Problem set 2, Part 1: Generalized Roy Model 1 Factor

Remark: Some of the results you will obtain will probably discourage you as to the generality of the current method. Don’t be. As we will see later, the same sensitivity to the exogeneity assumption is shared by almost every other method. As you go through the problem sets and the class pay attention to the economic assumptions made by each method. No method is free of assumptions or free of problems, you just need to pick the one you think makes most sense. Some problems can only be tackled with certain methods. Also, more general versions of the methods that will be presented exist or can be developed (thesis anyone?). You are supposed to be learning about the existing issues and about computation and estimation.

Take a generalized Roy model of the form:

\[
I = Z\gamma + V
\]

\[
Y_1 = X_1\beta_1 + \varepsilon_1
\]

\[
Y_0 = X_0\beta_0 + \varepsilon_0
\]

\[
D = 1 (I > 0)
\]

\[
Y = DY_1 + (1 - D)Y_0
\]

Assume that

\[
V = f_1\alpha_1 + U_V
\]

\[
\varepsilon_1 = f_1\alpha_1 + U_1
\]

\[
\varepsilon_0 = f_1\alpha_0 + U_0
\]

and that

\[
(U_1, U_0, U_V) \text{ mutually independent (} U_1 \perp U_0 \perp U_V) \]

\[
f_1 \perp (U_0, U_1, U_V)
\]

\[
U_1 \sim N (0, \sigma_{U_1}^2)
\]

\[
U_0 \sim N (0, \sigma_{U_0}^2)
\]

\[
U_V \sim N (0, \sigma_{U_V}^2)
\]

Suppose that additionally, you have an external test equation which we observe regardless of \( D \). (on part b of the problem set we will see how panel
data can be used instead of the test equation to identify the model). The test takes the form

\[ T_1 = Q\theta_1 + f_1\delta_{11} + U_{T_1} \]

and

\[ U_{T_1} \sim N\left(0, \sigma^2_{T_1}\right). \]

1. Write down the likelihood function for this problem assuming that \( f_1 \) has some distribution say \( \Pr(f_1) \).

Notice that conditional on \( f_1 \) everything is independent, so take advantage of this when writing the likelihood.

2. First assume that \( f_1 \sim N\left(0, \sigma^2_{f_1}\right) \) and that \( \sigma^2_{U_V} = 1 \) and \( \delta_{11} = 1 \).

Why do we need to assume that \( \delta_{11} = 1 \)? Could we change the assumption to \( \sigma^2_{f_1} = 1 \) and \( \delta_{11} > 0 \). Why do we need to assume \( \sigma^2_{U_V} = 1 \)? Could we have identified all the pieces of the model without the test equation? What if we had two time periods of \( Y \)?

3. Program a Maximum Likelihood version of this model. If you choose to take advantage of the independence conditional on the factors use your favorite integration method (Gaussian quadrature, Simpson’s rule or Monte Carlo).

4. Program an MCMC version of this problem. Put a non-informative prior on \( \gamma, \theta_1, \beta_1 \) and \( \beta_0 \). Put normal \( (0, 10.0) \) (proper but with little information) priors on \( \alpha_{01}, \alpha_{11} \) and \( \alpha_{11} \); and gamma \( (2, 1) \) priors on \( \left(\frac{1}{\sigma^2_{U_1}}, \frac{1}{\sigma^2_{U_0}}, \frac{1}{\sigma^2_{T_1}}\right) \).

