This is a supplementary appendix to "Tail Events in Venture Capital Returns."

1 GLS with endogeneous variables

I discuss below a generalized model of an endogenous variable that is also explicitly impacts the variance of the outcome equation.

1.1 Simple model

The following discussion borrows from Hamilton (1994). Suppose the outcome equation of interest is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

and let x_1 be endogenous. There exists a variable x_3 that is exogenous in the outcome equation and correlated with x_1. Write

$$x_{1i} = \gamma_0 + \gamma_1 x_{3i} + \gamma_2 x_{2i} + u_i$$

Define $z' = (1, x_2, x_3)$. Then the estimate of $\gamma = [\gamma_0, \gamma_1, \gamma_2]$ is

$$\hat{\gamma} = \left[\sum_{i=1}^{N} z_i z_i'\right]^{-1} \left[\sum_{i=1}^{N} z_i x_{1i}\right]$$

From this estimate define

$$\hat{x} = [\hat{\gamma}' z, x_2]$$

The standard 2SLS estimator for $\beta = [\beta_1 \beta_2]$ is

$$\beta_{2SLS} = \left[\sum_{i=1}^{N} \hat{x}_i x_i'\right]^{-1} \left[\sum_{i=1}^{N} \hat{x}_i y_i\right]$$

1.2 Why GLS is biased

Suppose that we utilize the information that $\epsilon \sim N(0, \sigma^2 x_1)$ and divide all the regressors in the second stage by $\sqrt{x_{i1}}$. The derivation below shows that this leads to a biased estimate of β.

$$\tilde{\hat{x}}_i' = [\hat{\gamma} z_i / \sqrt{x_{1i}}, x_2 / \sqrt{x_{1i}}]$$

$$\tilde{x}_i = x_i / \sqrt{x_{1i}}$$

So the 2SLS-GLS estimator is

$$\beta_{2SLS-GLS} = \left[\sum_{i=1}^{N} \tilde{\hat{x}}_i \tilde{x}_i'\right]^{-1} \left[\sum_{i=1}^{N} \tilde{\hat{x}}_i \tilde{y}_i\right] = \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\hat{x}}_i \tilde{x}_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\hat{x}}_i (\tilde{x}_i \beta + \tilde{\epsilon}_i)\right] \tag{1}$$

$$= \beta + \left[\frac{1}{N}\sum \tilde{\hat{x}}_i \tilde{x}_i'\right]^{-1} \frac{1}{N}\sum_{i=1}^{N} \tilde{\hat{x}}_i \tilde{\epsilon}_i \tag{2}$$

This estimator is biased because

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \hat{\gamma} z_i / x_{1i} \\ x_{2i}/x_{1i} \end{bmatrix} \epsilon_i = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \hat{\gamma} z_i \\ x_{2i} \end{bmatrix} \frac{\epsilon_i}{x_{1i}} \neq 0$$

as $E[x_1 \epsilon] \neq 0$.

GLS with fitted values

I now repeat the exercise above, but replace the term $\sqrt{x_{1i}}$ with the fitted value from the first stage regression, $\sqrt{\hat{x}_{1i}}$. Note that

$$\hat{x}_{1i} = \hat{\gamma} z_i$$

We have

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \hat{\gamma} z_i / \sqrt{\hat{x}_{1i}} \\ x_{2i}/\sqrt{\hat{x}_{1i}} \end{bmatrix} \tilde{\epsilon}_i = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \hat{\gamma} z_i \\ x_{2i} \end{bmatrix} \frac{\epsilon_i}{\hat{x}_{1i}} \tag{3}$$

$$= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \hat{\gamma} z_i \\ x_{2i} \end{bmatrix} \frac{\epsilon_i}{\hat{\gamma} z_i} \tag{4}$$

$$= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_{2i}/\hat{\gamma} z_i \end{bmatrix} \epsilon_i = 0 \tag{5}$$

where the last step follows from $E[\epsilon] = E[x_2 \epsilon] = 0$. So the GLS correction should use the fitted values of the endogenous variable from the first stage regression. Next, I run some simple simulations showing the bias of each estimator.

1.3 Simulation results

The model presented above is simulated using N random draws S times. The simulation works as follows:

1. Generate N draws of $r \sim N(0, \sigma_r^2)$.

2. Generate N draws of $u = \sigma_u * r$

3. Generate N draws of $z = \mu_z + \sigma_z Z \qquad Z \sim N(0,1)$ and $x_2 = \mu_2 + \sigma_2 X_2, X_2 \sim N(0,1)$.

4. Compute $x_1 = \gamma_0 + \gamma_1 x_3 + \gamma_2 x_2 + u$ where γ is fixed.

5. Generate N draws of $\epsilon = \sigma_\epsilon x_1 E \qquad E \sim N(0,1)$

6. Compute $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \epsilon$

7. Run OLS of y on a constant, x_1 and x_2. Save $\hat{\beta}$.

8. Run 2SLS. Compute $\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_1 x_3 + \hat{\gamma}_3 x_2$ from OLS of x_1 on a constant, x_3 and x_2.

9. Run OLS of y on a constant, $\hat{x}_1$ and x_2. Save $\hat{\beta}_{IV}$.

10. Compute $\tilde{y} = y/\sqrt{x_1}$, $\tilde{c} = 1/\sqrt{x}$, $\tilde{x}_1 = \hat{x}_1/\sqrt{x_1}$ and $\tilde{x}_2 = x_2/\sqrt{x_1}$.

