The Frisch-Waugh-Lovell Theorem

It is often useful to take deviations from the mean. Suppose that

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{x} + \boldsymbol{u}. \tag{1}$$

If $\bar{x} \equiv N^{-1} \sum_{i=1}^{N} x_i$, we can define $\boldsymbol{z}$, the vector of deviations from the mean, as $\boldsymbol{z} \equiv \boldsymbol{x} - \bar{x}\boldsymbol{\iota}$.

The vector $\boldsymbol{z}$ is easily seen to be orthogonal to $\boldsymbol{\iota}$, because

$$\boldsymbol{\iota}^\top \boldsymbol{z} = \boldsymbol{\iota}^\top (\boldsymbol{x} - \bar{x}\boldsymbol{\iota}) = N\bar{x} - \bar{x}\boldsymbol{\iota}^\top \boldsymbol{\iota} = N\bar{x} - N\bar{x} = 0. \tag{2}$$

In this case, the vector $\boldsymbol{z}$ is the **centered** version of the vector $\boldsymbol{x}$.

Centering can be performed algebraically with the orthogonal projection matrix $\boldsymbol{M}_\iota$:

$$\boldsymbol{M}_\iota \boldsymbol{x} = (\mathbf{I} - \boldsymbol{P}_\iota)\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top \boldsymbol{x} = \boldsymbol{x} - \bar{x}\boldsymbol{\iota} = \boldsymbol{z}. \tag{3}$$

If we rewrite equation (1) in terms of $\boldsymbol{z}$, it becomes

$$\boldsymbol{y} = (\beta_1 + \beta_2 \bar{x})\boldsymbol{\iota} + \beta_2 \boldsymbol{z} + \boldsymbol{u} = \alpha_1 \boldsymbol{\iota} + \alpha_2 \boldsymbol{z} + \boldsymbol{u}. \tag{4}$$

From (4), it is evident that

$$\alpha_1 = \beta_1 + \beta_2 \bar{x}, \text{ and } \alpha_2 = \beta_2. \tag{5}$$

The estimated version of (4) is

$$\boldsymbol{y} = \hat{\alpha}_1 \boldsymbol{\iota} + \hat{\alpha}_2 \boldsymbol{z} + \hat{\boldsymbol{u}} = \hat{\alpha}_1 \boldsymbol{\iota} + \hat{\beta}_2 \boldsymbol{z} + \hat{\boldsymbol{u}}. \tag{6}$$

Since one of the regressors is a constant, $\boldsymbol{\iota}^\top \hat{\boldsymbol{u}} = \sum_{i=1}^{N} \hat{u}_i = 0$.

If we had centered $\boldsymbol{y}$ as well as $\boldsymbol{x}$, we could have regressed $\boldsymbol{y} - \bar{y}\boldsymbol{\iota}$ on $\boldsymbol{z} = \boldsymbol{x} - \bar{x}\boldsymbol{\iota}$. We would have obtained the same estimate $\hat{\beta}_2$.

Now consider a model with two groups of regressors:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{7}$$

where $\boldsymbol{X}_1$ is $N \times k_1$, $\boldsymbol{X}_2$ is $N \times k_2$, and $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, with $k = k_1 + k_2$.

We could just regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1. \tag{8}$$

The fitted values are then

$$\boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{X}_1(\boldsymbol{X}_1{}^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1{}^\top\boldsymbol{y}. \tag{9}$$

An extremely important property of $\boldsymbol{P}_1$ is that

$$\boldsymbol{P}_1\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_1 = \boldsymbol{P}_1. \tag{10}$$

This follows from the fact that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}_1 = \boldsymbol{X}_1$. This fact also implies that $\boldsymbol{M}_1\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_1 = \boldsymbol{M}_{\boldsymbol{X}}$.

Consider the regression (7) again. Suppose we regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and also regress every column of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$. These regressions yield residuals $\boldsymbol{M}_1\boldsymbol{y}$ and $\boldsymbol{M}_1\boldsymbol{X}_2$.

The **FWL regression** is

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}. \tag{11}$$

It yields the same vector of OLS estimates $\hat{\boldsymbol{\beta}}_2$ as regression (7), and also the same vector of residuals.

