# Control Function Approach

## Contents

**Basic Idea** **1**

**Example: Self-Selection and Heckman Two-Step** **1**

**Binary Choice with Endogenous variables** **2**

## Basic Idea

Consider The following set-up

$$y_1 = \alpha y_2 + u, \tag{1}$$

$$y_2 = z'\pi + v, \tag{2}$$

In which $cov(u, v) \neq 0, E(z'v) = 0, E(z'u) = 0$ Let's write down the the conditional $y_1$ as a function of $y_2$

$$E(y_1|y_2) = E(y_1|z'\pi + v = y_2) = E(\alpha y_2 + u|v = y_2 - z'\pi) = \alpha y_2 + E(u|v = y_2 - z'\pi)$$

The problem here is that, $E(u|v = y_2 - z'\pi)$ is a function of $y_2$ and $z'$. If we simply regress $y_1$ on $y_2$, then the estimation is not consistent. This logic is super simple.

If we further assume

$$u = \rho v + e, E(v'e) = 0$$

Then

$$E(y_1|y_2) = \alpha y_2 + \rho(y_2 - z'\pi)$$

Now we have got the real expression for the $E(y_1|y_2)$. This also implies that a good regression expression is

$$y_1 = \alpha y_2 + \rho(y_2 - z'\pi) + e$$

To get the consistent estimation of $\alpha$, we can do the following step:

- estimate $\hat{v}$ from (2)
- regress $y_1$ on $y_2$ and $\hat{v}$

This approach is called control function approach. The basic idea here, from another perspective, is that: since $y_2$ is endogenous due to $v$, then we put the $v$ directly into (1), i.e., 'control' for it. In this linear case, the estimation of the control function approach and the 2SLS are actually the same. But it is said such approach is useful under non-linear model.

## Example: Self-Selection and Heckman Two-Step

Consider The following set-up

$$y = \alpha D + u, \tag{1}$$

$$D = 1\{z'\pi + v > 0\}, \tag{2}$$

In which \$ E(z'v)=0\$, $E(z'u)=0$. The $u$ and $v$ follows joint normal distribution, with mean as $(0,0)$, and $var(u)=\sigma_u^2$,$var(v)=1$, and $cov(u,v)=\rho\sigma_u$. We are all familiar with this framework! Regard $D$ as the participation to a treatment,which is in turn determined by $z$ and unobservable factor $v$.

We still want to write down the real $E(y|D)$.when $D=1$

$$E(y|D)=E(y|z'\pi+v>0)=E(\alpha*1+u|v>-z'\pi)=\alpha+E(u|v>-z'\pi)$$

when $D=0$ then

$$E(y|D)=E(y|z'\pi+v<0)=E(\alpha*0+u|v<-z'\pi)=E(u|v<-z'\pi)$$

We know that

$$E(u|v>-z'\pi)=\frac{\int_{-z'\pi}^{\infty}E(u|v)f_v(v)dv}{pr(v>-z'\pi)}=\frac{\rho\sigma_u\int_{-z'\pi}^{\infty}vf_v(v)dv}{pr(v>-z'\pi)}=\frac{\rho\sigma_u\phi(-z'\pi)}{1-\Phi(-z'\pi)}=\rho\sigma_u\lambda(-z'\pi)$$

$$E(u|v<-z'\pi)=\frac{-\rho\sigma_u\phi(-z'\pi)}{\Phi(-z'\pi)}=\frac{-\rho\sigma_u\phi(z'\pi)}{1-\Phi(z'\pi)}=-\rho\sigma_u\lambda(z'\pi)$$

Here we used the fact that $E(u|v)=E(u)+\frac{cov(u,v)}{var(v)}(v-E(v))=\frac{\rho\sigma_u}{1}$. The true function of $E(y|D)$ therefore must be

$$\begin{aligned}&[\alpha+\rho\sigma_u\lambda(-z'\pi)]\,D+[-\rho\sigma_u\lambda(z'\pi)]\,(1-D)\\&=\alpha D+[\rho\sigma_u\lambda(-z'\pi)]\,D+[-\rho\sigma_u\lambda(z'\pi)]\,(1-D)\end{aligned} \tag{3}$$

Suppose $\rho\neq 0$. I|f we regress $y$ on $D$ directly, we would not be able to get a consistent estimation of $\alpha$.What should we do? We can again use control function approach(Which is exactly heckman two-step!).

