Economics 522A, Homework 10 Suggested Solutions

1. (a) We need to assume that the second moment matrix of (x_i, y_i, u_i) is finite and that $V(x_i) > 0$. Recall that we can write

$$\hat{\beta}_2 = \frac{\widehat{Cov}(\tilde{x}_i, y_i)}{\widehat{Var}(\tilde{x}_i)},$$

where $\widehat{Cov}$ is the sample covariance, and $\widehat{Var}$ is the sample variance. By the law of large numbers,

$$\hat{\beta}_2 \xrightarrow{p} \frac{Cov(\tilde{x}_i, y_i)}{Var(\tilde{x}_i)}.$$

Write

$$\begin{aligned} \tilde{x}_i &= x_i + u_i, \\ y_i &= \beta_1 + \beta_2 x_i + \epsilon_i. \end{aligned}$$

Then

$$Cov(\tilde{x}_i, y_i) = Cov(x_i + u_i, \beta_1 + \beta_2 x_i + \epsilon_i) = \beta_2 Var(x_i).$$

Also,

$$Var(\tilde{x}_i) = Var(x_i) + Var(u_i).$$

So

$$plim \hat{\beta}_2 = \frac{\beta_2 Var(x_i)}{Var(x_i) + Var(u_i)}.$$

If $Var(u_i) > 0$, then the probability limit of OLS will be closer to 0 ("attenuated") than β_2.

(b) Consider the alternative estimator

$$\tilde{\beta}_2 = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, \tilde{x}_i)}.$$

Note that

$$Cov(z_i, y_i) = Cov(x_i + v_i, \beta_1 + \beta_2 x_i + \epsilon_i) = \beta_2 Var(x_i).$$

Likewise

$$Cov(z_i, \tilde{x}_i) = Cov(x_i + v_i, x_i + u_i) = Var(x_i).$$

So, provided the second moments are finite and $Var(x_i) > 0$,

$$plim \tilde{\beta}_2 = \beta_2 \frac{Var(x_i)}{Var(x_i)} = \beta_2.$$

2. Suppose that

$$(x_i, u_i)' \overset{\text{iid}}{\sim} N(0, \Sigma),$$

where Σ is a 2×2 covariance matrix. Also suppose that

$$y_i = \delta_1 + \delta_2 x_i + u_i.$$

(a) Let

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

By the properties of the multivariate normal model,

$$u_i|x_i \sim N\left(\frac{\sigma_{12}}{\sigma_{11}} x_i, \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right).$$

Therefore,

$$y_i|x_i \sim \delta_1 + \delta_2 x_i + N\left(\frac{\sigma_{12}}{\sigma_{11}} x_i, \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right) \sim N\left(\delta_1 + \delta_2 x_i + \frac{\sigma_{12}}{\sigma_{11}} x_i, \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right).$$

(b) There are many possible answers for this part. Suppose that $\delta_1 = \delta_2 = 1$, and $\sigma_{12} = 0$ and $\sigma_{11} = \sigma_{22} = 1$. Then by our previous result,

$$y_i|x_i \sim N(1 + x_i, 1).$$

Now consider an alternative set of parameter values,

$$\tilde{\delta}_1 = 1, \quad \tilde{\delta}_2 = 0, \quad \tilde{\sigma_{12}} = \tilde{\sigma_{11}} = 1, \quad \tilde{\sigma_{22}} = 2.$$

Then

$$y_i|x_i \sim \tilde{\delta}_1 + \tilde{\delta}_2 x_i + N(x_i, 1) \sim N(1 + x_i, 1),$$

which is the same conditional distribution. This shows that the conditional distribution of y_i given x_i is the same under the two sets of parameter values. (In addition, the marginal distribution of x_i is the same.)

So two sets of parameter values yield the same distribution for the observable variables. There is no way we can tell, from observations on (x_i, y_i), which are the "true" parameter values. Hence, we cannot hope to estimate the parameters consistently.

3. The program `hw10_3.m` does the calculations for this part.

(a) By standard results for IV estimators, we have

$$\sqrt{n}(b - \beta) \xrightarrow{d} N(0, \Delta),$$

where

$$\Delta = \frac{E(u_i^2 z_i^2)}{(E(z_i x_i))^2}.$$

Note that by independence

$$E(u_i^2 z_i^2) = E(u_i^2)E(z_i^2) = 1 \cdot 1 = 1.$$

We can also calculate

$$E(z_i x_i) = E(z_i(\gamma z + \gamma v + \lambda u)) = \gamma E(z_i^2) + 0 = \gamma.$$

So, $\Delta = 1/\gamma^2$, and

$$\sqrt{n}(b - \beta) \xrightarrow{d} N(0, 1/\gamma^2).$$

Thus, the approximate distribution of b is

$$N(\beta, 1/(\gamma^2 n)) \sim N(1, 1/400) \sim N(1, .0025).$$

From the Monte Carlo study, the sample mean and variance of b are:

```
mean = .9973
var = .0027
```

This matches closely the theoretical counterparts. Also, the histogram of b, shown in Figure 1, appears to be normal. So the asymptotic result seems to give a good approximation in this particular case.

