# MIT Graduate Labor Economics 14.662 Spring 2015 Instructor Notes

David H. Autor and Peter D. Hull
MIT

February 11, 2015

## 1 Mundlak decomposition[1]

Consider a series of cross sectional regressions of earnings on educational attainment $S_i$ with establishment fixed effects

$$y_{it} = c_t + r_t S_i + \mu_{j(i,t)t} + v_{it}.$$

The $\mu_{j(i,t)t}$ captures any general effect of working in establishing $j$ in time $t$ that is shared by all workers, regardless of education. Estimating instead a population regression that ignores the establishment fixed effect:

$$y_{it} = c'_t + r_t^M S_i + v'_{it},$$

will yield a population regression coefficient

$$r_t^M \equiv \frac{Cov\left(y_{it}, S_i\right)}{Var\left(S_i\right)} = r_t + \lambda_t b_t,$$

where

$$b_t \equiv \frac{Cov\left(\bar{S}_{j(i,t)t}, S_i\right)}{Var\left(S_i\right)},$$

is an educational sorting index, and

$$\lambda_t \equiv \frac{Cov\left(u_{j(i,t)t}, \bar{S}_{j(i,t)t}\right)}{Var\left(\bar{S}_{j(i,t)t}\right)},$$

which is the coefficient of a regression of the establishment effect on mean schooling at the establishment.

Thus, the standard Mincerian return to schooling in each year $r_t^M$ is composed of three components:

1. the return to schooling conditional on job quality $r_t$;
2. the return to mean establishment schooling $\lambda_t$;
3. and the association between worker and establishment schooling $b_t$.

Note that we can derive this decomposition from the (hopefully familiar) omitted-variables

[1] Due to Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica,* 46, 69–85.

bias formula (see, e.g., *Mostly Harmless Econometrics*, section 3.2.2):

$$\begin{aligned}
\underbrace{r_t^M}_{\text{"short"}} &= \underbrace{r_t^l}_{\text{"long"}} + \underbrace{1}_{\text{"effect of omitted"}} \cdot \underbrace{\frac{Cov(\mu_{j(i,t)t}, S_i)}{Var(S_i)}}_{\text{"regression of omitted on included"}} \\
&= r_t^l + \frac{Cov(\mu_{j(i,t)t}, \bar{S}_{j(i,t)t} + \tilde{S}_i)}{Var(\bar{S}_{j(i,t)t} + \tilde{S}_i)} \\
&= r_t^l + \frac{Cov(\mu_{j(i,t)t}, \bar{S}_{j(i,t)t} + \tilde{S}_i)}{Var(\bar{S}_{j(i,t)t})} \frac{Var(\bar{S}_{j(i,t)t})}{Var(\bar{S}_{j(i,t)t} + \tilde{S}_i)} \\
&= r_t^l + \frac{Cov(\mu_{j(i,t)t}, \bar{S}_{j(i,t)t})}{Var(\bar{S}_{j(i,t)t})} \frac{Cov(\bar{S}_{j(i,t)t}, S_i)}{Var(S_i)}
\end{aligned}$$

Here we've defined $\tilde{S}_i \equiv S_i - \bar{S}_{j(i,t)t}$ as the "within-establishment" variation in $S_i$ (i.e. residual from regressing $S_i$ on establishment effects). By construction $Cov(\mu_{j(i,t)t}, \tilde{S}_i) = 0$ and

$$\begin{aligned}
Cov(\bar{S}_{j(i,t)t}, S_i) &= Cov(\bar{S}_{j(i,t)t}, \bar{S}_{j(i,t)t} + \tilde{S}_i) \\
&= Var(\bar{S}_{j(i,t)t})
\end{aligned}$$

We can use the same OVB logic to note that

$$\begin{aligned}
r_t^M &= r_t + \lambda_t b_t \\
&= r_t + \lambda_t \frac{Cov\left(\bar{S}_{j(i,t)t}, S_i\right)}{Var\left(S_i\right)}
\end{aligned}$$

relates the "short" regression coefficient to the "long" regression of

$$y_{it} = c_t + r_t S_i + \lambda_t \bar{S}_{j(i,t)t} + \omega_{it}.$$

We can thus estimate $\lambda_t$ by running this regression. Estimating the short and long regression coefficients $r_t^M$ and $r_t^l$ would then allow us to solve for $b$.

Figure 13 from CHK (NBER Working Paper #18522, November 2012) shows that a large chunk of the apparent rise in the return to schooling is due to the rising 'return' to mean establishment schooling, which effectively triples from 0.04 to 0.11 between 1985 to 2009.

14.662 Labor Economics II
Spring 2015