- First, estimate $\lambda(-\hat{z'}\pi)$ use probit model.
- Second, plug our estimation into (3) and estimate (3)

This control function approach is especially useful in dealing with the probit model with endogenous variables.

## Binary Choice with Endogenous variables

If we are familiar with the textbook the specification of the endogeneity problem, we can easily write down the following framework

$$y_1^*=y_2'\beta+u,\quad y_2=z'\gamma+v$$

$y_1=1$ if $y_1^*>0$, and $y_1=0$ if $y_1^*<0$.Assume that $u$ and $v$ are correlated unobservable terms (joint normal distribution with $\sigma_u$,$\sigma_v$,$\rho$), and $\sigma_u=1$. Also assume that $cov(z,u)=0$,$cov(z,v)=0$. Of course,if $\rho\neq 0$, we can see here that $y_2'$ is an endogenous variable, since $cov(y_2,u)=cov(z'\gamma+v,u)=cov(v,u)\neq 0$.

As in the basic model, we have

$$E(y_1^*|y_2)=\beta y_2+E(u|v=y_2-z'\gamma)$$

As before, we want to see the function form of $E(u|v>-z'\gamma)$. Statistic knowledge tells us that since $u$ and $v$ follows a joint normal distribution, we can write it down the relationship of $u$ and $v$ as

$$u=\delta v+\epsilon$$

In which $cov(v,\epsilon)=0$. Of course, since $u$ and $v$ are normal distribution with mean 0, $\epsilon$ must also follow a normal distribution with mean 0. Our task here is to find out the expression for $\delta$ and $var(\epsilon)$.

Notice that

$$cov(u,v)=\delta var(v)$$

. Therefore we have $\delta=\rho/\sigma$. And since

$$1=var(u)=\delta^2 var(v)+var(\epsilon)=\rho^2+var(\epsilon)$$

We have $var(\epsilon) = 1 - \rho^2$.

Now we know that the correction expression of $E(y_1^*|y_2)$ is

$$E(y_1^*|y_2) = \beta y_2 + \delta(y_2 - z'\gamma)$$

And

$$y_1^* = \beta y_2 + \delta(y_2 - z'\gamma) + e \tag{4}$$

But! Recall in probit model we need to make the error term be standard normal distribution. Here, the $e$, has variance $1 - \rho^2$. Therefore, before we apply the MLE for the probit model, we need to divide everything in (4) by $1 - \rho^2$:

$$\frac{y_1^*}{1-\rho^2} = y_2 \frac{\beta}{1-\rho^2} + \frac{\delta}{1-\rho^2}(y_2 - z'\gamma) + \frac{e}{1-\rho^2}$$

Therefore, control function approach here is

- estimate $\hat{v}$ by regress $y_2$ on $z$
- estimate The probit model using $y_1$,$y_2$ and $\hat{v}$.The log likelihood function for individual $i$ is

$$y_1 \log \Phi\left(-y_2 \frac{\beta}{1-\rho^2} - \frac{\delta}{1-\rho^2}\hat{v}\right) + (1 - y_1) log\left(1 - \Phi\left(-y_2 \frac{\beta}{1-\rho^2} - \frac{\delta}{1-\rho^2}\hat{v}\right)\right)$$

  .Notice that what we can identify is $\frac{\beta}{1-\rho^2}, \frac{\delta}{1-\rho^2}$

Reference:https://www.irp.wisc.edu/newsevents/workshops/appliedmicroeconometrics/participants/slides/Slides_14.pdf