Standard linear model, OLS approximation, Frisch–Waugh–Lovell theorem, FE estimator

Econometrics (35B206), Lecture 1

Tunga Kantarcı, TiSEM, Tilburg University, Spring 2019

Notation

The standard linear model (SLM) for the population is simplified in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Notation

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 & x_{12} & \dots & x_{1K} \\ 1 & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \dots & x_{nK} \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

Notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{y}$: dependent variable. n$\times$1. The bold font is for observations. A vector is always a column vector.

y_i: an observation in a row of $\boldsymbol{y}$.

i: unit of study.

Notation

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

$\boldsymbol{X}$: matrix of variables. $n \times K$. The bold font indicates multiple observations. The big font indicates multiple variables.

$\boldsymbol{x}_k$: a column in $\boldsymbol{X}$. $n \times 1$. It contains n observations for variable k. k, l, m are used to indicate different columns. The bold font indicates multiple observations.

$\boldsymbol{x}_i'$: a row in $\boldsymbol{X}$. $1 \times K$. It contains observations for K variables for unit i. i, j, t, s are used to indicate different rows. The bold font indicates multiple variables.

$\boldsymbol{x}_i$: column vector formed by the transpose of a row in $\boldsymbol{X}$. K$\times$1.

x_{ik}: an observation in row i, column k of $\boldsymbol{X}$.

Notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\beta}$: true, or population, coefficient vector. K$\times$1. Unobserved.

β_k: a coefficient in a row of $\boldsymbol{\beta}$. If you let $\boldsymbol{x}_0$ be a column of 1s, β_0 is the constant term.

Notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\varepsilon}$: error. n$\times$1. Unobserved.

ε_i: an element in a row of $\boldsymbol{\varepsilon}$.

SLM

What is a standard linear regression model? Which assumptions make it ‘standard’ ?

SLM, assumptions, linearity

A1. Linearity: the model is linear in the parameters.

SLM, assumptions, linearity

The model

$$y_i = \beta_1 + x_{i2}^2 \beta_2 + \varepsilon_i$$

is linear in the parameters, linear in the squared regressor, nonlinear in the regressor.

SLM, assumptions, linearity

The model

$$y_i = \exp(\beta_1 + x_{i2}\beta_2 + \varepsilon_i)$$

is nonlinear in the parameters. However, we can linearise it if we apply the logarithmic transformation:

$$\log y_i = \beta_1 + x_{i2}\beta_2 + \varepsilon_i.$$

SLM, assumptions, linearity

The model

$$y_i = x_{i2}^{\beta_2} + \varepsilon_i$$

is nonlinear in the parameter. We can linearise it using the logarithmic transformation. But we would end up with the logarithm of the error term meaning that we have to impose a distributional assumption on the error term. This is restrictive. Later in this course, we will not impose the linearity assumption, and study how we can still estimate a model that is non-linear in the parameters.

SLM, assumptions, full column rank

A2. Full column rank: $rank\,(\boldsymbol{X}) = K$. Remember that $\boldsymbol{X}$ is $n \times K$ matrix. It contains K columns. Hence, A2 means $\boldsymbol{X}$ has full column rank.

SLM, assumptions, full column rank

A2 is not satisfied in two cases.

SLM, assumptions, full column rank

First, if $n < K$. Note that $rank\,(\boldsymbol{X}) \leq min\,(n, K)$. Hence, $rank\,(\boldsymbol{X})$ cannot be K if $n < K$. In practice this is not likely.

SLM, assumptions, full column rank

Second is the case where there is an exact relationship among any of the columns of $\boldsymbol{X}$.

SLM, assumptions, full column rank

E.g., consider the regression

$$wage_i = x_{0i}\beta_0 + d_i^{female}\beta_1 + d_i^{male}\beta_2 + \varepsilon_i$$

where $x_{0i} = 1$, and

$$d_i^{female} = \begin{cases} 1 & if \quad i = female \\ 0 & if \quad i = male \end{cases}$$

$$d_i^{male} = \begin{cases} 0 & if \quad i = female \\ 1 & if \quad i = male \end{cases}$$

Sum of the values in each row of $\boldsymbol{d}^{female}$ and $\boldsymbol{d}^{male}$ is equal to the value in that row of $\boldsymbol{x}_0$. Hence, one value can be perfectly predicted from other values. $rank\,(\boldsymbol{X}) \neq K$. This is perfect multicollinearity.

SLM, assumptions, full column rank

$$\begin{bmatrix} \boldsymbol{x}_0 & \boldsymbol{d}^{female} & \boldsymbol{d}^{male} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

SLM, assumptions, full column rank

Perfect multicollinearity is a problem for estimating β. The OLS estimator of β is given by

$$\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Rank of $\boldsymbol{X}$ is not K. Hence rank of $\boldsymbol{X}'\boldsymbol{X}$ is not K. Square matrices are invertible if they have full rank. $\boldsymbol{X}'\boldsymbol{X}$ does not have full rank and hence it is not invertible. This implies that $\hat{\beta}$ has multiple solutions.

SLM, assumptions, exogeneity

A3. Strict exogeneity:

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{x}_k\right] = 0.$$

What does this moment condition say? Recall that $\boldsymbol{x}_k$ contains n observations for variable k. The stated condition says that the expected value of the error at observation i in the sample is independent of the explanatory variable k observed at any observation, including observation i. It says that the average of the error is the same across all observations of the independent variable, and that this average is 0. More on this later.

SLM, assumptions, exogeneity

Why is it strict? Take a look at the definition of weak exogeneity:

$$\mathrm{E}\left[\varepsilon_i \mid x_{ik}\right] = 0.$$

x_{ik} is observation i for variable k. That is, we do not consider all n observations of variable k, but just observation i. That is why

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{x}_k\right] = 0$$

is strict, since all n observations of variable k are considered.

SLM, assumptions, exogeneity

Note that strict exogeneity,

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{x}_k\right] = 0,$$

can be considered to apply to all K variables as

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = 0.$$

But this is beside the point. What makes it strict is about n not K.

SLM, assumptions, exogeneity

Why do we need A3? The model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Taking the expectation conditional on $\boldsymbol{X}$,

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{X}\right] &= \mathrm{E}\left[\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{X}\right] + \mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] \\ &= \boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

That is, A3 gives the conditional expectation function, allowing us to comment on $\boldsymbol{\beta}$: if x_k changes by one unit, y changes by β_k, on average, controlling for other factors.

Later we will study the other reasons for A3, and how we can relax it.

SLM, assumptions, exogeneity

A3 has two implications. First, by the LIE,

$$\mathsf{E}\left[\varepsilon_i\right] = \mathsf{E}_{\boldsymbol{X}}\left[\mathsf{E}\left[\varepsilon_i \mid \boldsymbol{X}\right]\right] = 0.$$

SLM, assumptions, exogeneity

Second, note that

$$\mathrm{Cov}\left[\varepsilon_i, \boldsymbol{X}\right] = \mathrm{E}\left[\varepsilon_i \boldsymbol{X}\right] - \mathrm{E}\left[\varepsilon_i\right]\mathrm{E}\left[\boldsymbol{X}\right]$$

and

$$\mathrm{E}\left[\varepsilon_i \boldsymbol{X}\right] = \mathrm{E}_{\boldsymbol{X}}\left[\mathrm{E}\left[\varepsilon_i \boldsymbol{X} \mid \boldsymbol{X}\right]\right] = \mathrm{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{X}\right]\right].$$

Hence, if

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = \boldsymbol{0},$$

then

$$\mathrm{Cov}\left[\varepsilon_i, \boldsymbol{X}\right] = \boldsymbol{0}.$$

It says that ε_i is not correlated with $\boldsymbol{X}$, or any function of $\boldsymbol{X}$.

