LECTURE 14: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 20, 2019

1. Multivariate linear regression

Consider the regression model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{iK} + \epsilon_i \tag{1}$$

Using Matrix Algebra:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

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

$\boldsymbol{X}\boldsymbol{\beta}$ is the matrix product of a $n \times K$ matrix with a $K \times 1$ matrix, resulting in a $n \times 1$ matrix.

2. Ordinary Least Squares (OLS) estimator

Our dataset is: $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ for $i = 1, \ldots, n$. In matrix form, the data matrix is $[\boldsymbol{y}, \boldsymbol{X}]$. Data-generating process (DGP): $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik}, \epsilon_i)$ are i.i.d random sample from $(Y, X_1, X_2, \ldots, X_k, \epsilon)$ such that $Y = \sum_{k=1}^{K} \beta_k X_k + \epsilon$. We want to estimate the vector of parameters $\boldsymbol{\beta}$.

Make additional assumption on the data-generating process, derive k number of moments conditions, and use Method of Moments: $\mathbb{E}[\epsilon|X_1] = 0$, $\mathbb{E}[\epsilon|X_2] = 0$, $\ldots$, $\mathbb{E}[\epsilon|X_k] = 0$. This is known as the conditional mean independence assumption, or the exogeneity assumption.

Recall that $\mathbb{E}[\epsilon|X] = 0$ implies that $\mathbb{E}[X\epsilon] = 0$, $\text{Cov}(X, \epsilon) = 0$, and also that $\mathbb{E}[\epsilon] = 0$.

The sample moment conditions can be written as: $\sum_{i=1}^{n} x_{i1}\epsilon_i = 0$, $\sum_{i=1}^{n} x_{i2}\epsilon_i = 0$, $\ldots$, $\sum_{i=1}^{n} x_{ik}\epsilon_i = 0$. In matrix notation:

$$\boldsymbol{X}_1^T \boldsymbol{\epsilon} = 0 \tag{2}$$

$$\vdots$$

$$\boldsymbol{X}_k^T \boldsymbol{\epsilon} = 0 \tag{3}$$

Where $\boldsymbol{X}_k$ is the k-th row of $\boldsymbol{X}$, and $\boldsymbol{X}^T$ is the matrix transpose of $\boldsymbol{X}$.

Finally, the moment conditions can be summarized as just:

$$\boldsymbol{X}^T \boldsymbol{\epsilon} = \boldsymbol{0} \tag{4}$$

Now, we can derive the OLS (Ordinary Least Square) estimators:

$$\boldsymbol{X}^T \boldsymbol{\epsilon} = \boldsymbol{0} \tag{5}$$

$$\boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0} \tag{6}$$

$$\boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \tag{7}$$

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \tag{8}$$

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{\beta} = \boldsymbol{0} \tag{9}$$

$$\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{10}$$

3. The geometry of OLS estimation

Consider the matrix of regressors:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_K \end{bmatrix} \tag{11}$$

Each column vector $\boldsymbol{x}_k$ is a point in the Euclidean space $\mathbb{R}^n$, where n is the number of observations (and the length of the vector $\boldsymbol{x}_k$). Now consider the set of all possible points (vectors) that can be achieved as a linear combination of $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_K$. That is, $\mathcal{S}(\boldsymbol{X}) = \{\boldsymbol{z} \in \mathbb{R}^n \big| \boldsymbol{z} = \sum_{k=1}^K \alpha_k \boldsymbol{x}_k, \, \alpha_k \in \mathbb{R}\}$. The object $\mathcal{S}(\boldsymbol{X}) \subseteq \mathbb{R}^n$ is called a subspace of $\boldsymbol{X}$ (or the column space of $\boldsymbol{X}$, or the space span by $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_K$).

