Notes for Paper on Mis-measured, Binary, Endogenous Regressors

Francis J. DiTraglia & Camilo García-Jimeno

April 28, 2016

1 Model and Notation

Probabilities

$$\begin{aligned} p^*_{tk} &= P(T^* = t, Z = k) \\ p_{tk} &= P(T = t, Z = k) \\ p^*_k &= P(T^* = 1 | Z = k) \\ p_k &= P(T = 1 | Z = k) \\ q &= P(Z = 1) \end{aligned}$$

$$\begin{aligned} p^*_{00} &= P(T^* = 0|Z = 0)P(Z = 0) = (1 - p^*_0)(1 - q) = \left(\frac{1 - p_0 - \alpha_1}{1 - \alpha_0 - \alpha_1}\right)(1 - q) \\ p^*_{10} &= P(T^* = 1|Z = 0)P(Z = 0) = p^*_0(1 - q) = \left(\frac{p_0 - \alpha_0}{1 - \alpha_0 - \alpha_1}\right)(1 - q) \\ p^*_{01} &= P(T^* = 0|Z = 1)P(Z = 1) = (1 - p^*_1)q = \left(\frac{1 - p_1 - \alpha_1}{1 - \alpha_0 - \alpha_1}\right)q \\ p^*_{11} &= P(T^* = 1|Z = 1)P(Z = 1) = p^*_1(1 - q) = \left(\frac{p_1 - \alpha_0}{1 - \alpha_0 - \alpha_1}\right)q \end{aligned}$$

CDFs For $t, Z \in \{0, 1\}$ define

$$\begin{aligned} F^*_{tk}(\tau) &= P(Y \leq \tau | T^* = t, Z = k) \\ F_{tk}(\tau) &= P(Y \leq \tau | T = t, Z = k) \\ F_k(\tau) &= P(Y \leq \tau | Z = k) \end{aligned}$$

Note that the second two are observed for all t, k while the first is never observed since it depends on the unobserved RV T^*.

2 Weakest Bounds on α_0, α_1

Assume that $\alpha_0 + \alpha_1 < 1$ that T is independent of Z conditional on T^*. These standard assumptions turn out to yield informative bounds on α_0 and α_1 without *any further restrictions of any kind*. In particular, we assume nothing about the validity of the instrument Z and nothing about the relationship between the mis-classification error and the outcome Y: we impose only that the mis-classification error rates do not depend on z and that the mis-classification is not so bad that $1 - T$ is a better measure of T^* than T.

By the Law of Total Probability and the assumption that T is conditionally independent of Z given T^*,

$$\begin{aligned} p_k &= P(T = 1 | Z = k, T^* = 0)(1 - p^*_k) + P(T = 1 | Z = k, T^* = 1) p^*_k \\ &= P(T = 1 | T^* = 0)(1 - p^*_k) + P(T = 1 | T^* = 1) p^*_k \\ &= \alpha_0 (1 - p^*_k) + (1 - \alpha_1) p^*_k \\ &= \alpha_0 + (1 - \alpha_0 - \alpha_1) p^*_k \end{aligned}$$

and similarly

$$\begin{aligned} 1 - p_k &= P(T = 0 | Z = k, T^* = 0)(1 - p^*_k) + P(T = 0 | Z = k, T^* = 1) p^*_k \\ &= P(T = 0 | T^* = 0)(1 - p^*_k) + P(T = 0 | T^* = 1) p^*_k \\ &= (1 - \alpha_0)(1 - p^*_k) + \alpha_1 p^*_k \\ &= \alpha_1 + (1 - p^*_k)(1 - \alpha_0 - \alpha_1) \end{aligned}$$

and hence

$$\begin{aligned} p_k - \alpha_0 &= (1 - \alpha_0 - \alpha_1)p_k^* \\ (1 - p_k) - \alpha_1 &= (1 - \alpha_0 - \alpha_1)(1 - p_k^*) \end{aligned}$$

Now, since p_k^* and $(1-p_k^*)$ are probabilities they are between zero and one which means that the sign of $p_k - \alpha_0$ as well as that of $(1 - p_k) - \alpha_1$ are both determined by that of $1 - \alpha_0 - \alpha_1$. Accordingly, provided that $1 - \alpha_0 - \alpha_1 < 1$, we have

$$\begin{aligned} \alpha_0 &< p_k \\ \alpha_1 &< (1 - p_k) \end{aligned}$$

so long as p_k^* does not equal zero or one, which is not a realistic case for any example that we consider. Since these bounds hold for all k, we can take the tightest bound over all values of Z.