SLM, assumptions, exogeneity

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = \mathbf{0}$$

can be easily violated. E.g., suppose

$$\varepsilon_i^* = \varepsilon_i + \boldsymbol{x}_k \beta_k,$$

where $\beta_k \neq 0$, and $\boldsymbol{x}_k$ is correlated with $\boldsymbol{X}$. Then, ε_i^* is correlated with $\boldsymbol{X}$ because

$$\mathrm{E}\left[\varepsilon_i^* \mid \boldsymbol{X}\right] \neq 0.$$

This is restrictive in practice. We would want to include $\boldsymbol{x}_k$ in the model so that

$$\mathrm{E}\left[\varepsilon_i^* \mid \boldsymbol{X}\right] = 0.$$

But what if $\boldsymbol{x}_k$ is unobserved? We cannot include it.

SLM, assumptions, exogeneity

$$\mathsf{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = \mathbf{0}$$

can be violated. Take another example. Suppose

$$\mathsf{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = c.$$

This is not restrictive in practice. We can subtract c from the error, ε_i, so that

$$\mathsf{E}\left[\varepsilon_i - c \mid \boldsymbol{X}\right] = 0,$$

and add c to the constant, β_0.

SLM, assumptions, spherical errors

A4. Errors are homoskedastic and non-autocorrelated.

Homoskedasticity: each ε_i has the same variance σ^2 conditional on $\boldsymbol{X}$:

$$\text{Var}\left[\varepsilon_i \mid \boldsymbol{X}\right] = \sigma^2, \ \forall \, i.$$

Nonautocorrelation: each ε_i is uncorrelated with every other disturbance ε_j conditional on $\boldsymbol{X}$:

$$\text{Cov}\left[\varepsilon_i, \varepsilon_j \mid \boldsymbol{X}\right] = 0, \ \forall \, i \neq j.$$

Later we will study how we can relax this assumption.

SLM, assumptions, spherical errors

If $\mathrm{E}[\varepsilon_i \mid \boldsymbol{X}] = 0$,

$$\mathrm{Var}[\varepsilon_i \mid \boldsymbol{X}] = \mathrm{E}\left[\varepsilon_i^2 \mid \boldsymbol{X}\right] - (\mathrm{E}[\varepsilon_i \mid \boldsymbol{X}])^2 = \mathrm{E}[\varepsilon_i\varepsilon_i \mid \boldsymbol{X}] = \sigma^2,$$

and

$$\mathrm{Cov}[\varepsilon_i, \varepsilon_j \mid \boldsymbol{X}] = \mathrm{E}[\varepsilon_i\varepsilon_j \mid \boldsymbol{X}] - \mathrm{E}[\varepsilon_i \mid \boldsymbol{X}]\,\mathrm{E}[\varepsilon_j \mid \boldsymbol{X}] = \mathrm{E}[\varepsilon_i\varepsilon_j \mid \boldsymbol{X}] = 0.$$

The variance-covariance matrix for n errors is

$$\mathrm{Var}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] - \mathrm{E}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}]\,\mathrm{E}\left[\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right].$$

Note that $\boldsymbol{\varepsilon}$ is $n \times 1$, and hence $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ is $n \times n$. This implies that

$$\mathrm{Var}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] = \sigma^2 I_n = \sigma^2 \boldsymbol{I}$$

which is a $n \times n$ matrix.

SLM, assumptions, spherical errors

$$
\begin{aligned}
\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] &= \begin{bmatrix} \mathrm{E}\left[\varepsilon_1\varepsilon_1 \mid \boldsymbol{X}\right] & \mathrm{E}\left[\varepsilon_1\varepsilon_2 \mid \boldsymbol{X}\right] & \dots & \mathrm{E}\left[\varepsilon_1\varepsilon_n \mid \boldsymbol{X}\right] \\ \mathrm{E}\left[\varepsilon_2\varepsilon_1 \mid \boldsymbol{X}\right] & \mathrm{E}\left[\varepsilon_2\varepsilon_2 \mid \boldsymbol{X}\right] & \dots & \mathrm{E}\left[\varepsilon_2\varepsilon_n \mid \boldsymbol{X}\right] \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{E}\left[\varepsilon_n\varepsilon_1 \mid \boldsymbol{X}\right] & \mathrm{E}\left[\varepsilon_n\varepsilon_2 \mid \boldsymbol{X}\right] & \dots & \mathrm{E}\left[\varepsilon_n\varepsilon_n \mid \boldsymbol{X}\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & 1 \end{bmatrix}}_{I_n} \sigma^2
\end{aligned}
$$

SLM, assumptions, random sampling

A5. Random sampling: the data $\{(\boldsymbol{x}_i, y_i) : i = 1, 2, \ldots, n\}$ is a random sample following the population model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. It says that all elements of the data have the same probability of being selected from the population. That is, the observations are i.i.d. This implies that the data have been chosen to be representative of the population. The sample selection model deals with situations where this assumption fails. This course does not study this model.

SLM, assumptions, random sampling

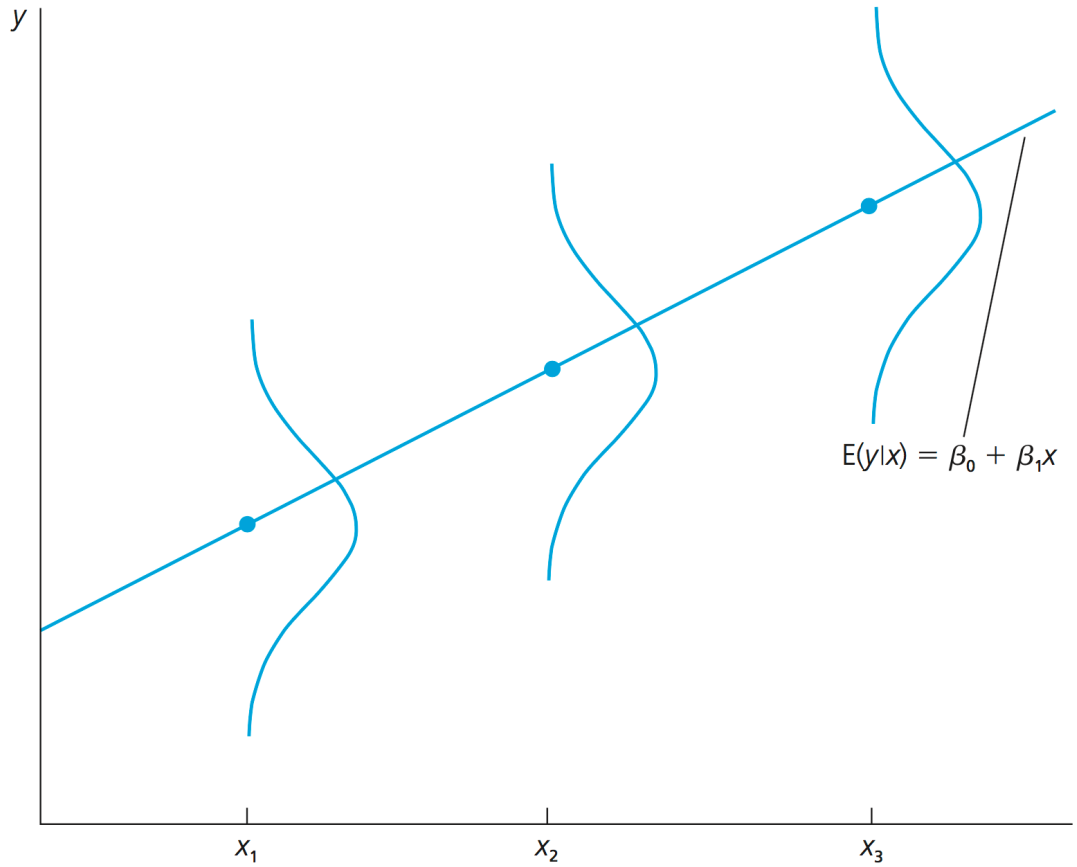
A6. ε_i is normal. That is, ε_i has the mean and variance given by A3 and A4, and has a normal distribution. That is,

$$\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N\left[\mathbf{0}, \sigma^2 \boldsymbol{I}\right].$$

We will use this assumption if n is small. We will drop this assumption if n is large.