In geometric terms, we have projected $\boldsymbol{y}$ and $\boldsymbol{X}_2$ off the subspace $\mathcal{S}(\boldsymbol{X}_1)$, or onto the subspace $\mathcal{S}^{\perp}(\boldsymbol{X}_1)$.

- To be more colloquial, we have **partialed out** the regressors in $\boldsymbol{X}_1$.
- This is extremely useful in many theoretical contexts, and it can also make computation much faster in many cases.
- Cross-section and panel regressions very often use **fixed effects**, which are usually partialed out for computational reasons.

The FWL Theorem has two parts:

1. OLS estimates of $\boldsymbol{\beta}_2$ from regressions (7) and (11) are identical.
2. OLS residuals from regressions (7) and (11) are identical.

The estimate of $\boldsymbol{\beta}_2$ from (11) is

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2{}^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2{}^\top \boldsymbol{M}_1 \boldsymbol{y}. \tag{12}$$

Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ denote the two vectors of OLS estimates from (7). Then

$$\boldsymbol{y} = \boldsymbol{P}_X \boldsymbol{y} + \boldsymbol{M}_X \boldsymbol{y} = \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_X \boldsymbol{y}. \tag{13}$$

Premultiplying the leftmost and rightmost expressions in (13) by $\boldsymbol{X}_2{}^\top \boldsymbol{M}_1$ yields

$$\boldsymbol{X}_2{}^\top \boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{X}_2{}^\top \boldsymbol{M}_1 \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2. \tag{14}$$

The r.h.s. of (14) has only one term, whereas there were three in (13). What happened?

The term $\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1$ dropped out because $\boldsymbol{M}_1$ annihilates $\boldsymbol{X}_1$.

The term $\boldsymbol{M}_X\boldsymbol{y}$ dropped out because

$$\boldsymbol{X}_2{}^\top\boldsymbol{M}_1\boldsymbol{M}_X\boldsymbol{y} = \boldsymbol{X}_2{}^\top\boldsymbol{M}_X\boldsymbol{y} = \boldsymbol{0}. \tag{15}$$

We can now solve (14) for $\hat{\boldsymbol{\beta}}_2$:

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2{}^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2{}^\top\boldsymbol{M}_1\boldsymbol{y}, \tag{16}$$

which is (12).

This proves part 1 of the theorem.

If we had premultiplied (13) by $\boldsymbol{M}_1$ instead of by $\boldsymbol{X}_2^\top \boldsymbol{M}_1$, we would have obtained

$$\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_X \boldsymbol{y}, \tag{17}$$

where the last term is unchanged from (13) because $\boldsymbol{M}_1 \boldsymbol{M}_X = \boldsymbol{M}_X$.

The regressand in (17) is the regressand from the FWL regression (11).

The first term on r.h.s. of (17) is vector of fitted values from (11), so the second term must be the residual vector.

But $\boldsymbol{M}_X \boldsymbol{y}$ is also the residual vector from (7)!

This proves part 2 of the theorem.

The FWL Theorem has many useful applications in both theory and computation. These go far beyond the linear regression model.

Using FWL often makes it unnecessary to invert partititioned matrices.

Applications of the FWL Theorem

A simple way to model seasonality is to use **seasonal dummies**.

With quarterly data, we need four dummies, s_1 through s_4. s_j is 1 in the j^{th} quarter and 0 in all other quarters. Evidently,

$$s_1 + s_2 + s_3 + s_4 = \iota. \tag{18}$$

We must drop either the constant term or one of the seasonal dummies.

It does not matter what we drop, since

$$\mathcal{S}(s_1, s_2, s_3, s_4) = \mathcal{S}(\iota, s_2, s_3, s_4) = \mathcal{S}(\iota, s_1, s_3, s_4) = \dots. \tag{19}$$

We can retain the constant and use the dummies

$$s_1' = s_1 - s_4, \quad s_2' = s_2 - s_4, \quad s_3' = s_3 - s_4. \tag{20}$$

For each complete year, these new dummy variables sum to 0.