Now observe that $\boldsymbol{X}\boldsymbol{\beta}$ is just a linear combination of the columns of $\boldsymbol{X}$:

$$\boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_K \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} = \sum_{k=1}^K \beta_k \boldsymbol{x}_k \tag{12}$$

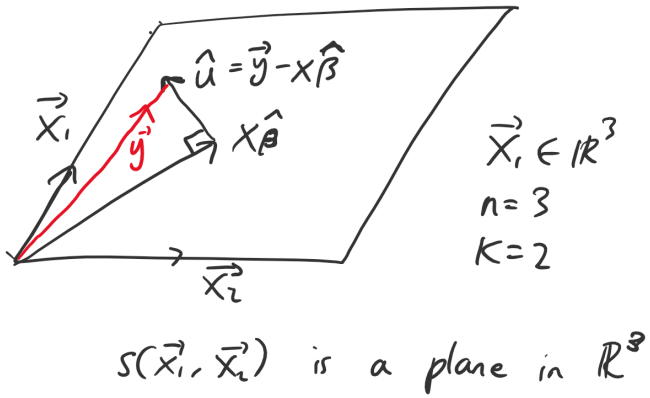
Therefore, the fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ lies in the subspace of $\boldsymbol{X}$. Recall that the OLS estimator $\hat{\boldsymbol{\beta}}$ is a solution to $\boldsymbol{X}^T(y - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, which implies that $\boldsymbol{x}_i^T(y - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$ for all $k = 1, \dots, K$, and that $\boldsymbol{x}_k$ is *orthogonal* to $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. We term $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the vector of fitted residuals. Therefore, we have $\sum_{i=1}^n x_{ik}\hat{u}_i = 0$ for $k = 1, \dots, n$.

Since the vector $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is orthogonal to all the column vectors of $\boldsymbol{X}$, it is orthogonal to the subspace of $\boldsymbol{X}$. As such, the vector of fitted value $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is also orthogonal to $\hat{\boldsymbol{u}}$, i.e., $\sum_{i=1}^n \hat{y}_i\hat{u}_i = 0$.

Therefore, for a given vector $\boldsymbol{y}$, OLS finds $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, the vector in the subspace of $\boldsymbol{X}$, that is the closest to $\boldsymbol{y}$. In another words, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{y}$ onto the subspace of $\boldsymbol{X}$.

The implication is that OLS decomposes $\boldsymbol{y}$ into a systematic part explained by $\boldsymbol{X}$, and a residual part that is unrelated/orthogonal to $\boldsymbol{X}$. Moreover, $\boldsymbol{y}$, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ forms a right-angled triangle in Euclidean space, and we can apply the Pythagoras' theorem:

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\hat{\boldsymbol{u}}\|^2 \tag{13}$$



In words, the total sum of squares is equal to the explained sum of squares plus the sum of squared residuals. This motivates the use of R^2 as the goodness-of-fit of OLS.

$$R^2 = \frac{\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y}\|^2} = \frac{(\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\boldsymbol{y}^T\boldsymbol{y}} = \frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\boldsymbol{y}^T\boldsymbol{y}} \tag{14}$$

Which always lies between 0 and 1. $\|\boldsymbol{y}\|^2$ is the Euclidean length of the vector $\boldsymbol{y}$.

3.1. Linear transformation of regressors

Suppose we take a regressor $\boldsymbol{x}_k$ and transform it as $a\mathbf{1} + b\boldsymbol{x}_k$, where $\mathbf{1}$ is a vector of ones of the same dimension as $\boldsymbol{x}_k$.

How does OLS estimation change? *The fitted OLS values and the residuals do not change! (The estimated coefficients would change however, but we can just apply the inverse of the transformation on the estimated coefficients)* This is a consequence of OLS as an orthogonal projection.

If we apply a linear transformation on $\boldsymbol{X}$, the subspace of $\boldsymbol{X}$ does not change. Therefore the orthogonal projection of $\boldsymbol{y}$ onto the same subspace would give the same OLS fitted values and residuals.

To see this, let $\boldsymbol{A}$ be a $k \times k$ non-singular matrix. Then, $\boldsymbol{XA}$ is a nonsingular linear transformation. For instance, $\boldsymbol{X}$ consists of two regressors, $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{T}]$, where $\boldsymbol{T}$ is the recorded temperature in Celsius. The transformation $\boldsymbol{F} = 32 \cdot \mathbf{1} + \frac{9}{5}\boldsymbol{T}$ can then be expressed as:

$$\begin{bmatrix}\mathbf{1} & \boldsymbol{F}\end{bmatrix} = \begin{bmatrix}\mathbf{1} & \boldsymbol{T}\end{bmatrix}\begin{bmatrix}1 & 32\\ 0 & 9/5\end{bmatrix} \tag{15}$$

Now verify that regressing $\boldsymbol{y}$ on $\boldsymbol{X}$ is equivalent to regressing $\boldsymbol{y}$ on $\boldsymbol{XA}$.