(b) Now we set $\gamma = .5$. Then the approximate mean of b is still 1, and the approximate variance is $(1/.5^2) \cdot (1/100) = .04$.

```
mean =     0.9548
var =      0.0564
```

Again these are fairly close to the asymptotic approximations. However, the histogram in Figure 2 shows that the distribution of b appears to be asymmetric. So the normal approximation fails to capture this aspect of the sampling distribution of b.

Note: in part (b) and especially part (c), one can occasionally get extreme samples for b, which lead to varying results for the mean and the variance of b. So your simulation results might differ from what is obtained here.

(c) Now we set $\gamma = .1$. Then the approximate variance of b is $(1/.1^2) \cdot (1/100) = 1$.

```
mean =     0.1914
var =    1.6574e+03
```

Now the mean and variance are well off from the approximations. Also, examining the histogram in Figure 3, we see that the normal approximation is quite poor. The histogram appears to be bimodal, with one mode near 1 and another mode close to 2.5. (This is not an artifact of the sampling - it can be shown that the exact distribution is bimodal.) So the normal approximation is quite poor in this case.

It appears that if the correlation between x and z is low, the asymptotic distribution may give a poor approximation to the finite sample properties of b.

Figure 1: Question 3 (a)

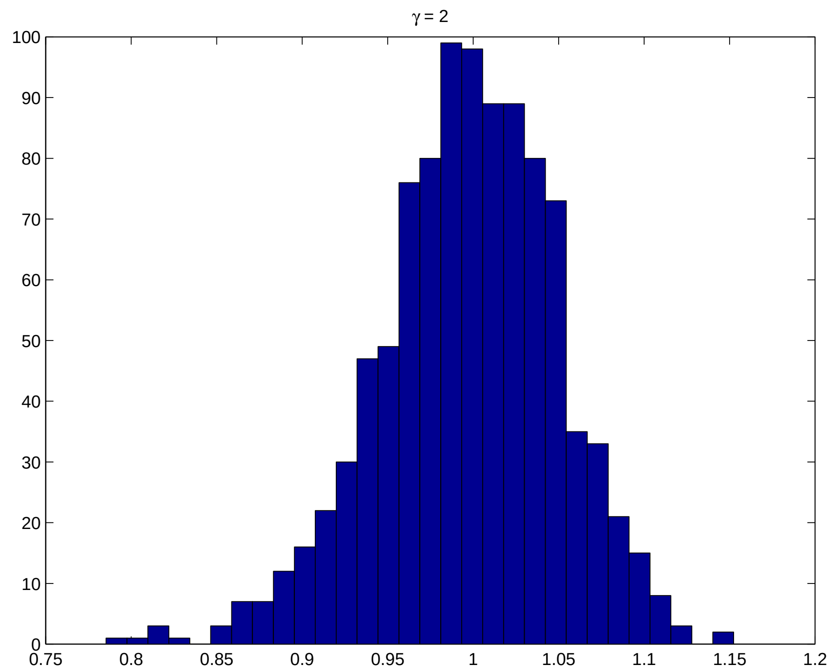


Figure 2: Question 3 (b)