The regression including the seasonal dummies is

$$\boldsymbol{y} = \delta_0 \boldsymbol{\iota} + \delta_1 \boldsymbol{s}_1' + \delta_2 \boldsymbol{s}_2' + \delta_3 \boldsymbol{s}_3' + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{21}$$

For any sample whose size is a multiple of 4, each of the $\boldsymbol{s}_j'$, $j = 1, 2, 3$, is orthogonal to the constant.

For t in season i, $j = 1, 2, 3$, the constant term is $\delta_0 + \delta_i$. For t belonging to season 4, it is $\delta_0 - \delta_1 - \delta_2 - \delta_3$. Thus the average of the constants for all four seasons is just δ_0, the coefficient of the constant, $\boldsymbol{\iota}$.

Let $\boldsymbol{S}$ denote any $N \times 4$ matrix that spans the constant and the four seasonal variables $\boldsymbol{s}_i$. Then regression (21) can be written as

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{\delta} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{22}$$

So can all of the equivalent regressions that write the constant and the seasonal dummies in different ways.

The FWL Theorem implies that the estimates $\hat{\boldsymbol{\beta}}$ and the residuals $\hat{\boldsymbol{u}}$ from (22) can also be obtained by running the FWL regression

$$\boldsymbol{M_S y} = \boldsymbol{M_S X \beta} + \text{residuals}, \tag{23}$$

where, as the notation suggests, $\boldsymbol{M_S} \equiv \mathbf{I} - \boldsymbol{S}(\boldsymbol{S}^\top \boldsymbol{S})^{-1}\boldsymbol{S}^\top$.

The effect of the projection $\boldsymbol{M_S}$ on $\boldsymbol{y}$ and on the explanatory variables in the matrix $\boldsymbol{X}$ can be considered as a crude form of **seasonal adjustment**, which also removes the sample means.

Notice that $\boldsymbol{M_S x}$ just subtracts the vector $\bar{\boldsymbol{x}}_q$ from $\boldsymbol{x}$, where the j^{th} element of $\bar{\boldsymbol{x}}_q$ is the quarterly mean of all the elements of $\boldsymbol{x}$ that correspond to whatever quarter observation j belongs to.

Partialing out the seasonal dummy variables is extremely cheap and easy. We just calculate four quarterly means and subtract them.

Partialing out other types of fixed effect can be done in the same way.

The equivalence of (22) and (23) is sometimes used to claim that it does not matter whether one uses "raw" data, along with seasonal dummies, or seasonally adjusted data.

This is nonsense, because actual seasonal adjustment procedures are far more complicated than using seasonal dummies.

Seasonally adjusting any time series that includes 2020-1 and 2020-2 is going to be extremely difficult!

We can remove the effect of a linear time trend in the same way that we remove the effect of seasonality.

The original equation is

$$\boldsymbol{y} = \gamma_1 \boldsymbol{\iota} + \gamma_2 \boldsymbol{T} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{24}$$

where $\boldsymbol{T}_t = t$. If we partial out $\boldsymbol{T}$, we obtain data that have been detrended (badly). This is what Frisch and Waugh (1933) was about.

Data with Two Subscripts

Let x_{gi} denote the i^{th} observation within group g on x. If there are G groups and the g^{th} group has N_g observations, a regression model is

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{25}$$

where y_{gi} and u_{gi} are scalars, $\boldsymbol{X}_{gi}$ is a row vector of length k, and $N = \sum_{g=1}^{G} N_g$.

The data can be ordered in any way, but for computational purposes it is easiest to order them first by g and then by i.

Suppose the constant term differs across each of the G groups. Then

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + \eta_g + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{26}$$

where the η_g are scalars that are often called **fixed effects**.

Instead of having a single constant term for all groups, the **fixed-effects model** (26) can be thought of as having G different constant terms, one for each group.

Regression (26) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{D\eta} + \boldsymbol{u}, \tag{27}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ have typical elements y_{gi} and u_{gi}, $\boldsymbol{X}$ has typical row $\boldsymbol{X}_{gi}$, and $\boldsymbol{D}$ contains G dummy variables.

Each column of $\boldsymbol{D}$ corresponds to one fixed effect. The constant vector $\boldsymbol{\iota}$ is a linear combination of the columns of $\boldsymbol{D}$.