11. Run OLS of $\tilde{y}$ on $\tilde{c}$, $\tilde{x}_1$ and $\tilde{x}_2$. Save $\hat{\beta}_{IVGLS}$.

12. Compute $\bar{y} = y/\sqrt{\hat{x}_1}$, $\bar{c} = 1/\sqrt{\hat{x}_1}$, $\bar{x}_1 = \hat{x}_1/\sqrt{\hat{x}_1}$ and $\bar{x}_2 = x_2/\sqrt{\hat{x}_1}$.

13. Run OLS of $\bar{y}$ on $\bar{c}$, $\bar{x}_1$ and $\bar{x_2}$. Save $\hat{\beta}_{IVGLS2}$.

Repeat these steps S times. We then have S estimates of $(\hat{\beta}, \hat{\beta}_{IV}, \hat{\beta}_{GLS}, \hat{\beta}_{GLSIV})$. We are ultimately interested in any bias of these estimators. First, we know that the standard OLS estimate of β will be biased. I am interested in the the bias of β_1 and β_2 for a range of γ_1 and σ_r. These two constants measure the relevancy of the instrument and the covariance of the two error terms. The fixed parameters are

$$\beta_1 = 1.5 \quad \beta_2 = 4 \quad \mu_z = 1 \quad \mu_{x_2} = 2.5 \quad \sigma_{x2} = 2.4$$

$$\sigma_z = 1.1 \quad \sigma_\epsilon = .6\sigma \quad \gamma_2 = 2.2 \quad \sigma_u = .5$$

Finally, I set $N = S = 1000$. Figure 1 shows that the bias of the IV estimator increases (in absolute terms) as the variance of r increases, while for very large γ_1, the bias is minimal. Figure 2 shows the bias of the adjusted IV+GLS estimator that uses the fitted values of the endogenous variable in the denominator. The most significant absolute bias occurs when the instrument is weak: $\gamma_1 \approx 0$. Overall, using the fitted values for the GLS correction eliminates most of the pervasive bias.

References

Hamilton, James, *Time Series Analysis*, Princeton University Press, 1994.

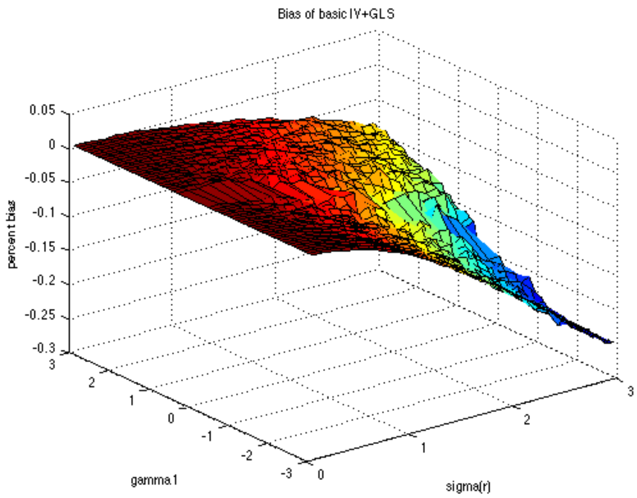

Figure 1: Bias of IV+GLS of β_1 with endogenous denominator

2 Figures

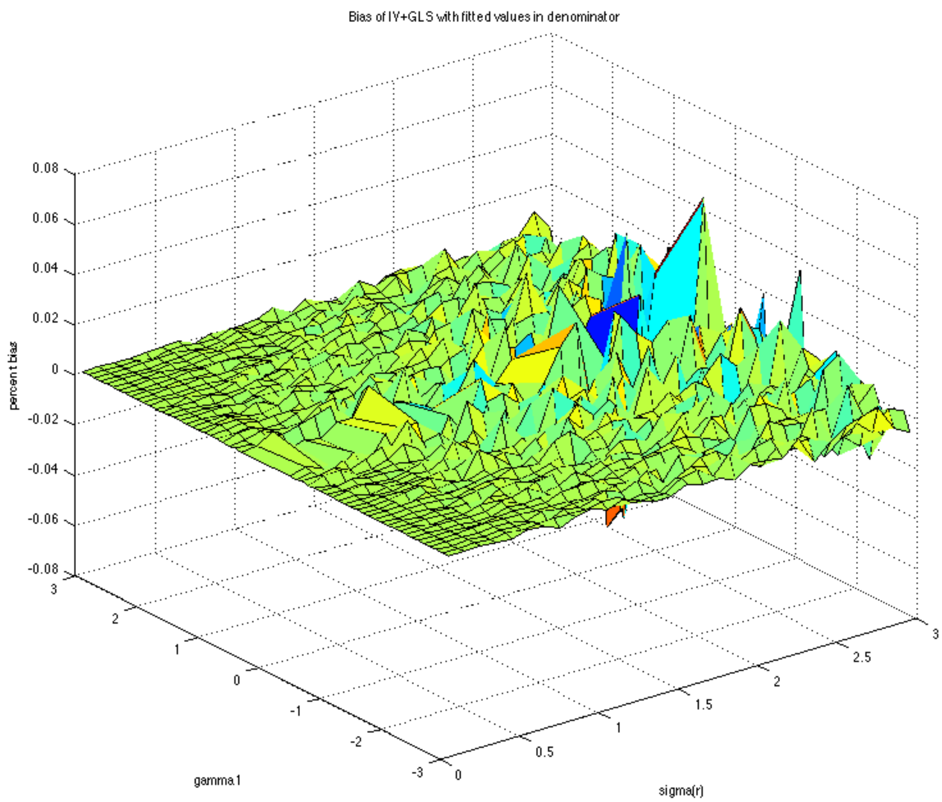

Figure 2: Bias of IV+GLS of β_1 with fitted denominator