histogram for $\hat{\epsilon}^T\hat{\epsilon}/(n-1)$

Table 2: Summary for $\hat{\sigma}^2$

	Mean	SD	RMSE
$\hat{\epsilon}^T\hat{\epsilon}/(n-1)$	1.007	0.154	0.154
$\hat{\eta}^T\hat{\eta}$	1.017	1.353	1.353

References

Ding, P. (2023). A first course in causal inference.

Ding, P. (2024). Linear model and extensions.

Freedman, D. A. (2009). *Statistical models: theory and practice.* cambridge university press.