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

Francis J. DiTraglia & Camilo García-Jimeno

April 25, 2016

## 1 Model and Notation

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

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

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 $\alpha_0, \alpha_1$

**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

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

this becomes

$$\widetilde{F}_{0k}(\tau) = (1-\alpha_0)(1-p^*_k)F^*_{0k}(\tau) + \alpha_1 p^*_k F^*_{1k}(\tau) \tag{2.1}$$

$$\widetilde{F}_{1k}(\tau) = \alpha_0(1-p^*_k)F^*_{0k}(\tau) + (1-\alpha_1)p^*_k F^*_{1k}(\tau) \tag{2.2}$$

Now, solving Equation 2.1 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 2.2,

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

and therefore

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

Equation 2.3 relates the observable $\widetilde{F}_{1k}(\tau)$ to the mis-classification error rate $\alpha_1$ 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

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

while substituting 1 gives

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

Rearranging Equation 2.4

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

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

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

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

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

The bounds given by equations ??? and ??? relate the mis-classification error rate $\alpha_1$ to observable quantities *only* and hold for all values of $\tau$.

## 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

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

for all $k$ and all $\tau$. Evaluating at two values $k$ and $\ell$ 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

$$F_k(\tau) - F_\ell(\tau) = p_\ell^* \left[F_{1\ell}^*(\tau + \beta) - F_{1\ell}^*(\tau)\right] - p_k^* \left[F_{1k}^*(\tau + \beta) - F_{1k}^*(\tau)\right] \tag{2.7}$$

for all $\tau$. Now we simply need to re-express all of the "star" quantities, namely $p_k^*, p_\ell^*$ and $F_{1k}^*, F_{1\ell}^*$ in terms of $\alpha_0, \alpha_1$ and the *observable* probability distributions $F_{1k}$ and $F_{1\ell}$ and observable probabilities $p_k, p_\ell$. 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 2.7, we find that

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

Now, define

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

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

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

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

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

Each value of $\tau$ yields a moment condition.

# 3 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 $\tau$.