Let $\boldsymbol{\beta}$ be the OLS estimate of regressing $\boldsymbol{y}$ on $\boldsymbol{X}$, and let $\hat{\boldsymbol{\beta}}_A$ be the OLS estimate of regressing $\boldsymbol{y}$ on $\boldsymbol{XA}$. Show that $\hat{\boldsymbol{\beta}}_A = \boldsymbol{A}^{-1}\boldsymbol{\beta}$, and therefore the fitted values are the same.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_A &= ((\boldsymbol{XA})^T(\boldsymbol{XA}))^{-1}(\boldsymbol{XA})^T\boldsymbol{y}\\ &= (\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{XA})^{-1}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

We have used the fact that $(AB)^T = B^TA^T$. Moreover, suppose we have two (square) invertible matrices A and B. Then, $(AB)^{-1} = B^{-1}A^{-1}$. Since $\boldsymbol{A}$ and $\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{X}$ are both *square* invertible,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_A &= ((\boldsymbol{XA})^T(\boldsymbol{XA}))^{-1}(\boldsymbol{XA})^T\boldsymbol{y}\\ &= (\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{XA})^{-1}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y}\\ &= \boldsymbol{A}^{-1}(\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y}\\ &= \boldsymbol{A}^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{A}^T)^{-1}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y}\\ &= \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}\end{aligned}$$

Therefore the fitted value is $\boldsymbol{XAA}^{-1}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

3.2. Orthogonal projection

Define the projection matrix as:

$$P_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{16}$$

Multiplying $P_{\boldsymbol{X}}$ with any vector $\boldsymbol{y}$ results in the orthogonal projection of $\boldsymbol{y}$ on to the subspace of $\boldsymbol{X}$:

$$P_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{17}$$

Suppose that $\boldsymbol{y}$ can be perfectly explained by the regressors $\boldsymbol{X}$, i.e. $\boldsymbol{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ for some vector $\tilde{\boldsymbol{\beta}}$ without any error term $\boldsymbol{\epsilon}$. then check that $P_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{y}$, and so the OLS estimator is $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{u}} = \mathbf{0}$.

Projection matrix $P_{\boldsymbol{X}}$ is idempotent: $P_{\boldsymbol{X}}P_{\boldsymbol{X}} = P_{\boldsymbol{X}}$.

Define another projection matrix $M_{\boldsymbol{X}} = \boldsymbol{I} - P_{\boldsymbol{X}}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. While $P_{\boldsymbol{X}}$ creates the fitted value, $M_{\boldsymbol{X}}$ is the residual maker. We know from the last section that:

$$\boldsymbol{y} = P_{\boldsymbol{X}}\boldsymbol{y} + M_{\boldsymbol{X}}\boldsymbol{y} \tag{18}$$

Hence $P_{\boldsymbol{X}}\boldsymbol{y}$ is the vector in the subspace of $\boldsymbol{X}$ closest to $\boldsymbol{y}$, while $M_{\boldsymbol{X}}\boldsymbol{y}$ is the vector that is unrelated to $\boldsymbol{X}$.

Consider the regression model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$, where we have partitioned $\boldsymbol{X}$ into two groups of regressors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. For example, $\boldsymbol{X}_1$ is a $n \times k_1$ matrix while $\boldsymbol{X}_2$ is a $n \times k_2$ matrix, with $K = k_1 + k_2$, and $\boldsymbol{X}$ is a $n \times K$ matrix.

Consider the projection matrix $P_{X_1} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$ and $P_X = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. As a shorthand, $P_1 = P_{X_1}$. A very useful property of projection matrices is that:

$$P_1P_X = P_XP_1 = P_1 \tag{19}$$

Therefore if we fit $\boldsymbol{y}$ to $\boldsymbol{X}$, then take the fitted value $P_X\boldsymbol{y}$ and fit it to $\boldsymbol{X}_1$, we get $P_1\boldsymbol{y}$, which is the fitted value of $\boldsymbol{y}$ on $\boldsymbol{X}_1$. Similarly, if we fit $\boldsymbol{y}$ to $\boldsymbol{X}_1$, then take the fitted value $P_1\boldsymbol{y}$ and fit it to $\boldsymbol{X}$, we get $P_1\boldsymbol{y}$.