3 Stronger Bounds for α_0, α_1

Now suppose we add the assumption that T is conditionally independent of Y given T^*. This is essentially the non-differential measurement error assumption although it is slightly stronger than the version used by Mahajan (2006) who assumes only conditional mean independence. This assumption allows us to considerably strengthen the bounds from the preceding section by exploiting information contained in the conditional distribution of Y given T and Z. The key ingredient is a relationship that we can derive between the unobservable distributions F_{tk}^* and the observable distributions F_{tk} using this new conditional independence assumption. To begin, note that by Bayes' rule we have

$$\begin{aligned} P(T^* = 1|T = 1, Z = k) &= P(T = 1|T^* = 1)\left(\frac{p_k^*}{p_k}\right) = (1 - \alpha_1)\left(\frac{p_k^*}{p_k}\right) \\ P(T^* = 1|T = 0, Z = k) &= P(T = 0|T^* = 1)\left(\frac{p_k^*}{1 - p_k}\right) = \alpha_1\left(\frac{p_k^*}{1 - p_k}\right) \\ P(T^* = 0|T = 1, Z = k) &= P(T = 1|T^* = 0)\left(\frac{1 - p_k^*}{p_k}\right) = \alpha_0\left(\frac{1 - p_k^*}{p_k}\right) \\ P(T^* = 0|T = 0, Z = k) &= P(T = 0|T^* = 0)\left(\frac{1 - p_k^*}{1 - p_k}\right) = (1 - \alpha_0)\left(\frac{1 - p_k^*}{1 - p_k}\right) \end{aligned}$$

Now, by the conditional independence assumption

$$P(Y \leq \tau | T^* = 0, T = t, Z = k) = P(Y \leq \tau | T^* = 0, Z = k) = F^*_{0k}(\tau)$$
$$P(Y \leq \tau | T^* = 1, T = t, Z = k) = P(Y \leq \tau | T^* = 1, Z = k) = F^*_{1k}(\tau)$$

Finally, putting everything together using the Law of Total Probability, we find that

$$(1 - p_k)F_{0k}(\tau) = (1 - \alpha_0)(1 - p^*_k)F^*_{0k}(\tau) + \alpha_1 p^*_k F^*_{1k}(\tau)$$
$$p_k F_{1k}(\tau) = \alpha_0(1 - p^*_k)F^*_{0k}(\tau) + (1 - \alpha_1)p^*_k F^*_{1k}(\tau)$$

for all k. Defining the shorthand

$$\begin{aligned} \widetilde{F}_{0k}(\tau) &\equiv (1 - p_k)F_{0k}(\tau) \\ \widetilde{F}_{1k}(\tau) &\equiv p_k F_{1k}(\tau) \end{aligned}$$

this becomes

$$\widetilde{F}_{0k}(\tau) = (1 - \alpha_0)(1 - p^*_k)F^*_{0k}(\tau) + \alpha_1 p^*_k F^*_{1k}(\tau) \tag{3.1}$$
$$\widetilde{F}_{1k}(\tau) = \alpha_0(1 - p^*_k)F^*_{0k}(\tau) + (1 - \alpha_1)p^*_k F^*_{1k}(\tau) \tag{3.2}$$

Now, solving Equation 3.1 for $p^*_k F^*_{1k}(\tau)$ we have

$$p^*_k F^*_{1k}(\tau) = \frac{1}{\alpha_1}\left[\widetilde{F}_{0k}(\tau) - (1 - \alpha_0)(1 - p^*_k)F^*_{0k}(\tau)\right]$$