To ensure that the matrix of regressors $[\boldsymbol{X} \;\; \boldsymbol{D}]$ has full rank, $\boldsymbol{X}$ must not contain either a constant or any group of variables that collectively add up to a constant vector.

The FWL Theorem applies directly to (27). If we let $\boldsymbol{M_D}$ denote $\mathbf{I} - \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$, then, by the FWL Theorem,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{D}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{D}} \boldsymbol{y}. \tag{28}$$

Each element of $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{x}$ is the deviation of x_{gi} from its group mean $\bar{x}_g$. Of course, we *never* actually compute $\boldsymbol{M}_{\boldsymbol{D}}$.

Using (for example) the `areg` command in Stata, it is easy to run regressions with hundreds or even thousands of fixed effects.

In many cases, we do not care about the $\hat{\boldsymbol{\eta}}$. If we do care, they are not hard to obtain.

Replacing $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in (27) by their estimates from the FWL regression and rearranging yields the equation

$$\boldsymbol{D}\hat{\boldsymbol{\eta}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{u}}. \tag{29}$$

Premultiplying both sides of this equation by $\boldsymbol{D}^\top$, we obtain

$$\boldsymbol{D}^\top \boldsymbol{D}\hat{\boldsymbol{\eta}} = \boldsymbol{D}^\top \boldsymbol{y} - \boldsymbol{D}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{D}^\top \hat{\boldsymbol{u}} = \boldsymbol{D}^\top \boldsymbol{y} - \boldsymbol{D}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{30}$$

The second equality in (30) holds because the residuals sum to zero over each of the G groups.

Solving equations (30) yields the result that

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{31}$$

This is just the vector of OLS estimates from a regression of $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on $\boldsymbol{D}$. It can be computed easily as

$$\hat{\boldsymbol{\eta}} = \begin{bmatrix} \bar{y}_1 - \bar{\boldsymbol{X}}_1 \hat{\boldsymbol{\beta}} \\ \bar{y}_2 - \bar{\boldsymbol{X}}_2 \hat{\boldsymbol{\beta}} \\ \vdots \\ \bar{y}_G - \bar{\boldsymbol{X}}_G \hat{\boldsymbol{\beta}} \end{bmatrix}. \tag{32}$$

Thus the estimated fixed effect $\hat{\eta}_g$ is simply the sample mean of $y_{gi} - \boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}}$ over the observations that belong to group g.

Two-Way Fixed Effects

The **two-way fixed effects**, or **TWFE**, model is extremely popular. It is often written as

$$y_{ghi} = \beta_1 + \boldsymbol{X}_{ghi}\boldsymbol{\beta} + \gamma_g + \delta_h + u_{ghi}. \tag{33}$$

If there are K explanatory variables and controls in the row vector $\boldsymbol{X}_{ghi}$, then the total number of regressors is $K + G + H - 1$.

It is possible, but tricky, to partial out both sets of fixed effects. We just need to find an efficient way to compute

$$M_{[\boldsymbol{G}\ \boldsymbol{H}]}\boldsymbol{x} = \left(\mathbf{I} - [\boldsymbol{G}\ \ \boldsymbol{H}] \begin{bmatrix} \boldsymbol{G}^\top\boldsymbol{G} & \boldsymbol{G}^\top\boldsymbol{H} \\ \boldsymbol{H}^\top\boldsymbol{G} & \boldsymbol{H}^\top\boldsymbol{H} \end{bmatrix}^{-1} [\boldsymbol{G}\ \ \boldsymbol{H}]^\top\right)\boldsymbol{x} \tag{34}$$

for any N-vector $\boldsymbol{x}$. Here $\boldsymbol{G}$ is the matrix of fixed effects for the G dimension, and $\boldsymbol{H}$ is the matrix of fixed effects for the H dimension.