Proof:

$$\begin{aligned} P_XP_1 &= P_X\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T \\ &= \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T \\ &= P_1 \end{aligned}$$

$P_X\boldsymbol{X}_1 = \boldsymbol{X}_1$ because all columns of $\boldsymbol{X}_1$ belong to the subspace of $\boldsymbol{X}$, so the orthogonal projection of $\boldsymbol{X}_1$ onto the subspace of $\boldsymbol{X}$ is just itself. In another words, if we regress (each column of) $\boldsymbol{X}_1$ on $\boldsymbol{X}$, then the fitted value is just $\boldsymbol{X}_1$ since $\boldsymbol{X}_1$ can be explained perfectly by $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$.

Finally, we have:

$$\begin{aligned} P_1^T &= (P_XP_1)^T \\ P_1 &= P_1^TP_X^T \\ P_1 &= P_1P_X \end{aligned}$$

It is straightforward to verify that $P_1 P_X = P_X P_1 = P_1$ implies that $M_1 M_X = M_X M_1 = M_X$.

4. Frisch-Waugh-Lovell (FWL) theorem

Consider the regression model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}$, where we have partitioned $\boldsymbol{X}$ into two groups of regressors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals} \tag{20}$$

$$M_{\boldsymbol{X}_1}\boldsymbol{y} = M_{\boldsymbol{X}_1}\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals} \tag{21}$$

Where $M_{\boldsymbol{X}_1} = \boldsymbol{I} - P_{\boldsymbol{X}_1}$, $P_{\boldsymbol{X}_1} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

The OLS estimates $\boldsymbol{\beta}_2$ from (20) and (21) are identical.

When we apply $M_{\boldsymbol{X}_1}$ to $\boldsymbol{y}$, it removes and purges $\boldsymbol{X}_1$ from $\boldsymbol{y}$, as such $M_{\boldsymbol{X}_1}\boldsymbol{y}$ is orthogonal and unrelated to $\boldsymbol{X}_1$.

FWL theorem is very important, as it means that we can first regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$, collect the residual as $\tilde{\boldsymbol{y}}$. Then regress each of the column vector of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, collect the residuals as $\tilde{\boldsymbol{X}}_2$. Finally, regress $\tilde{\boldsymbol{y}}$ on $\tilde{\boldsymbol{X}}_2$ as if $\boldsymbol{X}_1$ has been controlled for.

For example, a regression involving time-series data that have been de-trend and de-seasonanalize is equivalent to a regression involving the original time-series data plus seasonality dummy and time trend variables. FWL also underlies the 2-stage least square (2SLS) procedure in instrumental variable estimation.

As a shortcut, let $M_1 = M_{\boldsymbol{X}_1}\boldsymbol{X}_2$. Projection matrices have the property that $M_1 M_1 = M_1$.

The OLS from the second regression is:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_2 &= ((M_1\boldsymbol{X}_2)^T M_1\boldsymbol{X}_2)^{-1}(M_1\boldsymbol{X}_2)^T M_1\boldsymbol{y} \\
&= (\boldsymbol{X}_2^T M_1^T M_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T M_1^T M_1\boldsymbol{y} \\
&= (\boldsymbol{X}_2^T M_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T M_1\boldsymbol{y}
\end{aligned}$$

Let $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ be the OLS from the first regression. As such,

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + M_X\boldsymbol{y} \\
\boldsymbol{X}_2^T M_1\boldsymbol{y} &= \boldsymbol{X}_2^T M_1\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2^T M_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2^T M_1 M_X\boldsymbol{y} \\
&= \boldsymbol{X}_2^T M_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2^T M_1 M_X\boldsymbol{y} \\
&= \boldsymbol{X}_2^T M_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + (M_X M_1\boldsymbol{X}_2)^T\boldsymbol{y}
\end{aligned}$$

Now $M_X M_1 \boldsymbol{X}_2 = \boldsymbol{0}$. Recall from the last section that $M_X M_1 = M_X$. As such, $M_X \boldsymbol{X}_2 = \boldsymbol{0}$ because $\boldsymbol{X}_2$ can be perfectly explained by $\boldsymbol{X}$ and so there is no residual left.