Substituting this into Equation 3.2,

$$\begin{aligned} \widetilde{F}_{1k}(\tau) &= \alpha_0(1 - p^*_k)F^*_{0k}(\tau) + \frac{1 - \alpha_1}{\alpha_1}\left[\widetilde{F}_{0k}(\tau) - (1 - \alpha_0)(1 - p^*_k)F^*_{0k}(\tau)\right] \\ &= \frac{1 - \alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) + \left[\alpha_0 - \frac{(1 - \alpha_1)(1 - \alpha_0)}{\alpha_1}\right](1 - p^*_k)F^*_{0k}(\tau) \\ &= \frac{1 - \alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) + \left[\frac{\alpha_0\alpha_1 - (1 - \alpha_1)(1 - \alpha_0)}{\alpha_1}\right](1 - p^*_k)F^*_{0k}(\tau) \\ &= \frac{1 - \alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) - \left[\frac{(1 - \alpha_1)(1 - \alpha_0) - \alpha_0\alpha_1}{\alpha_1}\right](1 - p^*_k)F^*_{0k}(\tau) \\ &= \frac{1 - \alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) - \left[\frac{1 - \alpha_1 - \alpha_0}{\alpha_1}\right]\left(\frac{1 - p_k - \alpha_1}{1 - \alpha_0 - \alpha_1}\right)F^*_{0k}(\tau) \end{aligned}$$

and therefore

$$\widetilde{F}_{1k}(\tau) = \frac{1-\alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) - \frac{1-p_k-\alpha_1}{\alpha_1}F^*_{0k}(\tau) \tag{3.3}$$

Equation 3.3 relates the observable $\widetilde{F}_{1k}(\tau)$ to the mis-classification error rate α_1 and the unobservable CDF $F^*_{0k}(\tau)$. Since $F^*_{0k}(\tau)$ is a CDF, however, it lies in the interval $[0,1]$. Accordingly, substituting 0 in place of $F^*_{0k}(\tau)$ gives

$$\widetilde{F}_{1k}(\tau) \leq \frac{1-\alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) \tag{3.4}$$

while substituting 1 gives

$$\widetilde{F}_{1k}(\tau) \geq \frac{1-\alpha_1}{\alpha_1}\widetilde{F}_{0k}(\tau) - \frac{1-p_k-\alpha_1}{\alpha_1} \tag{3.5}$$

Rearranging Equation 3.4

$$\begin{aligned}
\alpha_1\widetilde{F}_{1k}(\tau) &\leq (1-\alpha_1)\widetilde{F}_{0k}(\tau) \\
\alpha_1\widetilde{F}_{1k}(\tau) &\leq \widetilde{F}_{0k}(\tau) - \alpha_1\widetilde{F}_{0k}(\tau) \\
\alpha_1\left[\widetilde{F}_{1k}(\tau) + \widetilde{F}_{1k}(\tau)\right] &\leq \widetilde{F}_{0k}(\tau)
\end{aligned}$$

since $\alpha_1 \in [0,1]$ and therefore

$$\alpha_1 \leq \frac{\widetilde{F}_{0k}(\tau)}{\widetilde{F}_{1k}(\tau) + \widetilde{F}_{1k}(\tau)} \tag{3.6}$$

since $\widetilde{F}_{1k}(\tau) + \widetilde{F}_{1k}(\tau) \geq 0$. Proceeding similarly for Equation 3.5,

$$\begin{aligned}
\alpha_1\widetilde{F}_{1k}(\tau) &\geq (1-\alpha_1)\widetilde{F}_{0k}(\tau) - (1-p_k-\alpha_1) \\
\alpha_1\left[\widetilde{F}_{1k}(\tau) + \widetilde{F}_{0k}(\tau) - 1\right] &\geq \widetilde{F}_{0k}(\tau) - (1-p_k) \\
-\alpha_1\left[1 - \widetilde{F}_{1k}(\tau) - \widetilde{F}_{0k}(\tau)\right] &\geq -\left[1 - \widetilde{F}_{0k}(\tau) - p_k\right] \\
\alpha_1\left[1 - \widetilde{F}_{1k}(\tau) - \widetilde{F}_{0k}(\tau)\right] &\leq 1 - \widetilde{F}_{0k}(\tau) - p_k
\end{aligned}$$

