Partial Solutions

1. Table 1.2 of Chapter 1 of Deaton (1997), reproduced on the next page, contains summary statistics about the joint distribution of income and consumption for Ivorian households over the two years 1985-1986. Use the notation y_{it} for the income of household i in year t, and c_{it} for the consumption of household i in year t.

   (a) What is the correlation between consumption in the two years? [Hint: use the fact that $V[c_{i,t+1} - c_{i,t}] = V[c_{i,t+1}] + V[c_{i,t}] - 2\text{cov}(c_{i,t+1}, c_{i,t})$].]
   Answer: We have $V[\Delta c] = 987^2$; $V[c_{t+1}] = 1513^2$; $V[c_t] = 1236^2$. So,

$$\text{cov}(c_{t+1}, c_t) = \frac{-1}{2}\left\{987^2 - 1513^2 - 1236^2\right\}. \quad (1)$$

   And, dividing by $\sqrt{V[c_t]V[c_{t+1}]} = 1513 \cdot 1236$, we get

$$\begin{aligned} \rho &= \frac{-1}{2}\left\{\frac{987^2}{1513 \cdot 1236} - \frac{1513}{1236} - \frac{1236}{1513}\right\} \\ &\approx 0.76. \end{aligned} \quad (2)$$

   (b) Suppose these data come from a simple random sample of households. Further, suppose there is some measurement error in the income data, so that

$$y_{it} = y_{it}^* + \varepsilon_{it} \quad (3)$$

   where y_{it}^* is true income, and ε_{it} is "noise" that is independent of y_{it}^* each period, with $E[\varepsilon_{it}] = 0$ and $V[\varepsilon_{it}] = \sigma^2$. However, measurement error may be persistent over time (perhaps because it is driven by negligence on the part of the reporting household), so let ρ be the correlation between $\varepsilon_{i,t+1}$ and ε_{it}.

   Using the notation $\Delta y_{i,t+1} = y_{i,t+1} - y_{it}$ for the observed change in income for household i, and similarly for true income y_{it}^*, how would you calculate the variance in true income growth from panel data on y_{it}, if you knew ρ and σ^2?

   Answer: The variance in observed income is $V[\Delta y] = V[\Delta y^* + \Delta\varepsilon] = V[\Delta y^*] + V[\Delta\varepsilon]$, by the independence of ε and y^*. Then, notice that $V[\Delta\varepsilon] = V[\varepsilon_{t+1}] + V[\varepsilon_t] - 2\text{cov}(\varepsilon_{t+1}, \varepsilon_t) = 2\sigma^2 - 2\sigma^2\rho = 2\sigma^2(1-\rho)$. So if we have panel data on y and we know ρ and σ^2, we can compute $V[\Delta y^*]$ as $V[\Delta y] - 2\sigma^2(1-\rho)$.

   (c) Consider two methods of estimating the mean change in income, $E[\Delta y_{i,t+1}^*]$, over this period: (i) taking two independent cross-sectional surveys, each of size n, and (ii) collecting a two-period panel, also of size n.

   For method (i), say the households surveyed at time t are labelled $i = 1, \ldots n$ and those surveyed at $t+1$ are labelled $i = n+1, \ldots 2n$. The first method leads to the estimator

$$\overline{\Delta} = \frac{1}{n}\sum_{i=n+1}^{2n} y_{i,t+1} - \frac{1}{n}\sum_{i=1}^{n} y_{i,t} \quad (4)$$

while the second leads to the estimator

$$\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n} \Delta y_{i,t+1} \tag{5}$$

Under what conditions - on σ^2, ρ, and the joint distribution of $(y^*_{i,t+1}, y^*_{i,t})$ - will the first method be more precise than the second? (Here, "precise" means having a lower variance.) Does your answer depend on the sample size? If so, why?

Answer: We have

$$\begin{aligned} V[\overline{\Delta}] &= n^{-1}V[y_{i,t+1}] + n^{-1}V[y_{it}] \\ &= n^{-1}(V[y^*_{t+1}] + V[y^*_t] + 2\sigma^2) \end{aligned} \tag{6}$$

where the first equality follows from the independence of the two cross-sections. The variance of the panel estimator is

$$\begin{aligned} V[\widehat{\Delta}] &= n^{-1}V[\Delta y] \\ &= n^{-1}(V[\Delta y^*] + 2\sigma^2(1-\rho)) \end{aligned} \tag{7}$$

Thus, $V[\overline{\Delta}] \leq V[\widehat{\Delta}]$ whenever

$$V[y^*_{t+1}] + V[y^*_t] + 2\sigma^2 \leq V[\Delta y^*] + 2\sigma^2(1-\rho) \tag{8}$$

which we can rearrange to give

$$\rho\sigma^2 \leq -\mathrm{cov}(y^*_{t+1}, y^*_t) \iff \rho\sigma^2 + \mathrm{cov}(y^*_{t+1}, y^*_t) \leq 0 \tag{9}$$

i.e when the sum of the autocovariances (of the measurement error ε_t and true income y^*_t) is less than zero, we should prefer two independent samples to a panel.[1]

To explore the intuition for this result, imagine measurement errors are independent over time, so $\rho = 0$. Then we should prefer a panel to two independent cross-sections whenever $\mathrm{cov}(y^*_{t+1}, y^*_t) > 0$, i.e true incomes are persistent (which is typically the case). This is because the cross-sectional variation in y^* represents mainly permanent differences across households (and measurement error). Thus, the changes in income are less variable than the levels.

On the other hand, if $\rho < 0$ so the measurement error is mean-reverting, repeated observations on the same households are less useful for measuring changes. In this case, the observed changes in y will tend to be dominated by fluctuations in measurement error rather than real changes in y^*.

[5 + 5 + 10 = 20 points]

This version: March 13, 2019

[1] You didn't need to prove this, but you should confirm that both estimators are *unbiased*, meaning that their expected values are $E[\Delta y^*]$.

Table 1.2. Consumption and income for panel households, Côte d'Ivoire, 1985–86
(thousands of CFAs per month)

	Mean	*Standard deviation*	*Median*	*Interquartile range*
Levels				
Consumption, 1985	1,561	1,513	1,132	1,344
Consumption, 1986	1,455	1,236	1,070	1,090
Income, 1985	1,238	1,464	780	1,137
Income, 1986	1,332	1,525	871	1,052
Differences 1986–85				
Consumption	-106	987	-17	679
Income	94	1,128	92	723

Notes: The figures shown are for total consumption and disposable income, both including imputed rental values of housing and durable goods. Data from 730 panel households.
Source: Author's and World Bank calculations using the Côte d'Ivoire Living Standards Surveys, 1985–86.