SLM, summary of assumptions

A1: regression line is linear in $\boldsymbol{\beta}$. A3: the conditional expectation function. A4: errors have a constant variance conditional on $\boldsymbol{X}$, and hence so do $\boldsymbol{y}$. The latter because $\text{Var}\left[y_i \mid \boldsymbol{X}\right] = \text{Var}\left[\varepsilon_i \mid \boldsymbol{X}\right]$. A6: errors are normal, and hence so do $\boldsymbol{y}$. The following figure demonstrates all of these assumptions:



OLS approximation

Consider the SLM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$\boldsymbol{\beta}$ is unknown and we want to estimate it. The best estimate is the one that makes $\boldsymbol{y}$ as close to $\boldsymbol{X}\boldsymbol{\beta}$ as possible since our aim is to explain $\boldsymbol{y}$ with $\boldsymbol{X}\boldsymbol{\beta}$ as much as possible. Let $\hat{\boldsymbol{\beta}}$ be a candidate for $\boldsymbol{\beta}$ that intends to minimise the sum of squared residuals

$$S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

The necessary condition for a minimum is

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0}.$$

$\partial S(\hat{\boldsymbol{\beta}})/\partial \hat{\boldsymbol{\beta}}$ is calculated using matrix differentiation. Checkpoint. If A2 holds, $S(\hat{\boldsymbol{\beta}})$ attains a minimum at $\hat{\boldsymbol{\beta}}_{OLS}$ which takes the form

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

OLS approximation

The solution to the least squares problem is

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta}_{OLS} - \boldsymbol{X}'\boldsymbol{y} = -\boldsymbol{X}'(\boldsymbol{y} - \underbrace{\boldsymbol{X}\hat{\beta}_{OLS}}_{\hat{\boldsymbol{y}}}) = -\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = 0.$$

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta}_{OLS} = \boldsymbol{X}'\boldsymbol{y}$$

are also called the normal equations.

OLS approximation, implications

The solution has three implications:

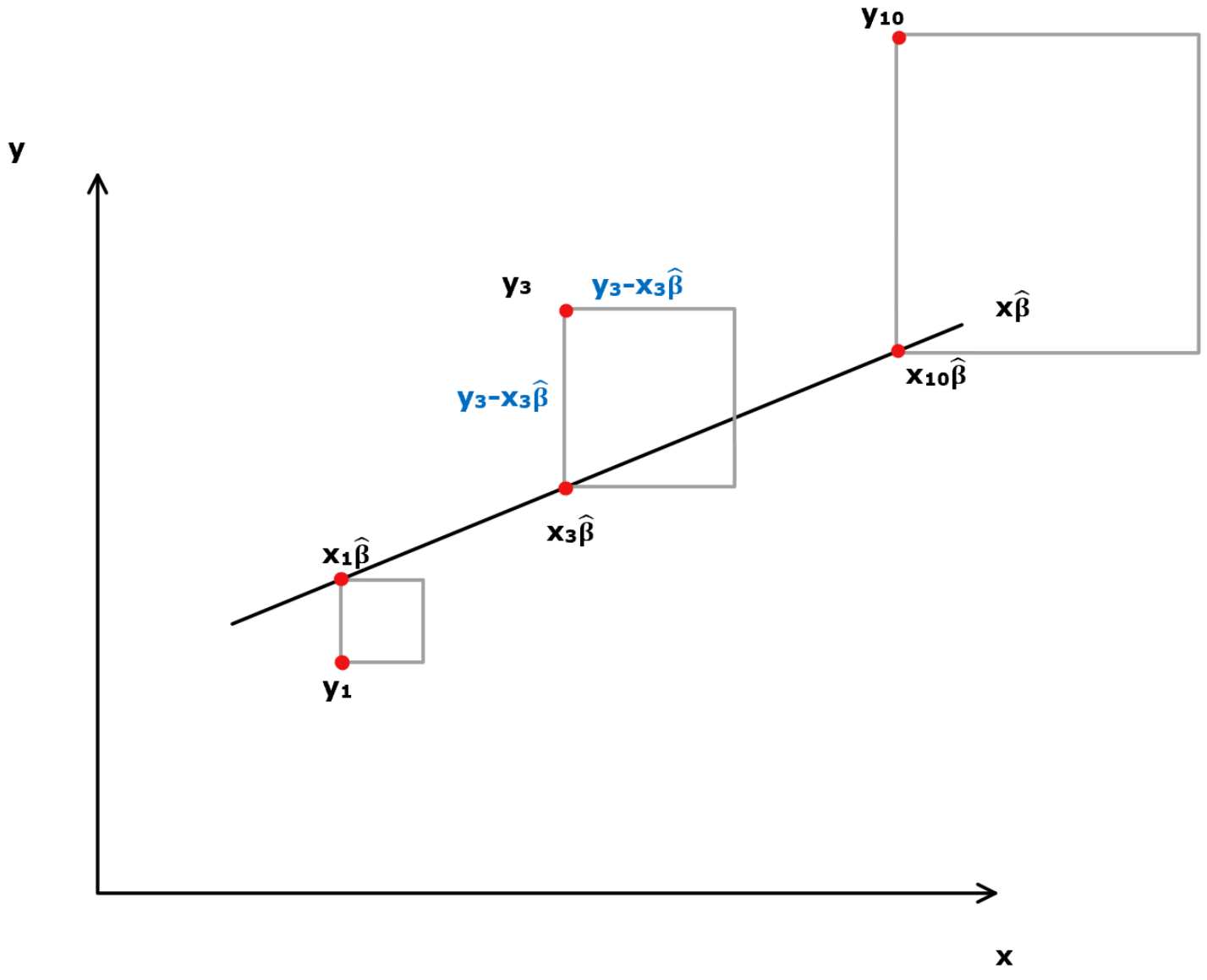
1. If the first column of $\boldsymbol{X}$, $\boldsymbol{x}_0$, is a column of 1s, i.e. the regression includes a constant, the residuals sum to zero:

$$\boldsymbol{x}_0'\hat{\boldsymbol{\varepsilon}} = \sum_i^n \hat{\varepsilon}_i = 0.$$