In the simple case where there is just one observation for each gh pair (think of g as indexing places and h as indexing time), we can define

$$\bar{z}_{g\cdot} = \frac{1}{H}\sum_{h=1}^{H} z_{gh}, \quad \bar{z}_{\cdot h} = \frac{1}{G}\sum_{g=1}^{G} z_{gh}, \quad \text{and} \quad \bar{z} = \frac{1}{GH}\sum_{g=1}^{G}\sum_{h=1}^{H} z_{gh}. \tag{35}$$

Then, where z_{gh} represents both a typical element of $\boldsymbol{y}$ and a typical element of every column of $\boldsymbol{X}$, compute

$$\ddot{z}_{gh} = z_{gh} - \bar{z}_{g\cdot} - \bar{z}_{\cdot h} + \bar{z}. \tag{36}$$

Regressing the $\ddot{y}_{gh}$ on the $\ddot{x}^k_{gh}$ for $k = 1, \ldots, K$ is equivalent to running the TWFE regression (33).

It is common to use t and T instead of h and H when discussing the TWFE model, especially for the special case of $N = GT$. Using H makes it clear that the second dimension does not have to be time.

Influential Observations and Leverage

Each element of $\hat{\boldsymbol{\beta}}$ is a weighted average of the elements of $\boldsymbol{y}$.

Define $\boldsymbol{c}_i$ as the i^{th} row of the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Then $\hat{\beta}_i = \boldsymbol{c}_i \boldsymbol{y}$.

Because each element of $\hat{\boldsymbol{\beta}}$ is a weighted average, some observations may affect the value of $\hat{\boldsymbol{\beta}}$ much more than others do.

If we run the regression

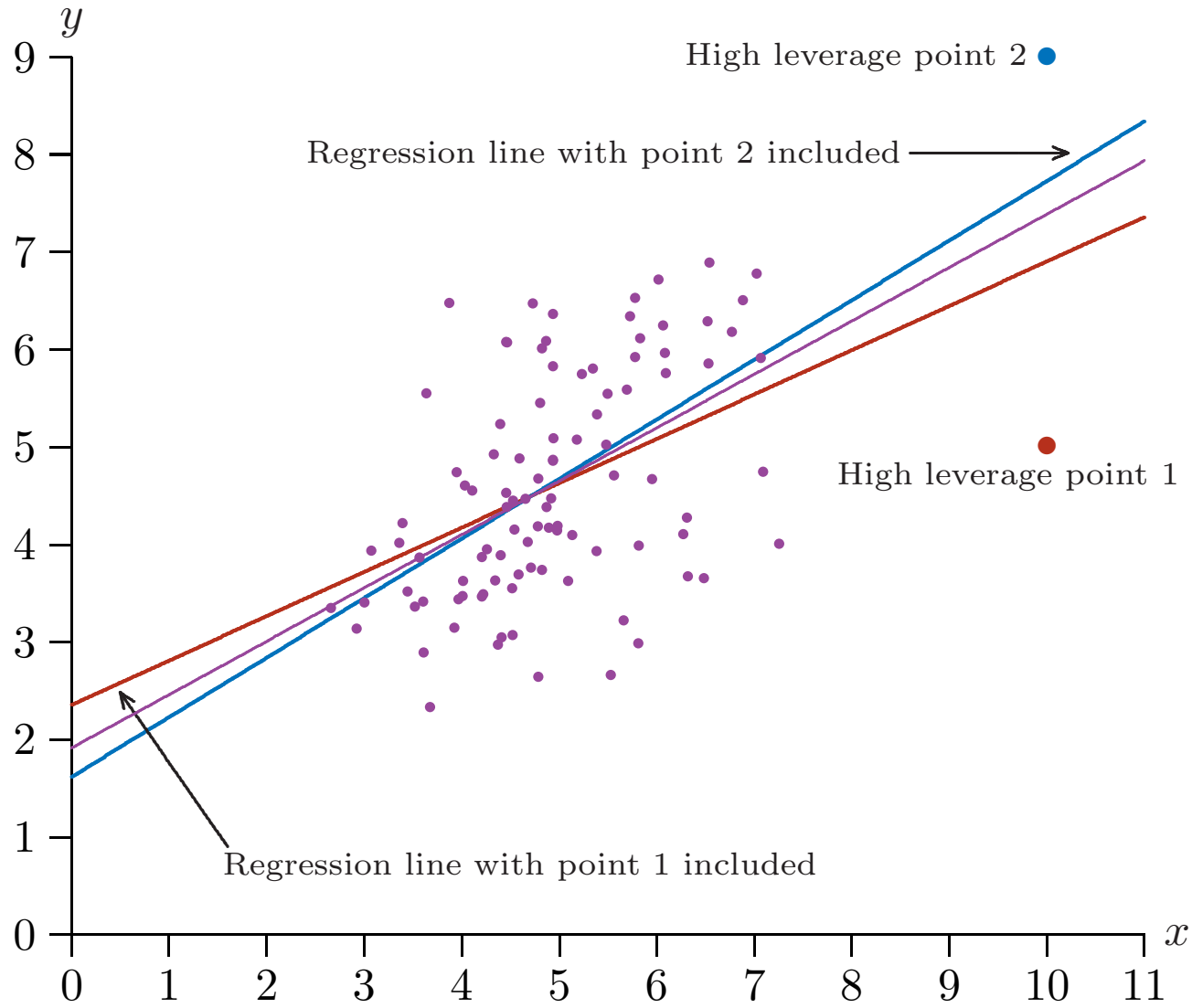
$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{x} + \boldsymbol{u} \tag{37}$$