Now since $\widetilde{F}_{1k}(\tau) = p_k F_{1k}(\tau) \leq p_k$ and $\widetilde{F}_{0k}(\tau) = (1-p_k)F_{0k}(\tau) \leq (1-p_k)$ it follows that $1 - \widetilde{F}_{1k}(\tau) - \widetilde{F}_{0k}(\tau) \geq 0$ and hence

$$\alpha_1 \leq \frac{1 - \widetilde{F}_{0k}(\tau) - p_k}{1 - \widetilde{F}_{1k}(\tau) - \widetilde{F}_{0k}(\tau)} \tag{3.7}$$

The bounds given in Equations 3.6 and 3.7 relate α_1 to observable quantities *only* and hold for all values of τ for which their respective denominators are non-zero. Moreover, these bounds hold for any value k that the instrument takes on.

We can proceed similarly for α_0. First solve Equation 3.1 for $(1-p_k^*)F_{0k}^*(\tau)$:

$$(1-p_k^*)F_{0k}^*(\tau) = \frac{1}{1-\alpha_0}\left[\widetilde{F}_{0k}(\tau) - \alpha_1 p_k^* F_{1k}^*(\tau)\right]$$

and then substitute into Equation 3.2:

$$\begin{aligned}
\widetilde{F}_{1k}(\tau) &= \frac{\alpha_0}{1-\alpha_0}\left[\widetilde{F}_{0k}(\tau) - \alpha_1 p_k^* F_{1k}^*(\tau)\right] + (1-\alpha_1)p_k^* F_{1k}^*(\tau) \\
&= \frac{\alpha_0}{1-\alpha_0}\widetilde{F}_{0k}(\tau) + \left[(1-\alpha_1) - \frac{\alpha_0\alpha_1}{1-\alpha_0}\right] p_k^* F_{1k}^*(\tau) \\
&= \frac{\alpha_0}{1-\alpha_0}\widetilde{F}_{0k}(\tau) + \left[\frac{(1-\alpha_1)(1-\alpha_0) - \alpha_0\alpha_1}{1-\alpha_0}\right] p_k^* F_{1k}^*(\tau) \\
&= \frac{\alpha_0}{1-\alpha_0}\widetilde{F}_{0k}(\tau) + \left[\frac{1-\alpha_0-\alpha_1}{1-\alpha_0}\right]\frac{p_k-\alpha_0}{1-\alpha_0-\alpha_1} F_{1k}^*(\tau)
\end{aligned}$$

and therefore

$$\widetilde{F}_{1k}(\tau) = \frac{\alpha_0}{1-\alpha_0}\widetilde{F}_{0k}(\tau) + \frac{p_k-\alpha_0}{1-\alpha_0} F_{1k}^*(\tau) \tag{3.8}$$

Now we can again obtain two bounds by substituting the smallest and largest possible values of $F_{1k}^*(\tau)$. Substituting zero gives

$$\widetilde{F}_{1k}(\tau) \geq \frac{\alpha_0}{1-\alpha_0}\widetilde{F}_{0k}(\tau) \tag{3.9}$$

while substituting one gives

$$\widetilde{F}_{1k}(\tau) \leq \frac{\alpha_0}{1-\alpha_0}\widetilde{F}_{0k}(\tau) + \frac{p_k-\alpha_0}{1-\alpha_0} \tag{3.10}$$

Now, rearranging Equation 3.9,

$$\begin{aligned}
(1-\alpha_0)\widetilde{F}_{1k}(\tau) &\geq \alpha_0\widetilde{F}_{0k}(\tau) \\
\widetilde{F}_{1k}(\tau) &\geq \alpha_0\left[\widetilde{F}_{0k}(\tau) + \widetilde{F}_{1k}(\tau)\right]
\end{aligned}$$

since $1 - \alpha_0 \geq 0$. Therefore,

$$\alpha_0 \leq \frac{\widetilde{F}_{1k}(\tau)}{\widetilde{F}_{0k}(\tau) + \widetilde{F}_{1k}(\tau)} \tag{3.11}$$