Therefore $\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^T M_1 \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T M_1\boldsymbol{y} = \tilde{\boldsymbol{\beta}}_2$

5. Unbiasedness of OLS estimators

Is the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ unbiased? What does unbiasedness mean here? We need a ground truth, and say that it is unbiased with respect to a data-generating process.

DGP: $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik}, \epsilon_i)$, for $i = 1, \ldots, n$, are generated from some joint distribution that obeys the equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$. We can be agnostic about this joint distribution, in particular, ϵ_i may not even be i.i.d across i.

Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator.

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \tag{22}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon})] \tag{23}$$

$$= \mathbb{E}[\boldsymbol{\beta}_0 + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}] \tag{24}$$

$$= \boldsymbol{\beta}_0 + \mathbb{E}[\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}|\boldsymbol{X}]] \tag{25}$$

$$= \boldsymbol{\beta}_0 + \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]] \tag{26}$$

The Law of Iterated Expectation is applied in the last two equations. It is clear that a sufficient condition for the unbiasedness of OLS estimator is that $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0}$. This expression means that $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$ for all $i = 1, \ldots, n$. Further unpacking, it means that ϵ_i for each $i = 1, \ldots, n$ is (conditionally mean) independent from the entire matrix $\boldsymbol{X}$, i.e. $\mathbb{E}[\epsilon_i|x_{11}, \ldots, x_{ik}, \ldots, x_{nk}] = 0$.

Two possible ways to satisfy $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$.

(1) $(\epsilon_i, x_{i1}, x_{i2}, \ldots, x_{iK})$ are independently and identically distributed across i from some probability distributions, *and*, $\mathbb{E}[\epsilon_i|x_{i1}, x_{i2}, \ldots, x_{iK}] = 0$. In the i.i.d case, we

can drop the i subscript, and write $\mathbb{E}[\epsilon|X_1, X_2, \ldots, X_K] = 0$. Now $\mathbb{E}[\epsilon|X_1, X_2, \ldots, X_K] = 0$ implies that $\mathbb{E}[\epsilon|X_k] = 0$ for $k = 1, \ldots, K$. This is what we assumed when we use the Method of Moments to derive the OLS estimator. This is a sufficient but not a necessary condition for unbiasedness.

(2)* $(\epsilon_i, x_{i1}, x_{i2}, \ldots, x_{iK})$ are *not necessarily* i.i.d across i, but $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$ for each i. Therefore, in the context of time-series where i.i.d does not hold true, OLS can still be unbiased if x_{tk} is independent of ϵ_t for all $t = 1, \ldots, T$. There can be no correlation between the error term at time t and your covariates at time $t + 1, t + 2, \ldots$.

Exogeneity alone guarantees unbiasedness. Exogeneity can be violated under many circumstances – whenever the regressor is correlated with the error term. Let us consider cross-sectional data, so that we are in the i.i.d scenario (the first scenario above). The cross-sectional data consist of gas stations' quantities sold (number of gallons) and prices, across multiple gas stations at a given hour. If we regress quantities sold on price, we expect the coefficient to be negative, i.e. high price causes lower demand. But often this regression gives us positive price coefficient! This is because price is an endogenous variable.

Prices are set strategically by firms. Suppose some gas station locations are popular because of their more friendly staffs. Customers are willing to pay more for service friendliness, and so firms respond by charging higher prices. Now it is highly unlikely that we can ever observe or even measure service friendliness. Therefore, service friendliness becomes part of the error term ϵ that can explain demand. It follows that ϵ correlates with price here (positive correlation according to our story). In general, even if we can measure and observe service friendliness, there could be some unobserved preference shocks that somehow drive higher demand at one location, and which correlates with prices.

If price is set randomly (experimentation or A/B testing), then it would be exogenous. This is the foremost concern in any empirical research. The branch of statistics/econometrics dealing with this concern is called *causal inference*. Tools that fall under causal inference include (1) Instrumental Variable approach, (2) Difference-in-difference, (3) Regression discontinuity, (4) Propensity score matching, (5) experimentation and A/B testing.