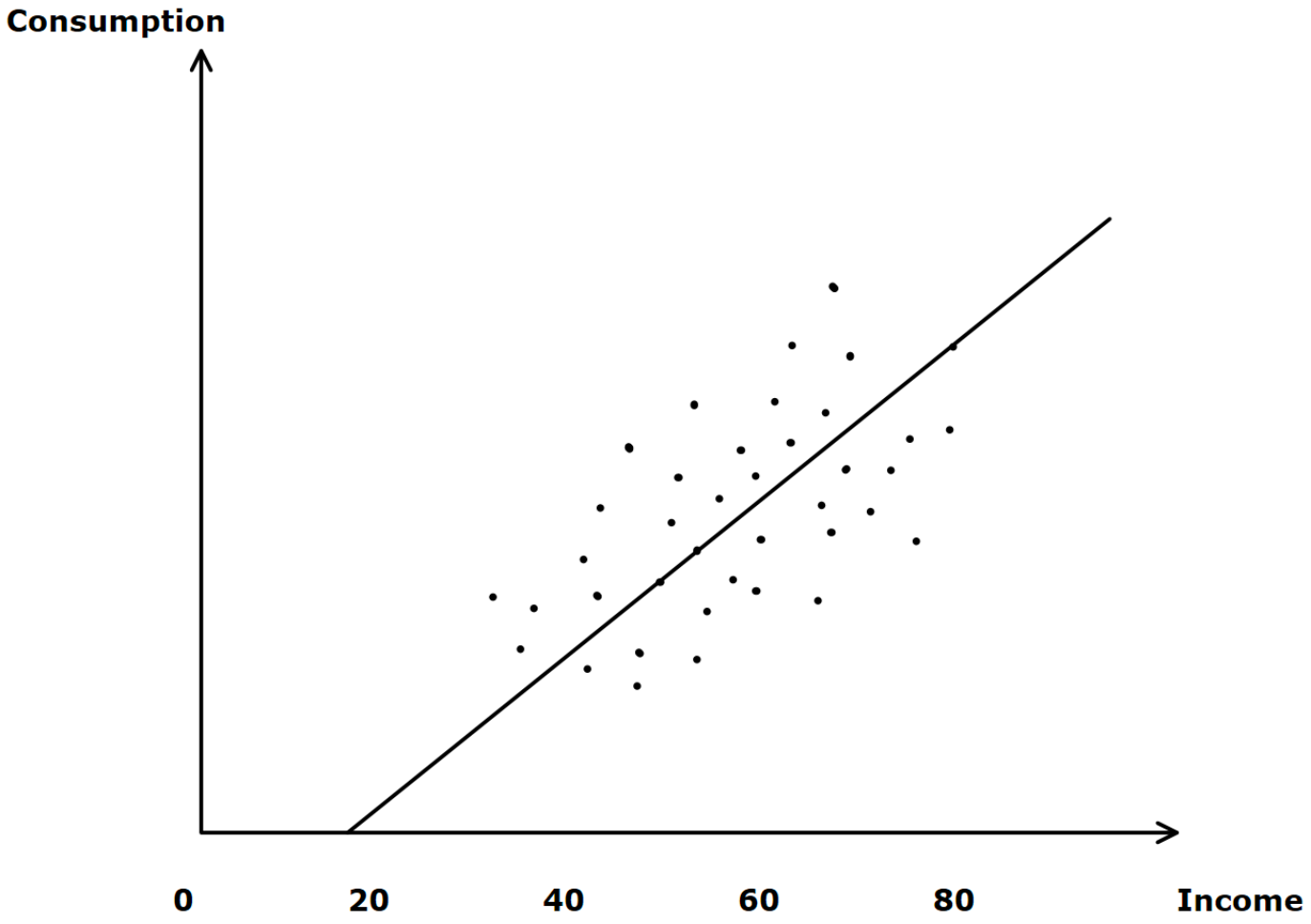
2. $\bar{y} = \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_{OLS} + \bar{\hat{\varepsilon}}$, and since $\bar{\hat{\varepsilon}} = 0$ by the first implication, $\bar{y} = \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_{OLS}$. This says that the regression hyperplane passes through the point of means of the data.

3. $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$ from the solution. Taking the means, we obtain $\bar{y} = \bar{\hat{y}} + \bar{\hat{\varepsilon}}$. Since $\bar{\hat{\varepsilon}} = 0$ by the first implication, we obtain $\bar{y} = \bar{\hat{y}}$.

OLS approximation, insights



OLS approximation, insights



OLS approximation, insights

For incomes between 40 and 80, the consumption function can be approximated by the line (model). Does the line describe the consumption-income relationship for all incomes, or only for the those in the center? Only in the center! What is the predicted consumption when income is 10? A negative value! Models are approximations. Approximations do not work well if we move too far away from the point of approximation. OLS is a good approximator around the average value of x.

OLS approximation, insights

$\hat{\boldsymbol{\beta}}$ intends to minimise the sum of squared residuals

$$(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

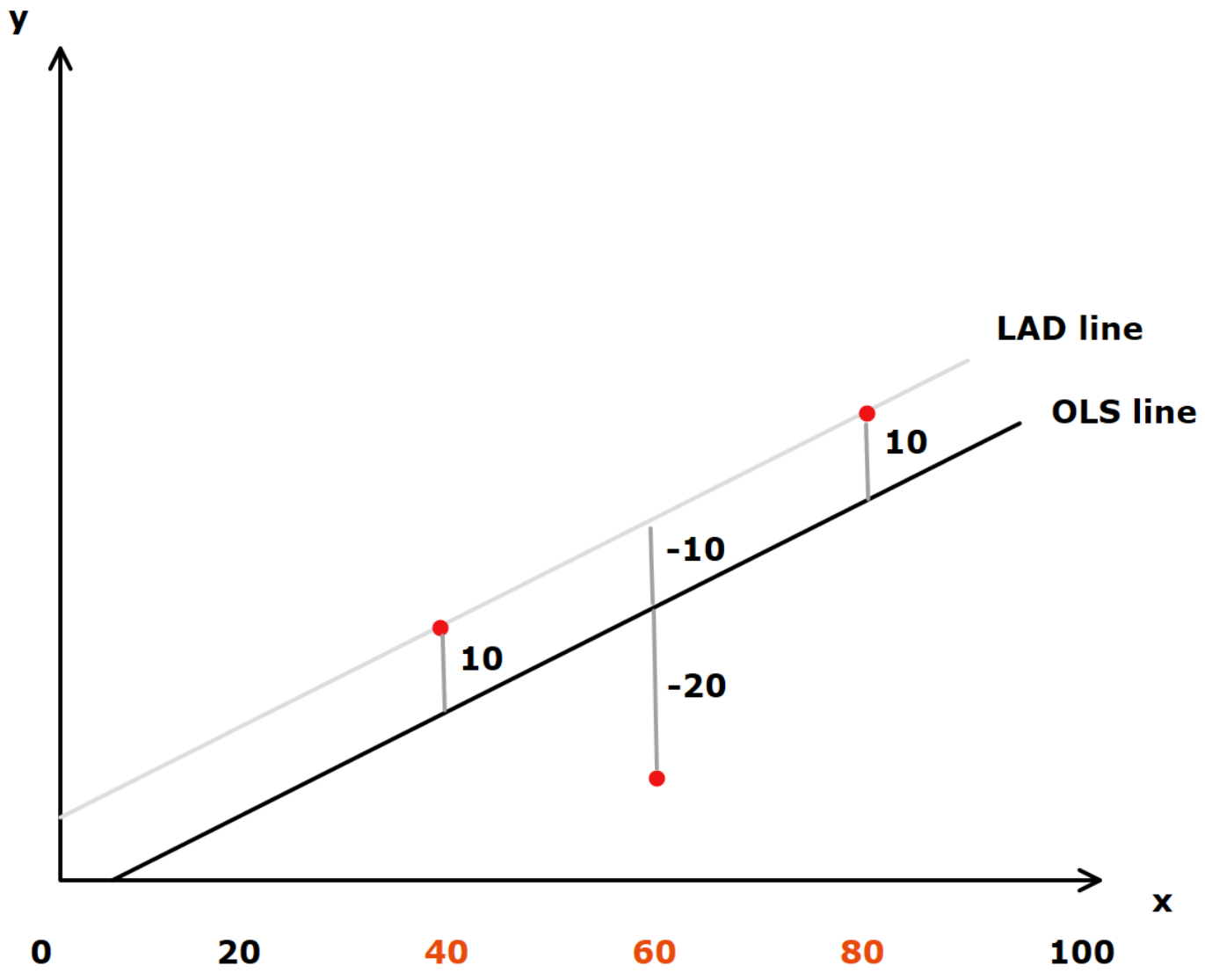
Instead, $\hat{\boldsymbol{\beta}}$ could intend to minimise the sum of absolute residuals, or deviations,

$$\boldsymbol{\iota}' \mid \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \mid .$$

$\boldsymbol{\iota}'$ is a row of ones.