since $\left[\widetilde{F}_{0k}(\tau) + \widetilde{F}_{1k}(\tau)\right] \geq 0$. Similarly, rearranging Equation 3.10

$$\begin{aligned}
(1-\alpha_0)\widetilde{F}_{1k}(\tau) &\leq \alpha_0\widetilde{F}_{0k}(\tau) + p_k - \alpha_0 \\
\widetilde{F}_{1k}(\tau) - p_k &\leq \alpha_0\left[\widetilde{F}_{0k}(\tau) + \widetilde{F}_{1k}(\tau) - 1\right] \\
-\left[1 - \widetilde{F}_{1k}(\tau) - (1-p_k)\right] &\leq -\alpha_0\left[1 - \widetilde{F}_{0k}(\tau) - \widetilde{F}_{1k}(\tau)\right] \\
\left[1 - \widetilde{F}_{1k}(\tau) - (1-p_k)\right] &\geq \alpha_0\left[1 - \widetilde{F}_{0k}(\tau) - \widetilde{F}_{1k}(\tau)\right]
\end{aligned}$$

Therefore

$$\alpha_0 \leq \frac{1 - \widetilde{F}_{1k}(\tau) - (1-p_k)}{1 - \widetilde{F}_{0k}(\tau) - \widetilde{F}_{1k}(\tau)} \tag{3.12}$$

4 Even Stronger Bounds on α_0, α_1

Try applying the stochastic dominance conditions from our simulation study.

5 Independent Instrument

Assume that $Z \perp U$. The model is $Y = \beta T^* + U$ and

$$F_U(\tau) = P(U \leq \tau) = P(Y - \beta T^* \leq \tau)$$

but if Z is independent of U then it follows that

$$\begin{aligned}
F_U(\tau) &= F_{U|Z=k}(\tau) = P(U \leq \tau | Z = k) = P(Y - \beta T^* \leq \tau | Z = k) \\
&= P(Y \leq \tau | T^* = 0, Z = k)(1 - p_k^*) + P(Y \leq \tau + \beta | T^* = 1, Z = k)p_k^* \\
&= (1 - p_k^*)F_{0k}^*(\tau) + p_k^* F_{1k}^*(\tau + \beta)
\end{aligned}$$

for all k by the Law of Total Probability. Similarly,

$$F_k(\tau) = (1 - p_k^*)F_{0k}^*(\tau) + p_k^* F_{1k}^*(\tau)$$

and rearranging

$$(1-p_k^*)F_{0k}^*(\tau) = F_k(\tau) - p_k^* F_{1k}^*(\tau)$$

Substituting this expression into the equation for $F_U(\tau)$ from above, we have

$$F_U(\tau) = F_k(\tau) + p_k^* \left[F_{1k}^*(\tau+\beta) - F_{1k}^*(\tau)\right]$$

for all k and all τ. Evaluating at two values k and ℓ in the support of Z and equating

$$F_k(\tau) + p_k^* \left[F_{1k}^*(\tau+\beta) - F_{1k}^*(\tau)\right] = F_\ell(\tau) + p_\ell^* \left[F_{1\ell}^*(\tau+\beta) - F_{1\ell}^*(\tau)\right]$$

or equivalently

$$F_k(\tau) - F_\ell(\tau) = p_\ell^* \left[F_{1\ell}^*(\tau+\beta) - F_{1\ell}^*(\tau)\right] - p_k^* \left[F_{1k}^*(\tau+\beta) - F_{1k}^*(\tau)\right] \tag{5.1}$$

for all τ. Now we simply need to re-express all of the "star" quantities, namely p_k^*, p_ℓ^* and $F_{1k}^*, F_{1\ell}^*$ in terms of α_0, α_1 and the *observable* probability distributions F_{1k} and $F_{1\ell}$ and observable probabilities p_k, p_ℓ. To do this, we use the fact that