using only the 99 observations represented by small dots in the figure, the fitted values all lie on the **regression line**,

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x. \tag{38}$$

The extra points in the figure have high leverage because their x coordinate is much larger than that of any other point in the sample.

Two Observations with High Leverage



Observations are said to be **influential** if deleting them from the sample changes some elements of $\hat{\boldsymbol{\beta}}$ substantially.

The two high-leverage points in the figure are influential.

- We can "remove" observation i by using the dummy vector $\boldsymbol{e}_i$, which has i^{th} element 1 and all other elements 0.
- The vector $\boldsymbol{e}_i$ is called a **unit basis vector**, because its norm is 1, and because all the $\boldsymbol{e}_i$, for $i = 1, \ldots, N$, span the full space E^N.

Including $\boldsymbol{e}_i$ as a regressor leads to a regression of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \alpha_i \boldsymbol{e}_i + \boldsymbol{u}, \tag{39}$$

where $\hat{\alpha}_i$ will be the residual when observation i is deleted.

This is equivalent to the FWL regression

$$\boldsymbol{M}_i \boldsymbol{y} = \boldsymbol{M}_i \boldsymbol{X}\boldsymbol{\beta} + \text{residuals}, \tag{40}$$

where $\boldsymbol{M}_i \equiv \boldsymbol{M}_{\boldsymbol{e}_i} = \mathbf{I} - \boldsymbol{e}_i(\boldsymbol{e}_i^\top \boldsymbol{e}_i)^{-1}\boldsymbol{e}_i^\top$.

$\boldsymbol{M}_i \boldsymbol{y}$ is just $\boldsymbol{y}$ with its i^{th} component replaced by 0.

Since $\boldsymbol{e}_i^\top \boldsymbol{e}_i = 1$, and because $\boldsymbol{e}_i^\top \boldsymbol{y}$ is just the i^{th} component of $\boldsymbol{y}$,

$$\boldsymbol{M}_i \boldsymbol{y} = \boldsymbol{y} - \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{y} = \boldsymbol{y} - y_i \boldsymbol{e}_i. \tag{41}$$

Thus y_i is subtracted from $\boldsymbol{y}$ for the i^{th} observation only.

Similarly, $\boldsymbol{M}_i \boldsymbol{X}$ is just $\boldsymbol{X}$ with its i^{th} row replaced by 0s.

Running regression (40) gives the same parameter estimates as those that would be obtained if observation i were deleted from the sample.

The fitted values and residuals from regression (39) are given by

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(i)} + \hat{\alpha}_i \boldsymbol{e}_i + \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}, \tag{42}$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is the vector of OLS estimates based on all observations except the i^{th}, and $\boldsymbol{M}_{\boldsymbol{Z}}$ projects off $\mathcal{S}(\boldsymbol{X}, \boldsymbol{e}_i)$.