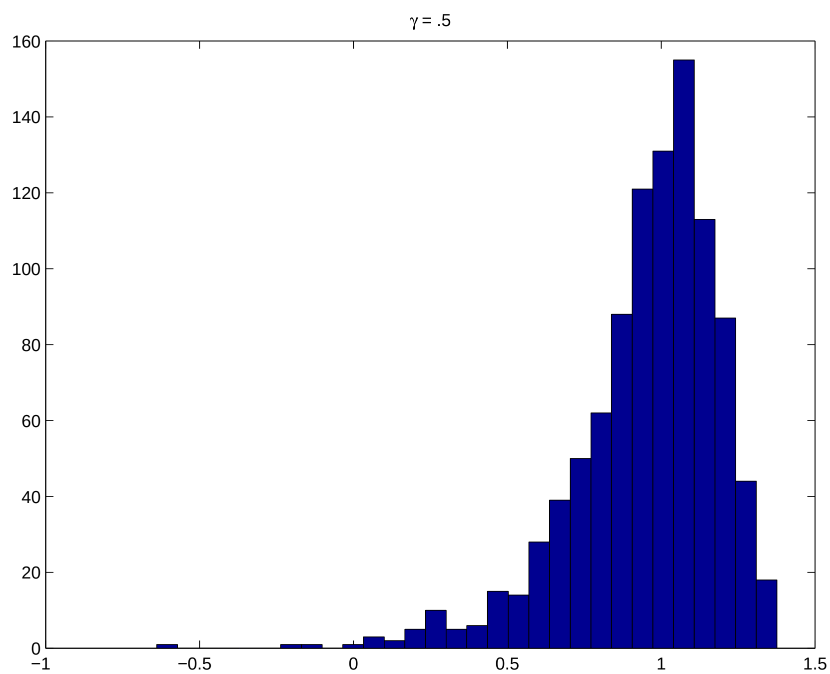
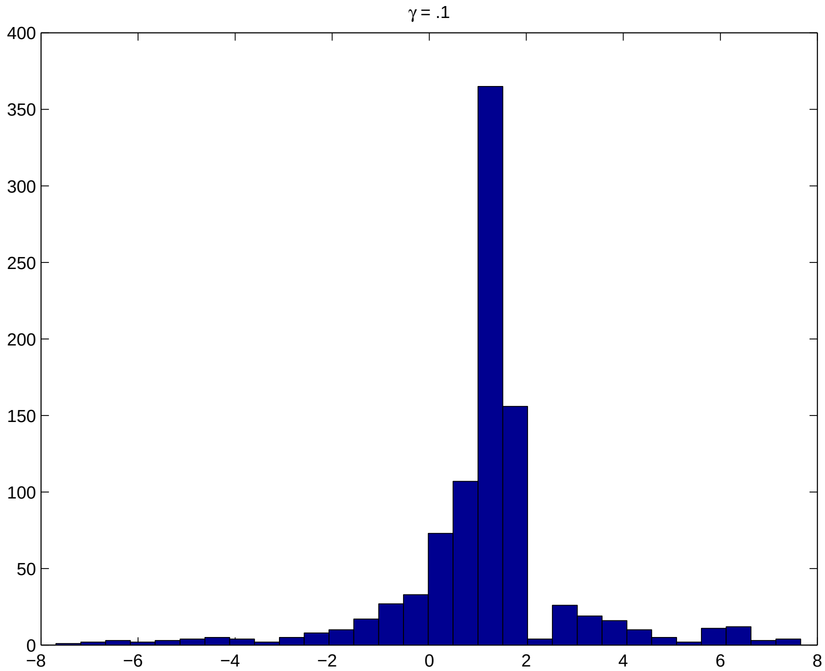


Figure 3: Question 3 (c)



4. The M-file `hw10_4.m` solves this question. It calls `tsls.m` to calculate the 2SLS estimator and associated standard errors.

   (a) The mean for the patients who received the flu shot was 0.0844, while the mean for the patients who did not receive the flu shot was 0.0848. A slightly smaller fraction of patients who got the flu shot were hospitalized.

   (b) The following results give the LS coefficients and two sets of standard errors: the "usual" standard errors and heteroskedasticity-robust standard errors.

	b	se	robust_se
constant	0.1285	0.0271	0.0278
flushot	0.0010	0.0120	0.0119
age	-0.0007	0.0004	0.0004

   Holding age constant, getting the flu shot appears to be associated with a .1% greater probability of being hospitalized. However, this is estimated very imprecisely. Also, holding flushot constant, an extra year of age is associated with a .07% lower probability of being hospitalized. This is also not estimated very precisely.

   (c) From the previous results, we might be tempted to conclude that the flu shot has basically no effect on hospitalization. However, this conclusion would implicitly assume that the association measured by the coefficient on flushot corresponded to the *effect* of the flu shot. There are a number of reasons why this might not hold. For example, it might be that the kind of people who get flu shots are more likely to get sick to begin with, even after controlling for age. If that were the case, even if the flu shot were effective, it might look as if the flu shot led to higher hospitalizations.

(d) The first-stage F statistic was 34.6, suggesting a fairly strong relationship between the instrumental variable and the treatment. The 2SLS results appear below:

	b_iv	se	se(heteroskedasticity)
constant	0.1386	0.0286	0.0289
flushot	-0.1151	0.0906	0.0908
age	-0.0004	0.0005	0.0005

Now it appears that the flu shot is associated with a relatively large effect on hospitalization, with the expected negative sign (ie getting the flu shot appears to lead to less chance of being hospitalized). However, the standard errors are too large to be able to reject the hypothesis of no effect at conventional significance levels. (This is a common problem in IV studies - large standard errors.) Still, it is interesting to see how much of a difference using an instrumental variable can make in the point estimates.

(e) It is conceivable that the exclusion restriction might be violated. Suppose for example that individuals who receive the letter, become motivated to take other, unobserved actions to improve their health (for example, getting more rest or changing their diet) and reduce the probability of hospitalization. If this were true, then the letter should appear in the main equation determining hospitalization.