Why not minimise the latter but the former?

OLS approximation, insights



OLS approximation, insights

Considering the lower line as the reference line, according to the sum of squared residuals approach,

$$\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \hat{\varepsilon}_3 \end{bmatrix} = \begin{bmatrix} 10 \\ -20 \\ 10 \end{bmatrix}$$

together with

$$\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 1 & 1 & 1 \\ 40 & 60 & 80 \end{bmatrix} \begin{bmatrix} 10 \\ -20 \\ 10 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

prove that the lower line is indeed the OLS line.

OLS approximation, insights

Considering still the lower line as the reference line, according to the sum of absolute residuals approach,

$$\boldsymbol{\iota}' \mid \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \mid = 10 + 20 + 10 = 40.$$

However, considering the upper line as the reference line, according to the sum of absolute residuals approach,

$$\boldsymbol{\iota}' \mid \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \mid = 0 + 30 + 0 = 30.$$

Hence, from the point of view of absolute deviations, the upper line is better than the lower line, and there exists no other line where the sum of the three absolute deviations is smaller than 30.

OLS approximation, insights

So why not minimise the absolute deviations but the squared deviations? One reason is that, from the point of view of absolute deviations, the upper line is better than the lower line. Most people find this counterintuitive, because one would expect the best line to lie somewhere in-between the points and not at the edge. A second reason is that absolute deviations are not differentiable which is a big disadvantage in theoretical derivations. Another reason is that minimising the squared residuals leads to and estimator, $\hat{\beta}_{OLS}$, with better statistical properties.

Geometry of OLS

Definition: the vector space spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Assume a linear model of the form

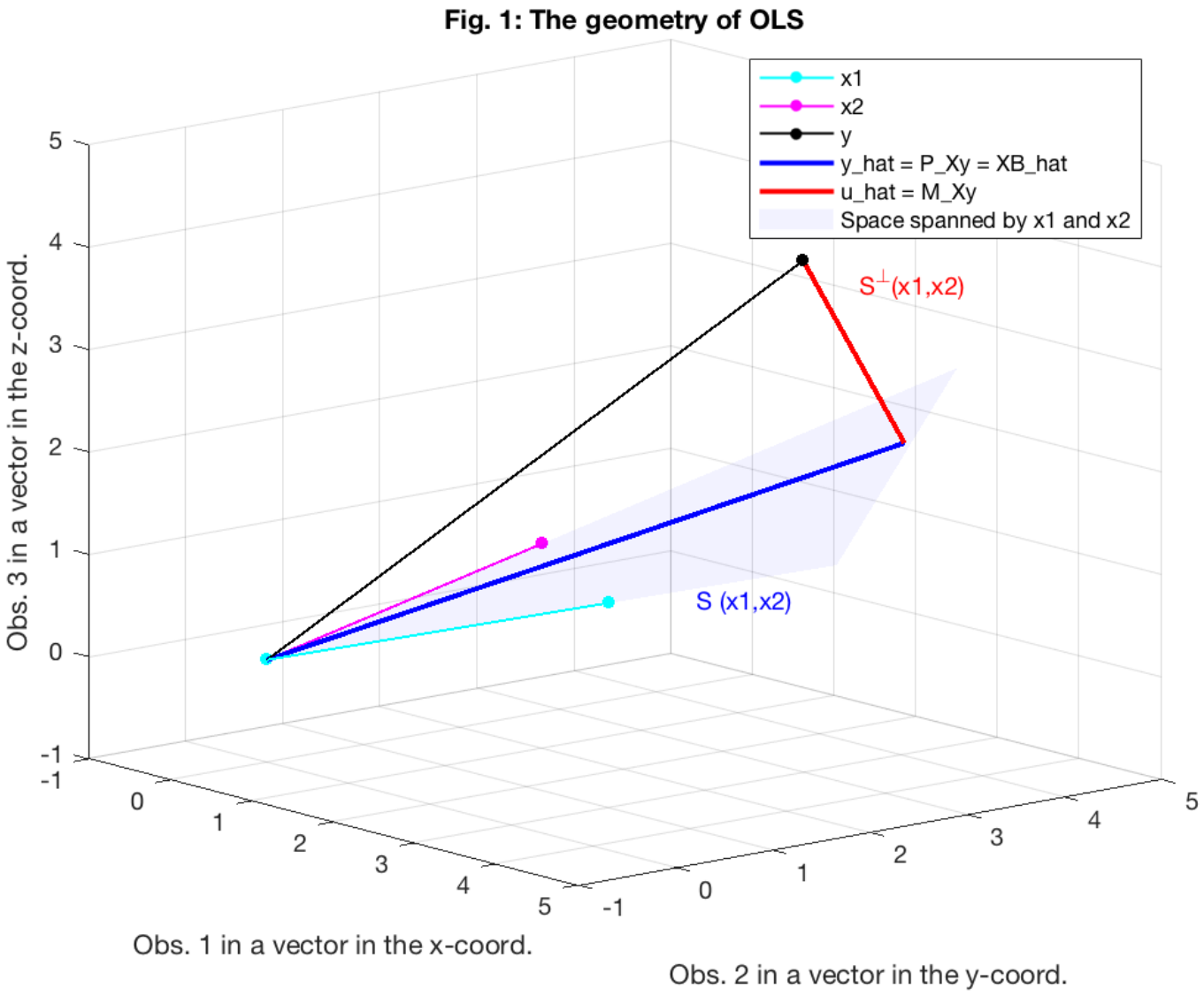
$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\varepsilon}$$

where there are 3 observations for each of the two variables. Hence, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both 3×1. The three observations of $\boldsymbol{x}_1$ are the coordinates of a point in the three dimensional space $\mathbb{R}^3$. Hence, $\boldsymbol{x}_1$ represents a vector in this space. $\boldsymbol{x}_2$ represents another vector in this space. A linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\boldsymbol{X}\boldsymbol{\beta}$, creates a new vector in this space. This vector space is called the span of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Geometry of OLS

$\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ also span a space in $\mathbb{R}^2$. The space defined by the two variables in $\mathbb{R}^2$ is a subspace of the space defined by the three observations in $\mathbb{R}^3$. That is, in a linear regression, the columns of $\boldsymbol{X}$ form a space in $\mathbb{R}^k$. This space is a subspace of the n dimensional space defined by observations.

Geometry of OLS



Geometry of OLS

What if there are two observations and three variables? That is, what if $n < K$? Draw the three vectors in the two dimensional space and see the problem!

Geometry of OLS

Definition: the projection matrix $\boldsymbol{P}$. Consider the matrix

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Two properties of $\boldsymbol{P}$ are that it is symmetric meaning that $\boldsymbol{P} = \boldsymbol{P}'$, and idempotent meaning that $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{P}$.

Premultiply $\boldsymbol{y}$ with $\boldsymbol{P}$

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{X}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}_{\hat{\boldsymbol{\beta}}_{OLS}} = \hat{\boldsymbol{y}}$$

where $\hat{\boldsymbol{y}}$ is a n$\times$1 vector containing the predictions. Note that $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$ is in the vector space spanned by $\boldsymbol{X}$. This shows that $\boldsymbol{P}$ has projected $\boldsymbol{y}$ into the vector space spanned by $\boldsymbol{X}$.