$$\begin{aligned} F_{0k}(\tau) &= \frac{1-\alpha_0}{1-p_k}(1-p_k^*)F_{0k}^*(\tau) + \frac{\alpha_1}{1-p_k}p_k^* F_{1k}^*(\tau) \\ F_{1k}(\tau) &= \frac{\alpha_0}{p_k}(1-p_k^*)F_{0k}^*(\tau) + \frac{1-\alpha_1}{p_k}p_k^* F_{1k}^*(\tau) \end{aligned}$$

for all k by Bayes' rule. Solving these equations,

$$p_k^* F_{1k}^*(\tau) = \frac{1-\alpha_0}{1-\alpha_0-\alpha_1}p_k F_{1k}(\tau) - \frac{\alpha_0}{1-\alpha_0-\alpha_1}(1-p_k)F_{0k}(\tau)$$

for all k. Combining this with Equation 5.1, we find that

$$\begin{aligned} (1-\alpha_0-\alpha_1)\left[F_k(\tau) - F_\ell(\tau)\right] = \alpha_0 \{&(1-p_k)\left[F_{0k}(\tau+\beta) - F_{0k}(\tau)\right] - (1-p_\ell)\left[F_{0\ell}(\tau+\beta) - F_{0\ell}(\tau)\right]\} \\ &- (1-\alpha_0)\{p_k\left[F_{1k}(\tau+\beta) - F_{1k}(\tau)\right] - p_\ell\left[F_{1\ell}(\tau+\beta) - F_{1\ell}(\tau)\right]\} \end{aligned}$$

Now, define

$$\Delta_{tk}^\tau(\beta) = F_{tk}(\tau+\beta) - F_{tk}(\tau) = E\left[\frac{\mathbf{1}\{T=t, Z=k\}}{p_{tk}}\left(\mathbf{1}\{Y \le \tau+\beta\} - \mathbf{1}\{Y \le \tau\}\right)\right]$$

and note that we can express $F_k(\tau) - F_\ell(\tau)$ similarly as

$$F_k(\tau) - F_\ell(\tau) = E\left[\mathbf{1}\{Y \leq \tau\}\left(\frac{\mathbf{1}\{Z=k\}}{q_k} - \frac{\mathbf{1}\{Z=\ell\}}{q_\ell}\right)\right]$$

Using this notation, we can write the preceding as

$$(1-\alpha_0-\alpha_1)\left[F_k(\tau) - F_\ell(\tau)\right] = \alpha_0\left[(1-p_k)\Delta^\tau_{0k}(\beta) - (1-p_\ell)\Delta^\tau_{0\ell}(\beta)\right] - (1-\alpha_0)\left[p_k\Delta^\tau_{1k}(\beta) - p_\ell\Delta^\tau_{1\ell}(\beta)\right]$$

or in moment-condition form

$$E\Bigg[(1-\alpha_0-\alpha_1)\mathbf{1}\{Y\leq\tau\}\left(\frac{\mathbf{1}\{Z=k\}}{q_k} - \frac{\mathbf{1}\{Z=\ell\}}{q_\ell}\right) - (\mathbf{1}\{Y\leq\tau+\beta\} - \mathbf{1}\{Y\leq\tau\})\Bigg\{ \\ \alpha_0\left((1-p_k)\frac{\mathbf{1}\{T=0,Z=k\}}{p_{0k}} - (1-p_\ell)\frac{\mathbf{1}\{T=0,Z=\ell\}}{p_{0\ell}}\right) \\ -(1-\alpha_0)\left(p_k\frac{\mathbf{1}\{T=1,Z=k\}}{p_{1k}} - p_\ell\frac{\mathbf{1}\{T=1,Z=\ell\}}{p_{1\ell}}\right)\Bigg\}\Bigg] = 0$$

Each value of τ yields a moment condition.

6 Special Case: $\alpha_0 = 0$

In this case the expressions from above simplify to

$$(1-\alpha_1)\left[F_k(\tau) - F_\ell(\tau)\right] + \left\{p_k\left[F_{1k}(\tau+\beta) - F_{1k}(\tau)\right] - p_\ell\left[F_{1\ell}(\tau+\beta) - F_{1\ell}(\tau)\right]\right\} = 0$$

for all τ.