Labor Supply: Wages and Hours Worked, Ziliak (1997)

Eyayaw Beze

July 05, 2020

Data Summary

Data on **id, year, lnhrs, lnwg, kids, age, and disab**

```
(N <- length(unique(labor_ss$id))) # cross-sections
```

```
#> [1] 532
```

```
(TT <- length(unique(labor_ss$year))) # length of Time
```

```
#> [1] 10
```

```
from <- min(labor_ss$year)
to <- max(labor_ss$year)
(NT <- N*TT) # total number of observations
```

```
#> [1] 5320
```

```
mainvars <- c("lnhrs", "lnwg")
K <- 1 # num of dep var; we lnwg as the only regressor
grand_mean <- my_round(sapply(labor_ss[mainvars], mean))
```

The dimension of the data set is N = 532 cross-sections by T = 10 (years) from 1979 to 1988. Thus, there are $N \times T = 5320$ observations. The sample means of lnhrs and lnwg are respectively, 7.66 and 2.61, implying geometric means of 2,122 hours and $13.6.

```
summ_stats <- sapply(labor_ss[-grep("^id|year", names(labor_ss))],
                     function(x, na.rm = FALSE) list(min = min(x),
                                                     mean = mean(x),
                                                     sd = sd(x),
                                                     max = max(x)), na.rm = TRUE)
dims <- dim(summ_stats)
dimnames <- dimnames(summ_stats)
summ_stats <- my_round(unlist(summ_stats))
dim(summ_stats) <- dims
dimnames(summ_stats) <- dimnames

summ_stats
```

```
#>      lnhrs  lnwg kids   age disab
#> min   2.77 -0.26 0.00 22.00  0.00
#> mean  7.66  2.61 1.56 38.92  0.06
#> sd    0.29  0.43 1.20  8.45  0.24
#> max   8.56  4.69 6.00 60.00  1.00
```

The sample standard deviations are respectively, 7.66 and 8.56 indicating considerably greater variability in percentage terms in wages rather than hours.

Decompose total variation of a series

We can decompose the total variation of a series x_{it} around its grand mean $\bar{x}$ into **within sum of squares** and **between sum of squares**:

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x})^2 &= \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it}-\bar{x}_i)+(\bar{x}_i-\bar{x})]^2 \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2+\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{x}_i-\bar{x})^2
\end{aligned}$$

.

This leads to within standard deviation S_{W} and between standard deviation s_{B}, where

$$s_{\mathrm{W}}^2 = \frac{1}{NT-N}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2$$

and

$$s_{\mathrm{B}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i-\bar{x})^2$$

.

```
df_between <-
  labor_ss %>%
  group_by(id) %>%
  summarise(across(contains("ln"), mean, .names = "mean_{col}")) %>%
   select(id, contains("mean_"))

df_within <-
  labor_ss %>%
  group_by(id) %>%
  mutate(across(contains("ln"), function(x) x-mean(x), .names = "demeaned_{col}")) %>%
  select(id, year, contains("demeaned_")) %>%
   ungroup()
df_fd <-
   labor_ss %>%
   group_by(id) %>%
   mutate(across(contains("ln"), function(x) x-dplyr::lag(x), .names = "fd_{col}")) %>%
   select(id, year, contains("fd_")) %>%
   ungroup()

ssw <- sapply(df_within[c("demeaned_lnhrs", "demeaned_lnwg")],
```

```r
              function(x) sum(x**2))
ssb <- sapply(df_between[c("mean_lnhrs", "mean_lnwg")],
              function(x) sum((x - mean(x))**2))

sst <- sapply(labor_ss[c("lnhrs", "lnwg")],
              function(x) sum((x - mean(x))**2))
ss <- rbind(ssw, ssb, sst) # sum of squares between, within and total variations.
dimnames(ss)[[2]] <- mainvars # in terms of the original names of variables
ss
```

```
#>       lnhrs   lnwg
#> ssw 263.677 152.18
#> ssb  17.015  81.26
#> sst 433.831 964.78
```

```r
rm(sst, ssb, ssw)
```

```r
touch_names <- function(nms, left = TRUE) {
   if (left)
      sub("(\\w+)_(\\w+)", "\\2", nms) # if stat name first e.g mean_x
   else
      sub("(\\w+)_(\\w+)", "\\1", nms) # if stat name second e.g. x_mean
   }

# to have identical variable names (against POLS) for stargazer output
names(df_between) <- touch_names(names(df_between))
names(df_within) <- touch_names(names(df_within))
names(df_fd) <- touch_names(names(df_fd))

denominator <- c(NT-N, N-1)
sd_lnhrs <- Map(function(x, y) sqrt(x/y), ss[-3, 1], denominator)
sd_lnwg <- Map( function(x, y) sqrt(x/y), ss[-3, 2], denominator)
```