We are going to use dataset 2a for this part of the problem set. The way this dataset was generated is the following:

\[ f_1 \sim N\left(0, 1\right) \]
\[ U_1 \sim N\left(0, 1\right) \]
\[ U_0 \sim N\left(0, 1\right) \]
\[ U_V \sim N\left(0, 1\right) \]
\( Z_0 = X_0 \) are just a constant equal to 1. Next we generate

\[
\begin{align*}
X_1 &\sim N(0,2) \\
Z_1 &\sim N(0,2)
\end{align*}
\]

so we are in the case where \( Z = (Z_0, Z_1) \) and \( X = (X_0, X_1) \) are exogenous. We finally form

\[
\begin{align*}
Y_1 &= 2X_0 + X_1 + 2f_1 + U_1 \\
Y_0 &= X_0 + X_1 + f_1 + U_0 \\
I &= 0.5Z_0 + Z_1 + f_1 + U_V
\end{align*}
\]

and let

\[
D = 1(I > 0).
\]

so that the observed \( Y \) is

\[
Y = DY_1 + (1 - D)Y_0
\]

Finally the test equation was generated as

\[
U_{T_1} \sim N(0,1)
\]

\( Q_0 = 1, \)

\( Q_1 \sim N(0,1) \)

and

\[
T_1 = Q_0 + Q_1 + f_1 + U_{T_1}.
\]

5. Derive the results don’t simulate them. Suppose you have the following regression

\[
Y = b_0X_0 + b_1X_1 + b_DD + \varepsilon
\]

and forget about \( f_1 \) (i.e. assume it does not enter any equation, so there is no selection on unobservables). What should \( b_D \) be equal to if there is no selection on unobservables? (hint: it should equal 1). Notice that \( b_D \) measures the effect of going from \( D = 0 \) to \( D = 1 \).

Now run the regression on the data you have, why are your results different from one? How is this related to the problem of ability bias or to the problem of omitted variable bias?

6. Run your Maximum likelihood model.

7. Run your MCMC model.
If you did it correctly you MLE estimates and the mean of your MCMC estimates should be close to the values we assigned when we built the dataset.

8. So now you have results that estimate all the parameters correctly. Could you define a unique parameter that takes the place of \( b_D \) (i.e., that measures the effect of going from \( D = 0 \) to \( D = 1 \)) now that you can estimate the parameters of the model without biases?

Now, let’s make \( Z \) endogenous.

\[
\varepsilon_Z \sim N(0, 1)
\]

\[
Z^*_1 = f_1 + \varepsilon_Z
\]

so that \( Z^*_1 \) is also distributed normal \((0, 2)\) like \( Z_1 \) but now it is endogenous. Then, define

\[
I^* = 0.5Z_0 + Z^*_1 + f_1 + U_V
\]

\[
D^* = 1(I^* > 0).
\]

and

\[
Y^* = D^*Y_1 + (1 - D^*)Y_0.
\]

9. Suppose you now run

\[
Y^* = b_0X_0 + b_1X_1 + b_DD^* + \varepsilon
\]

your results should be significantly different again from both you results in 5 and from the results with no selection. Why?

10. Run either your MLE or your MCMC model on this data (i.e. use \( Z^*_1, D^* \) and \( Y^* \) instead of \( Z_1, D \) and \( Y \)). What happens with your results? Can you explain why?

Now, let’s make \( X \) endogenous instead of \( Z \).

\[
\varepsilon_X \sim N(0, 1)
\]

\[
\hat{X}_1 = f_1 + \varepsilon_X
\]

so that \( \hat{X}_1 \) is also distributed normal \((0, 2)\) like \( X_1 \) but now it is endogenous. Then, define

\[
\hat{Y}_1 = 2X_0 + \hat{X}_1 + 2f_1 + U_1
\]

\[
\hat{Y}_0 = X_0 + \hat{X}_1 + f_1 + U_0
\]

and

\[
\hat{Y} = D\hat{Y}_1 + (1 - D)\hat{Y}_0
\]
11. Again, run the regression

\[ \hat{Y} = b_0X_0 + b_1\hat{X}_1 + b_D D + \varepsilon \]

again your results are different from all before. Why?

12. Run either your MLE or your MCMC model on this data (i.e. use \( \hat{X}_1 \) and \( \hat{Y} \) instead of \( X_1 \) and \( Y \) but go back to using \( D \) and \( Z_1 \)). What happens with your results? Can you explain why?