

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

Francis J. DiTraglia & Camilo García-Jimeno

April 26, 2016

1 Model and Notation

Fill in material from earlier notes so we have everything in one document!

$$p_{jk}^* = P(T^* = t, Z = k)$$

$$p_{jk} = P(T = t, Z = k)$$

$$p_k^* = P(T^* = 1|Z = k)$$

$$p_k = P(T = 1|Z = k)$$

$$q = P(Z = 1)$$

Thus,

$$\begin{aligned} p_{00}^* &= P(T^* = 0|Z = 0)P(Z = 0) = (1 - p_0^*)(1 - q) \\ &= \left(\frac{1 - p_0 - \alpha_1}{1 - \alpha_0 - \alpha_1} \right) (1 - q) \end{aligned}$$

$$\begin{aligned} p_{10}^* &= P(T^* = 1|Z = 0)P(Z = 0) = p_0^*(1 - q) \\ &= \left(\frac{p_0 - \alpha_0}{1 - \alpha_0 - \alpha_1} \right) (1 - q) \end{aligned}$$

$$\begin{aligned} p_{01}^* &= P(T^* = 0|Z = 1)P(Z = 1) = (1 - p_1^*)q \\ &= \left(\frac{1 - p_1 - \alpha_1}{1 - \alpha_0 - \alpha_1} \right) q \end{aligned}$$

$$\begin{aligned} p_{11}^* &= P(T^* = 1|Z = 1)P(Z = 1) = p_1^*(1 - q) \\ &= \left(\frac{p_1 - \alpha_0}{1 - \alpha_0 - \alpha_1} \right) q \end{aligned}$$

2 CDF Conditions

2.1 Notation

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^* .

2.2 Bounds on α_0, α_1

Case I: No Assumptions on Z We begin by considering the bounds that we can derive for the mis-classification error rates without imposing any conditions on Z . In other words we use only the assumption that the measurement error is non-differential and the structure of the model, namely $Y = \beta T^* + U$. The bounds we obtain in this

way will be applicable even if the instrument is invalid. To begin, note that we can express the observable CDFs F_{tk} in terms of the unobservable CDFs F_{tk}^* according to

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

for all k by Bayes' rule. Defining the shorthand

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

this becomes

$$\tilde{F}_{0k}(\tau) = (1 - \alpha_0)(1 - p_k^*)F_{0k}^*(\tau) + \alpha_1 p_k^* F_{1k}^*(\tau) \quad (2.1)$$

$$\tilde{F}_{1k}(\tau) = \alpha_0(1 - p_k^*)F_{0k}^*(\tau) + (1 - \alpha_1)p_k^* F_{1k}^*(\tau) \quad (2.2)$$

Now, solving Equation 2.1 for $p_k^* F_{1k}^*(\tau)$ we have

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

Substituting this into Equation 2.2,

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

and therefore

$$\tilde{F}_{1k}(\tau) = \frac{1 - \alpha_1}{\alpha_1} \tilde{F}_{0k}(\tau) - \frac{1 - p_k - \alpha_1}{\alpha_1} F_{0k}^*(\tau) \quad (2.3)$$

Equation 2.3 relates the observable $\tilde{F}_{1k}(\tau)$ 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

$$\tilde{F}_{1k}(\tau) \leq \frac{1 - \alpha_1}{\alpha_1} \tilde{F}_{0k}(\tau) \quad (2.4)$$

while substituting 1 gives

$$\tilde{F}_{1k}(\tau) \geq \frac{1 - \alpha_1}{\alpha_1} \tilde{F}_{0k}(\tau) - \frac{1 - p_k - \alpha_1}{\alpha_1} \quad (2.5)$$

Rearranging Equation 2.4

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

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

$$\alpha_1 \leq \frac{\tilde{F}_{0k}(\tau)}{\tilde{F}_{1k}(\tau) + \tilde{F}_{1k}(\tau)} \quad (2.6)$$

since $\tilde{F}_{1k}(\tau) + \tilde{F}_{1k}(\tau) \geq 0$. Proceeding similarly for Equation 2.5,

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

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

$$\alpha_1 \leq \frac{1 - \tilde{F}_{0k}(\tau) - p_k}{1 - \tilde{F}_{1k}(\tau) - \tilde{F}_{0k}(\tau)} \quad (2.7)$$

The bounds given in Equations 2.6 and 2.7 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 2.1 for $(1 - p_k^*)F_{0k}^*(\tau)$:

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

and then substitute into Equation 2.2:

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

and therefore

$$\tilde{F}_{1k}(\tau) = \frac{\alpha_0}{1 - \alpha_0} \tilde{F}_{0k}(\tau) + \frac{p_k - \alpha_0}{1 - \alpha_0} F_{1k}^*(\tau) \quad (2.8)$$

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

$$\tilde{F}_{1k}(\tau) \geq \frac{\alpha_0}{1 - \alpha_0} \tilde{F}_{0k}(\tau) \quad (2.9)$$

while substituting one gives

$$\tilde{F}_{1k}(\tau) \leq \frac{\alpha_0}{1 - \alpha_0} \tilde{F}_{0k}(\tau) + \frac{p_k - \alpha_0}{1 - \alpha_0} \quad (2.10)$$

2.3 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

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

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

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

or equivalently

$$F_k(\tau) - F_\ell(\tau) = p_\ell^*[F_{1\ell}^*(\tau + \beta) - F_{1\ell}^*(\tau)] - p_k^*[F_{1k}^*(\tau + \beta) - F_{1k}^*(\tau)] \quad (2.11)$$

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,

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

for all k . Combining this with Equation 2.11, we find that

$$(1 - \alpha_0 - \alpha_1) [F_k(\tau) - F_\ell(\tau)] = \alpha_0 \{ (1 - p_k) [F_{0k}(\tau + \beta) - F_{0k}(\tau)] - (1 - p_\ell) [F_{0\ell}(\tau + \beta) - F_{0\ell}(\tau)] \} \\ - (1 - \alpha_0) \{ p_k [F_{1k}(\tau + \beta) - F_{1k}(\tau)] - p_\ell [F_{1\ell}(\tau + \beta) - F_{1\ell}(\tau)] \}$$

Now, define

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

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

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

or in moment-condition form

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

Each value of τ yields a moment condition.

3 Special Case: $\alpha_0 = 0$

In this case the expressions from above simplify to

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

for all τ .