Linear regression - an example

Mankiw, Romer and Weil (QJE, 1992) present cross-section regression models for the log of income per capita, estimated using a sample of 98 countries

The Solow (QJE, 1956) version of the neoclassical growth model predicts that the steady state level of income per capita depends positively on the saving rate, and depends negatively on the rate of population growth

The Solow model predicts the linear relationship

$$\ln\left(\frac{Y}{L}\right)_i = \beta_1 + \beta_2 \ln\left(\frac{I}{Y}\right)_i + \beta_3 \ln(n+g+\delta)_i + u_i$$

or

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

where $y_i = \ln\left(\frac{Y}{L}\right)_i$ is the log of income per capita

$x_{2i} = \ln\left(\frac{I}{Y}\right)_i$ is the log of investment as a share of GDP

$x_{3i} = \ln(n+g+\delta)_i$ is the log of the population growth rate (n_i) plus the steady state growth rate (g) plus the rate at which capital depreciates (δ)

$$\ln\left(\frac{Y}{L}\right)_i = \beta_1 + \beta_2 \ln\left(\frac{I}{Y}\right)_i + \beta_3 \ln(n+g+\delta)_i + u_i$$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

The Solow model predicts that $\beta_2 > 0$ and that $\beta_3 < 0$

Moreover the model predicts that $\beta_3 = -\beta_2$

Mankiw, Romer and Weil (1992) estimate this linear model using OLS, and test the validity of this restriction

They use data from the Penn World Table

$\left(\frac{Y}{L}\right)_i$ is the ratio of real GDP in 1985 to the population of working age (aged 15-64) in 1985

$\left(\frac{I}{Y}\right)_i$ is the ratio of real investment to real GDP, averaged over the period 1960-85

n_i is the average rate of growth of the working age population over the period 1960-85

$g + \delta$ is set to 0.05 for all countries

They use a sample of 98 countries for which this data is available, excluding the major oil producing countries

The data is provided in the Appendix to the paper

Results for the unrestricted and restricted versions of this model are presented in the first column of Table 1

The note *OLS Regression Using Stata* explains how these results can be reproduced using Stata

What the OLS results show

The OLS estimates show the pattern of **partial correlations** in the data

At a given rate of population growth, there is a positive correlation between income per capita ($\ln(Y/L)_i$) and investment as a share of GDP ($\ln(I/Y)_i$)

At a given saving rate, there is a negative correlation between income per capita ($\ln(Y/L)_i$) and the rate of population growth ($\ln(n_i + g + \delta)$)

This pattern of partial correlations is *consistent with* the predictions of the Solow growth model

What the OLS results do not show (1)

This pattern of partial correlations does not *prove* that the data is generated by the Solow growth model

Other models may make similar predictions

An important assumption of the Solow model is that the aggregate production function is constant returns to scale Cobb-Douglas

$$\ln Y_i = \gamma_0 + \gamma_1 \ln K_i + (1 - \gamma_1) \ln L_i + \varepsilon_i$$

$$\text{or} \quad \ln Y_i - \ln L_i = \gamma_0 + \gamma_1 (\ln K_i - \ln L_i) + \varepsilon_i$$

$$\text{or} \qquad \ln\left(\frac{Y}{L}\right)_i = \gamma_0 + \gamma_1 \ln\left(\frac{K}{L}\right)_i + \varepsilon_i$$

So a key factor in explaining differences across countries in income per worker ($\ln(Y/L)_i$) will be differences across countries in capital per worker ($\ln(K/L)_i$)

All else equal, a higher share of investment in GDP implies more capital accumulation, raising capital per worker

All else equal, a faster rate of population growth implies more labour accumulation, reducing capital per worker

Other models which assume this technology will make similar predictions

What the OLS results do not show (2)

The causal mechanisms that explain these partial correlations in the Solow growth model are from autonomous differences in the saving rate, or autonomous differences in the population growth rate, resulting in different levels of steady state income per capita

But the causal mechanisms that explain the partial correlations we find in the data could be quite different

Within a country, we know that richer households tend to save a higher share of their income than poorer households

If richer countries also tend to save a higher share of their national income than poorer countries, there may be a causal link from higher income per capita to higher (savings and) investment as a share of GDP

This could also explain the partial correlation that we see in the data

This is an example of '**reverse causation**', with variation in the dependent variable directly producing the co-variation in the explanatory variable

Alternatively richer countries may tend to have other characteristics we have not accounted for, such as better legal protection for investors, or more developed financial systems, that also favour higher (savings and) investment as a share of GDP

Similarly, richer countries may tend to have better pension systems, or better education (particularly for women), which reduce population growth

In these examples, causality may run from '**omitted variables**' (e.g. investor protection, or female education) which may result in *both* higher income per capita *and* higher investment/lower population growth

In general, it is challenging to establish directions of causation from the patterns of partial correlations that we estimate in cross-section regressions

The key message is that finding a significant (partial) correlation between two variables y_i and x_i does not establish a causal link, either from x_i to y_i or from y_i to x_i (since the (partial) correlation may be explained by a third variable that drives both y_i and x_i)

The conclusion that **correlation is not the same as causation** also applies to the partial correlations that we describe with OLS estimates of coefficients in multivariate regression models