Geometry of OLS

Definition: the projection matrix $\boldsymbol{M}$. Consider the matrix

$$\boldsymbol{M} = \boldsymbol{I} - \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{\boldsymbol{P}}.$$

$\boldsymbol{M}$ is symmetric and idempotent since it is a projection matrix.

Premultiply $\boldsymbol{y}$ with $\boldsymbol{M}$

$$\boldsymbol{M}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{y} - \underbrace{\boldsymbol{P}\boldsymbol{y}}_{\hat{\boldsymbol{y}}} = \hat{\varepsilon}$$

where $\hat{\varepsilon}$ is a n$\times$1 vector containing the residuals. $\boldsymbol{M}$ is indeed called the residual maker.

Geometry of OLS

$\boldsymbol{P}$ projected $\boldsymbol{y}$ into the vector space spanned by $\boldsymbol{X}$. $\boldsymbol{M}$ projected $\boldsymbol{y}$ into some other vector space. These two vector spaces are orthogonal to each other because by definition $\boldsymbol{P}$ is orthogonal to $\boldsymbol{M}$ because $(\boldsymbol{I} - \boldsymbol{P})'\boldsymbol{P} = 0$. Hence, $\hat{\boldsymbol{\varepsilon}} \perp \hat{\boldsymbol{y}}$.

Another result is that the predictions, $\hat{\boldsymbol{y}}$, and the residuals, $\hat{\boldsymbol{\varepsilon}}$, add up to the dependent variable, $\boldsymbol{y}$, because $\boldsymbol{I}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$.

FWL Theorem, Frisch and Waugh (1933), Lovell (1963)

Consider two linear models. The first model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

$\boldsymbol{X}_1$ is a matrix of variables. $\boldsymbol{X}_2$ is a matrix of other variables.

The second model is

$$\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{v}.$$

$\boldsymbol{M}_2\boldsymbol{X}_1$ are the residuals from the regression of (all columns of) $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$.

What is important is that $\boldsymbol{M}_2\boldsymbol{X}_1 \perp \boldsymbol{X}_2$!

FWL Theorem, Frisch and Waugh (1933), Lovell (1963)

To see that $\boldsymbol{M}_2\boldsymbol{X}_1$ are the residuals from the regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$, note that

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{P}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$$

where $\boldsymbol{P}_2$ is the projection matrix for $\boldsymbol{X}_2$. Post multiply by $\boldsymbol{X}_1$ to obtain

$$\underbrace{\boldsymbol{M}_2\boldsymbol{X}_1}_{\hat{\varepsilon}} = \underbrace{\boldsymbol{X}_1 - \boldsymbol{X}_2 \underbrace{(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1}_{\hat{\beta}^{OLS}_{2,auxiliary}}}_{\hat{\boldsymbol{x}}_1}$$

where $(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1$ are the OLS estimates on $\boldsymbol{X}_2$ in the regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$. This means that $\boldsymbol{M}_2$ projects $\boldsymbol{X}_1$ into the vector space that is orthogonal to the vector space spanned by $\boldsymbol{X}_2$. Hence, $\boldsymbol{M}_2\boldsymbol{X}_1 \perp \boldsymbol{X}_2$.

FWL Theorem, Frisch and Waugh (1933), Lovell (1963)

FWL Theorem: OLS estimates of β_1 in

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \varepsilon.$$

and in

$$\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{X}_1\beta_1 + \upsilon.$$

are the same and given by

$$\hat{\beta}_1^{OLS} = (\underbrace{(\boldsymbol{M}_2\boldsymbol{X}_1)'}_{\boldsymbol{X}_1^{*\prime}}\underbrace{(\boldsymbol{M}_2\boldsymbol{X}_1)}_{\boldsymbol{X}_1^{*}})^{-1}\underbrace{(\boldsymbol{M}_2\boldsymbol{X}_1)'}_{\boldsymbol{X}_1^{*\prime}}\boldsymbol{y}$$

The theorem means that in the first model $\hat{\beta}_1^{OLS}$ gives the effect of $\boldsymbol{X}_1$ on $\boldsymbol{y}$ controlling for the effect of $\boldsymbol{X}_2$. That is, $\boldsymbol{M}_2$ enters the formula of $\hat{\beta}_1^{OLS}$! This is the power of the multiple regression analysis. It allows to do in a nonexperimental economic setting what natural scientists are able to do in a controlled laboratory setting: keeping other factors fixed. It provides this ceteris paribus interpretation although the data have not been collected in a ceteris paribus fashion.

FWL Theorem, proof

Start with the first SLM

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \varepsilon,$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n \times k_1$ and $n \times k_2$, and $k_1 + k_2 = K$. What is the algebraic solution for $\hat{\beta}_1^{OLS}$? Define

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}.$$

Then,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}.$$

Recall that the normal equations are

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta}_{OLS} = \boldsymbol{X}'\boldsymbol{y}.$$

In terms of partitioning,

$$\begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1^{OLS} \\ \hat{\beta}_2^{OLS} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

FWL Theorem, proof

Solving for $\hat{\beta}_2^{OLS}$ gives

$$\begin{aligned}\hat{\beta}_2^{OLS} &= (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} - (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1\hat{\beta}_1^{OLS} \\ &= (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'(\boldsymbol{y} - \boldsymbol{X}_1\hat{\beta}_1^{OLS}).\end{aligned}$$

This solution states that $\hat{\beta}_1^{OLS}$ is the set of coefficients in the regression of $\boldsymbol{y}$ on $\boldsymbol{X}_2$, minus a correction vector. From the normal equations in partitioned form we have

$$\boldsymbol{X}_1'\boldsymbol{X}_2\hat{\beta}_2^{OLS} + \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\beta}_1^{OLS} = \boldsymbol{X}_1'\boldsymbol{y}.$$

Insert the result for $\hat{\beta}_2^{OLS}$ into this result to obtain

$$\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1\hat{\beta}_1^{OLS} + \boldsymbol{X}_1'\boldsymbol{X}_1\hat{\beta}_1^{OLS} = \boldsymbol{X}_1'\boldsymbol{y}.$$

After collecting terms, solving for $\hat{\beta}_1^{OLS}$ gives

$$\begin{aligned}\hat{\beta}_1^{OLS} &= \left[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2')\boldsymbol{X}_1\right]^{-1}\left[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2')\boldsymbol{y}\right] \\ &= (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}.\end{aligned}$$

FWL Theorem, proof

Consider the second SLM

$$\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{v}.$$

Viewing $\boldsymbol{M}_2\boldsymbol{X}_1$ as a certain variable, the OLS estimate of $\boldsymbol{\beta}_1$ is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1^{OLS} &= (\underbrace{(\boldsymbol{M}_2\boldsymbol{X}_1)'}_{\boldsymbol{X}_1^{*\prime}}\underbrace{(\boldsymbol{M}_2\boldsymbol{X}_1)}_{\boldsymbol{X}_1^{*}})^{-1}\underbrace{(\boldsymbol{M}_2\boldsymbol{X}_1)'}_{\boldsymbol{X}_1^{*\prime}}\boldsymbol{y} \\ &= (\boldsymbol{X}_1'\boldsymbol{M}_2'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2'\boldsymbol{y}.\end{aligned}$$

Since $\boldsymbol{M}_2$ is symmetric and idempotent, we have

$$\hat{\boldsymbol{\beta}}_1^{OLS} = (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}.$$

This completes the proof.