The within and between standard deviations are, respectively, 0.23467 and 0.17901 for lnhrs, and 0.17828 and 0.39119.

```r
POLS <- lm(lnhrs~lnwg, labor_ss)
Between <- lm(lnhrs~lnwg, df_between)
Within <- lm(lnhrs~lnwg+0
             , df_within)
fd <- lm(lnhrs~lnwg, df_fd) # first difference
```

RE variance matrix estimate requires consistent estimates of the variance components σ_ε^2 and σ_α^2. From the within or fixed effects regression of $(y_{it} - \bar{y}_i)$ on $(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)$ we obtain

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)-K} \sum_i \sum_t \left((y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)' \widehat{\boldsymbol{\beta}}_{\mathrm{W}} \right)^2 .$$

From the between regression of $\bar{y}_i$ on an intercept and $\overline{\mathbf{x}}_i$, an equation that has error term with variance $\sigma_\alpha^2 + \sigma_\varepsilon^2/T$, we obtain

$$\widehat{\sigma}_\alpha^2 = \frac{1}{N-(K+1)} \sum_i \left(\bar{y}_i - \widehat{\mu}_{\mathrm{B}} - \overline{\mathbf{x}}_i' \widehat{\boldsymbol{\beta}}_{\mathrm{B}} \right)^2 - \frac{1}{T} \widehat{\sigma}_\varepsilon^2$$

Then, $\lambda = 1 - \frac{\sigma_\varepsilon}{(T\sigma_\alpha^2 + \sigma_\varepsilon^2)^{1/2}}$.

```r
# variance components of the RE GLS estimator
(var_epsilon <- sum((df_within[["lnhrs"]] -
                      df_within[["lnwg"]]*coef(Within))**2)/(NT-N-K)
)
```

```
#> [1] 0.054188
```

```r
(var_alpha <- sum((df_between[["lnhrs"]]- coef(Between)[[1]] -
                    df_between[["lnwg"]]*coef(Between)[[2]])**2)/(N-(K+1)) -
      (1/TT) * var_epsilon)
```

```
#> [1] 0.026001
```

```r
(lambda <- 1 - (var_epsilon)**.5/ (TT * var_alpha + var_epsilon)**.5)
```

```
#> [1] 0.58471
```

The individual effects α_{it} can be estimated by

$$\hat{\alpha}_i = \bar{y}_i - \mathbf{x}_i'\beta_{\mathbf{W}}.$$

```r
# estimate individual effects in FE or within estimator
alpha <- df_between[["lnhrs"]]-df_between[["lnwg"]]*coef(Within)
alpha_bar <- mean(alpha)
```

Thus, $\hat{\alpha}_i = 7.22$.

```r
coeffs <- lapply(list(POLS, Between, Within, fd), coef) # all coeffs comparison
coeffs[[3]] <- c("(Intercept)" = alpha_bar, coef(Within))
coeffs <- do.call(cbind, coeffs)
covariate.labels <- c("$\\beta$", "$\\alpha$")
dimnames(coeffs)[[1]] <- rev(covariate.labels)
dimnames(coeffs)[[2]] <- c("POLS" ,   "Between", "Within" , "First Diff")
knitr::kable(coeffs)
```

	POLS	Between	Within	First Diff
α	7.44152	7.48302	7.21989	0.00083
β	0.08274	0.06684	0.16768	0.10899

```r
#
# how to add coefficients by myself?
# coef = list("$\\alpha$" =coeffs[1,], "$\\beta$" = coeffs[2,])

 stargazer::stargazer(list(POLS, Between, Within, fd),
                      covariate.labels = covariate.labels,
                      dep.var.caption = "",
                      dep.var.labels = "",
                      column.labels = c("POLS", "Between", "Within", "First Diff"),
```

```
          header = FALSE,
          model.numbers = FALSE,
          digits = 3L,
          font.size = "small",
          df = FALSE,
          omit.stat = "adj.rsq",
          title = "Hours and Wages: Standard Linear Panel Model Estimators")
```

Table 2: Hours and Wages: Standard Linear Panel Model Estimators

	POLS	Between	Within	First Diff
β	0.083^{***}	0.067^{***}	0.168^{***}	0.109^{***}
	(0.009)	(0.020)	(0.018)	(0.021)
α	7.442^{***}	7.483^{***}		0.001
	(0.024)	(0.052)		(0.004)
Observations	5,320	532	5,320	4,788
R^2	0.015	0.021	0.016	0.005
Residual Std. Error	0.283	0.177	0.221	0.296
F Statistic	82.222^{***}	11.554^{***}	87.733^{***}	26.094^{***}

Note: *p<0.1; **p<0.05; ***p<0.01