Starting with (42), a little algebra shows that

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}}) = -\hat{\alpha}_i \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i. \tag{43}$$

We can compute the difference between $\hat{\boldsymbol{\beta}}^{(i)}$ and $\hat{\boldsymbol{\beta}}$ using this equation.

We previously applied the FWL Theorem to (39) to find $\hat{\boldsymbol{\beta}}^{(i)}$. We can also use it to find that

$$\hat{\alpha}_i = \frac{\boldsymbol{e}_i^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}}{\boldsymbol{e}_i^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{e}_i}. \tag{44}$$

The numerator here is the i^{th} element of $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$. The denominator is the i^{th} diagonal element of $\boldsymbol{M}_{\boldsymbol{X}}$.

It follows that

$$\hat{\alpha}_i = \frac{\hat{u}_i}{1 - h_i}, \tag{45}$$

where h_i denotes the i^{th} diagonal element of $\boldsymbol{P}_{\boldsymbol{X}}$, sometimes called the **hat matrix**, because $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y} = \hat{\boldsymbol{y}}$.

Using (43) and (44), we find that

$$\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}} = \frac{-1}{1 - h_i} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_i^\top \hat{u}_i. \tag{46}$$

Thus how influential an observation is depends on both $\hat{u}_i$ and h_i.

Aside: Equations (45) and (46) make it very easy to compute **jackknife estimates** and perform **delete-one cross-validation**.

Another way to measure influence is to look at the impact of each observation on its own residual (and fitted value):

$$\hat{u}_i^{(i)} - \hat{u}_i = \frac{\hat{u}_i}{1 - h_i} - \hat{u}_i = \frac{\hat{u}_i - (1 - h_i)\hat{u}_i}{1 - h_i} = \frac{h_i}{1 - h_i} \hat{u}_i. \tag{47}$$

Observations with large h_i have **high leverage** or are **leverage points**. A leverage point has the potential to be influential, depending on $\hat{u}_i$.

Clusters can also have high leverage and perhaps be influential. See MacKinnon, Neilsen, and Webb (*Stata Journal*, 2023).

We can express h_i as

$$h_i = \boldsymbol{e}_i^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i = \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i\|^2. \tag{48}$$

Since the rightmost expression here is a square, $h_i \geq 0$.

Because $\|\boldsymbol{e}_i\| = 1$, we find that $h_i = \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i\|^2 \leq 1$. Thus

$$0 \leq h_i \leq 1. \tag{49}$$

Moreover, when there is a constant term, $h_i \geq 1/N$.

If $\boldsymbol{X}$ consisted of the constant vector $\boldsymbol{\iota}$, then $h_i = \boldsymbol{e}_i^\top \boldsymbol{P}_{\boldsymbol{\iota}} \boldsymbol{e}_i = 1/N$.

If other regressors are present, then

$$1/N = \|\boldsymbol{P}_{\boldsymbol{\iota}} \boldsymbol{e}_i\|^2 = \|\boldsymbol{P}_{\boldsymbol{\iota}} \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i\|^2 \leq \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i\|^2 = h_i. \tag{50}$$

There is a special case in which h_i equals 1. If one column of $\boldsymbol{X}$ is $\boldsymbol{e}_i$, then $h_i = \boldsymbol{e}_i^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_i = \boldsymbol{e}_i^\top \boldsymbol{e}_i = 1$.

It is shown in Section 2.6 that, when there are N observations and k regressors, the average of the h_i is k/N.

This uses the fact that $\text{Tr}(\boldsymbol{P_X}) = k$.

When the diagonal elements of $\boldsymbol{P_X}$ are all close to their average value, no observation has very much leverage. Such an $\boldsymbol{X}$ matrix is said to have a **balanced design**.

When some of the h_i are much larger than k/N, and others inevitably smaller, the $\boldsymbol{X}$ matrix is said to have an **unbalanced design**.

The h_i tend to be larger for values of the regressors that are farther away from their average over the sample.

The figure shows the h_i for 100 observations in a simple regression model. Average value is $2/100 = 0.02$, but h_i varies from 0.0100 for values of x_i near the sample mean to 0.0695 for the largest value of x_i, which is about 2.4 standard deviations above the sample mean.

Leverage as a Function of the Value of a Regressor