FWL Theorem, example

Consider the regression of ***wage*** on ***educ***

```
. regress wage educ
```

Source	SS	df	MS		
Model	7842.35455	1	7842.35455	Number of obs	= 997
Residual	31031.0745	995	31.1870095	F(1, 995)	= 251.46
				Prob > F	= 0.0000
				R-squared	= 0.2017
				Adj R-squared	= 0.2009
Total	38873.429	996	39.0295472	Root MSE	= 5.5845

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	1.135645	.0716154	15.86	0.000	.9951106	1.27618
_cons	-4.860424	.9679821	-5.02	0.000	-6.759944	-2.960903

FWL Theorem, example

Consider the regression of ***wage*** on ***educ*** and ***exper***

```
. regress wage educ exper
```

Source	SS	df	MS	Number of obs	=	997
				F(2, 994)	=	172.32
Model	10008.3629	2	5004.18147	Prob > F	=	0.0000
Residual	28865.0661	994	29.0393019	R-squared	=	0.2575
				Adj R-squared	=	0.2560
Total	38873.429	996	39.0295472	Root MSE	=	5.3888

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	1.246932	.0702966	17.74	0.000	1.108985	1.384879
exper	.1327808	.0153744	8.64	0.000	.1026108	.1629509
_cons	-8.833768	1.041212	-8.48	0.000	-10.87699	-6.790542

The coefficient of ***educ*** has changed, signalling that ***educ*** and ***exper*** are correlated, and that we should control for ***exper*** in our model.

FWL Theorem, example

Consider the regression of ***educ*** on ***exper***

```
. regress educ exper
```

Source	SS	df	MS		
Model	204.317954	1	204.317954	Number of obs	= 997
Residual	5876.48847	995	5.90601856	F(1, 995)	= 34.59
				Prob > F	= 0.0000
Total	6080.80642	996	6.10522733	R-squared	= 0.0336
				Adj R-squared	= 0.0326
				Root MSE	= 2.4302

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	-.0400901	.006816	-5.88	0.000	-.0534655	-.0267147
_cons	14.04201	.1493993	93.99	0.000	13.74884	14.33519

educ and ***exper*** are negatively correlated, which explains why the coefficient of ***educ*** has decreased when we have controlled for ***exper***. Obtain the residuals of this model, and call them ***MexperXeduc***, in analogy to $\boldsymbol{M}_1\boldsymbol{X}_2$, using the Stata command

```
. predict MexperXeduc, resid
```

FWL Theorem, example

Consider the regression of ***wage*** on ***MexperXeduc***

```
. regress wage MexperXeduc, noconstant
```

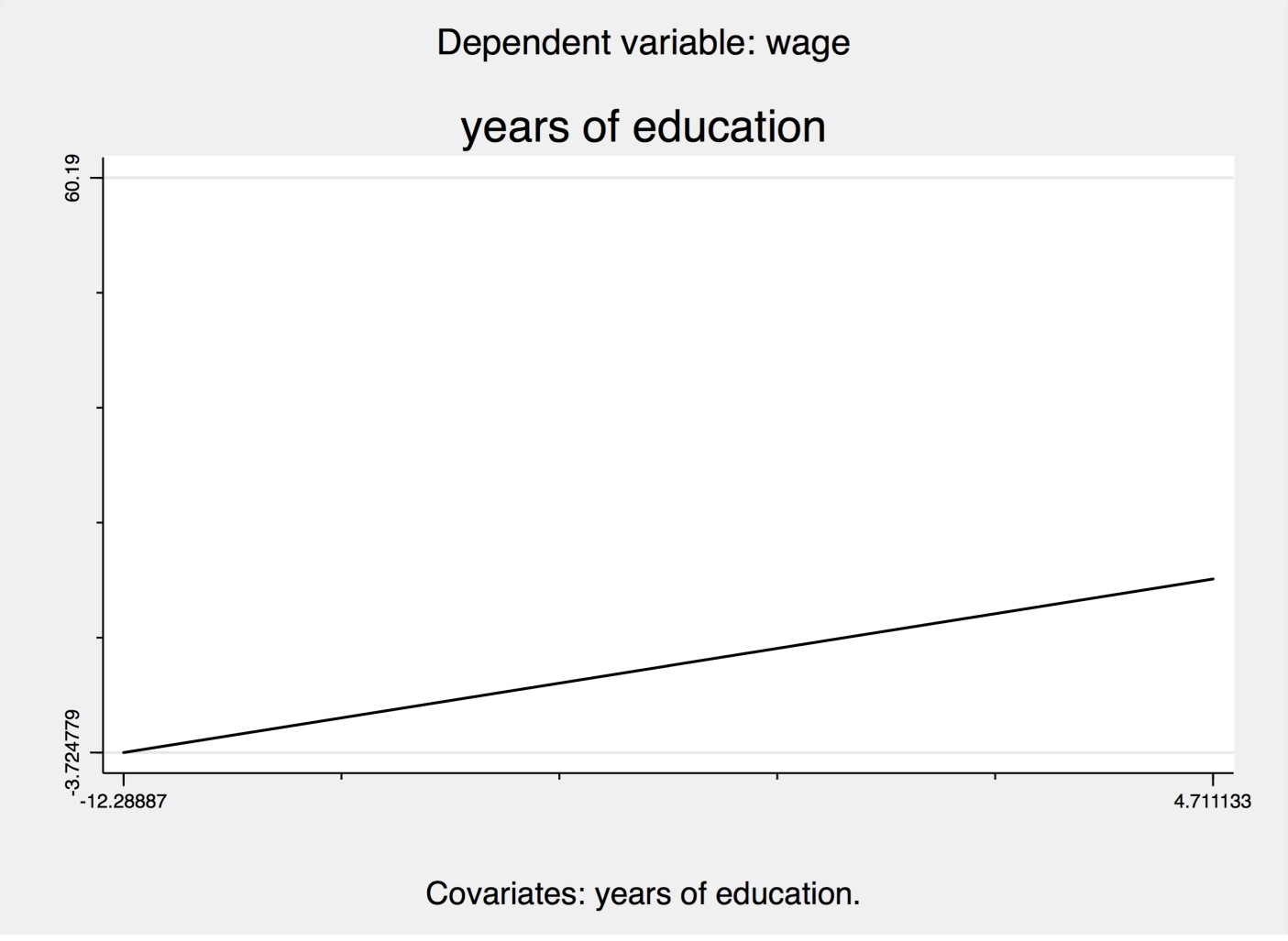
Source	SS	df	MS		
Model	9136.99599	1	9136.99599	Number of obs	= 997
Residual	134096.04	996	134.634579	F(1, 996)	= 67.87
				Prob > F	= 0.0000
				R-squared	= 0.0638
				Adj R-squared	= 0.0629
Total	143233.036	997	143.664028	Root MSE	= 11.603

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MexperXeduc	1.246932	.1513629	8.24	0.000	.9499053	1.543959

The coefficient of ***MexperXeduc*** in this regression and the coefficient of ***educ*** in the full model considered above are the same, as the FWL theorem requires.

FWL Theorem, example

The figure shows the fitted line from the regression of ***wage*** on ***educ***.



FWL Theorem, example

Adding to the figure the fitted line from the regression of ***wage*** on ***educ*** after partialling out the effect of ***exper*** (red line).

FWL Theorem, econometric application: FE estimator

Consider the linear model

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \iota\alpha_i + \varepsilon_{it}.$$

y_{it}: an observation for individual i at time t.

$\boldsymbol{x}'_{it}$: K observations for K regressors for individual i at time t. $1 \times K$.

$\boldsymbol{\beta}$: vector of true coefficients. $K \times 1$.

ι: scalar with a value of 1. Greek letter 'iota'.

α_i: time invariant constant term specific to individual i in the panel. Potentially correlated with $\boldsymbol{x}'_{it}$. It captures individual heterogeneity.

ε_{it}: error term. It meets the OLS assumptions.

FWL Theorem, econometric application: FE estimator

There are T observations available for each i. If we stack the T observations, we obtain

$$\mathbf{y}_i = \mathbf{X}_i'\boldsymbol{\beta} + \boldsymbol{\iota}\alpha_i + \boldsymbol{\varepsilon}_i.$$

$\mathbf{y}_i$: $T \times 1$.

$\mathbf{X}_i'$: T observations for i for K independent variables. $T \times K$.

$\mathbf{x}_{it}'$: row vector in row t of $\mathbf{X}_i'$. $1 \times K$. It contains k observations for k regressors for individual i at time t.

$\boldsymbol{\iota}$: column vector containing 1 in every row. $T \times 1$.

$\boldsymbol{\varepsilon}_i$: $T \times 1$.

FWL Theorem, econometric application: FE estimator

There are N individuals. If we stack the N individuals, we obtain

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

$\mathbf{y}$: $NT \times 1$.

$\mathbf{X}$: $NT \times K$.

$\mathbf{D}$: has N diagonal elements. Each element of the diagonal is a vector, is the same, and is given by the column vector $\boldsymbol{\iota}$. All of the off-diagonal elements are $\mathbf{0}$ column vectors of size $T \times 1$. Hence, $\mathbf{D}$ is $NT \times N$.

$\boldsymbol{\alpha}$: $N \times 1$ since there are N different α_is.

This is the Least Squares Dummy Variable (LSDV) model.

FWL Theorem, econometric application: FE estimator

For individual i, $T = 3$, $K = 3$,

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix} = \begin{bmatrix} k_{i1} & l_{i1} & m_{i1} \\ k_{i2} & l_{i2} & m_{i2} \\ k_{i3} & l_{i3} & m_{i3} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \alpha_i + \begin{bmatrix} \varepsilon_i \\ \varepsilon_i \\ \varepsilon_i \end{bmatrix}$$

where k, l, m represent three different regressors. Putting them into row vector $\mathbf{x}'_{it}$,

$$\underbrace{\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix}}_{\mathbf{y}_i} = \underbrace{\begin{bmatrix} \mathbf{x}'_{i1} \\ \mathbf{x}'_{i2} \\ \mathbf{x}'_{i3} \end{bmatrix}}_{\mathbf{X}'_i} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{\boldsymbol{\iota}} \alpha_i + \underbrace{\begin{bmatrix} \varepsilon_i \\ \varepsilon_i \\ \varepsilon_i \end{bmatrix}}_{\boldsymbol{\varepsilon}_i}$$

FWL Theorem, econometric application: FE estimator

Assume $N = 3$. Stack N individuals to obtain

$$\underbrace{\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_3\end{bmatrix}}_{\substack{\mathbf{y}\\ NT\times 1}} = \underbrace{\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\\ \mathbf{X}_3'\end{bmatrix}}_{\substack{\mathbf{X}\\ NT\times K}} \underbrace{\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}}_{\substack{\boldsymbol{\beta}\\ K\times 1}} + \underbrace{\begin{bmatrix}\boldsymbol{\iota} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\iota} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \boldsymbol{\iota}\end{bmatrix}}_{\substack{\boldsymbol{D}\\ NT\times N}} \underbrace{\begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix}}_{\substack{\boldsymbol{\alpha}\\ N\times 1}} + \underbrace{\begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\\ \boldsymbol{\varepsilon}_3\end{bmatrix}}_{\substack{\boldsymbol{\varepsilon}\\ N\times 1}}$$

where

$$\boldsymbol{D} = \begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 1\\ 0 & 0 & 1\end{bmatrix}$$

FWL Theorem, econometric application: FE estimator

The LSDV model has two problems. First, it requires the inversion of a very large matrix due to $\boldsymbol{D}$. Second, it requires estimation of a large number of intercept terms contained in $\boldsymbol{\alpha}$. Could we avoid these problems?

FWL Theorem, econometric application: FE estimator

Consider again the panel model for individual i at time t

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \iota\alpha_i + \varepsilon_{it}.$$

Take the average over all t for individual i to obtain

$$\bar{y}_i = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \iota\alpha_i + \bar{\varepsilon}_i,$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}.$$

Subtract the second equation from the first to obtain

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it}' - \bar{\boldsymbol{x}}_i')\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i.$$

The is the fixed effects transformation. The time invariant individual specific constant term α_i drops!

FWL Theorem, econometric application: FE estimator

We have carried out the fixed effects transformation for individual i using his T observations. We need to consider the fact that we have n individuals in the panel data.

FWL Theorem, econometric application: FE estimator

The panel model for N individuals described above is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

Consider the model

$$\mathbf{y} = \mathbf{M}_D\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon}.$$

where

$$\mathbf{M}_D = \mathbf{I} - \mathbf{P}_D.$$

and

$$\mathbf{P}_D = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'.$$

Apply the FWL theorem! The theorem states that the OLS estimator of $\boldsymbol{\beta}$ in the stated two models is the same and given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{OLS} &= ((\mathbf{M}_D\mathbf{X})'(\mathbf{M}_D\mathbf{X}))^{-1}(\mathbf{M}_D\mathbf{X})'\mathbf{y} = (\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_D\mathbf{y} \\ &= \hat{\boldsymbol{\beta}}_{FE}.\end{aligned}$$

This is the fixed effects estimator.

FWL Theorem, econometric application: FE estimator

Why does $\mathbf{M}_D$ demeans the data?

$$\mathbf{M}_D = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$$

where $\mathbf{D} = \mathbf{I}_n \otimes \boldsymbol{\iota}_T$.

FWL Theorem, econometric application: FE estimator

Assuming that $T = 3$ and $N = 2$,

$$\boldsymbol{M}_D = \begin{bmatrix} 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} & 0 & 0 & 0 \\ -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} & 0 & 0 & 0 \\ -\frac{1}{T} & -\frac{1}{T} & 1-\frac{1}{T} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} \\ 0 & 0 & 0 & -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} \\ 0 & 0 & 0 & -\frac{1}{T} & -\frac{1}{T} & 1-\frac{1}{T} \end{bmatrix}$$

FWL Theorem, econometric application: FE estimator

Then, assuming values for $\boldsymbol{X}$, $\boldsymbol{X}$ in deviation form is

$$\boldsymbol{M}_D\boldsymbol{X} = \begin{bmatrix} 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} & 0 & 0 & 0 \\ -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} & 0 & 0 & 0 \\ -\frac{1}{T} & -\frac{1}{T} & 1-\frac{1}{T} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} \\ 0 & 0 & 0 & -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} \\ 0 & 0 & 0 & -\frac{1}{T} & -\frac{1}{T} & 1-\frac{1}{T} \end{bmatrix} \begin{bmatrix} 1 & 12 \\ 1 & 13 \\ 1 & 11 \\ 3 & 43 \\ 4 & 46 \\ 4 & 41 \end{bmatrix}$$

$$= \begin{bmatrix} 1-\frac{1+1+1}{3} & 12-\frac{12+13+11}{3} \\ 1-\frac{1+1+1}{3} & 13-\frac{12+13+11}{3} \\ 1-\frac{1+1+1}{3} & 11-\frac{12+13+11}{3} \\ 3-\frac{3+4+4}{3} & 43-\frac{43+46+41}{3} \\ 4-\frac{3+4+4}{3} & 46-\frac{43+46+41}{3} \\ 4-\frac{3+4+4}{3} & 41-\frac{43+46+41}{3} \end{bmatrix}$$

This shows how the $\boldsymbol{M}_D$ matrix demeans the data. $\boldsymbol{M}_D$ is indeed called the centering matrix.