Forecasts Based on Linear Projection

DENG Yanfei
dengyf@fudan.edu.cn
www.scholat.com/idengyf.en

November 10, 2019

1 Principles of Forecasting

Suppose we want to know the value of y_{T+1} based on a sample of size T (i.e., $y_1 = Y_1, y_2 = Y_2, \ldots, y_T = Y_T$). Let $\mathbf{x}_T = \{y_1, y_2, \ldots, y_T\}$ and let $y^*_{T+1|T}$ denote a forecast of y_{T+1} based on $\mathbf{x}_T$. To evaluate the usefulness of this forecast, we need to specify a loss function (the mean squared error, MSE) and

$$\begin{aligned} \min_{\boldsymbol{\phi}} \mathrm{MSE}(y^*_{T+1|T}) &\equiv \mathbb{E}(y_{T+1} - y^*_{T+1|T})^2, \\ \Rightarrow \qquad y^*_{T+1|T} &= \mathbb{E}(y_{T+1}|\mathbf{x}_T), \end{aligned}$$

i.e., the forecast with the smallest mean squared error turns out to be the expectation of y_{T+1} conditional on $\mathbf{x}_T$.

Proof: See table 1.

The representation with general notations

Suppose we want to forecast Y_{t+1} based on its T most recent values, i.e., $(c, Y_t, Y_{t-1}, \ldots, Y_{t-T+1})' \equiv \mathbf{x}_t$,

$$\begin{aligned} \min_{\boldsymbol{\phi}} \mathrm{MSE}(y^*_{t+1|t}) &\equiv \mathbb{E}(y_{t+1} - y^*_{t+1|t})^2, \\ \Rightarrow \qquad y^*_{t+1|t} &= \mathbb{E}(y_{t+1}|\mathbf{x}_t). \end{aligned}$$

1) Recall forecasts of AR(1) without an initial condition

$$y_t = \phi y_{t-1} + \epsilon_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}, \quad |\phi| < 1.$$

$$\xRightarrow{\text{move fowward}} \begin{cases} y_{t+1} = \phi y_t + \epsilon_{t+1}, \\ y_{t+2} = \phi y_{t+1} + \epsilon_{t+2} = \phi(\phi y_t + \epsilon_{t+1}) + \epsilon_{t+2} = \phi^2 y_t + \phi\epsilon_{t+1} + \epsilon_{t+2}, \\ \vdots \\ y_{t+h} = \phi y_{t+h-1} + \epsilon_{t+h}; \end{cases}$$

$$\xrightarrow{\text{unconditional expectation}} \mathbb{E} y_t = 0,$$

$$\xrightarrow{\text{conditional expectation}} \mathbb{E}_t y_t = \phi y_{t-1},$$

$$\xRightarrow{\text{1 step ahead forecast}} \mathbb{E}_t y_{t+1} = \mathbb{E}_t(\phi y_t) + \mathbb{E}_t \epsilon_{t+1} = \phi \mathbb{E}_t y_t + 0 = \phi y_t,$$

$$\xRightarrow{\text{2 steps ahead forecasts}} \mathbb{E}_t y_{t+2} = \mathbb{E}_t(\phi y_{t+1}) = \phi \mathbb{E}_t y_{t+1} = \phi(\phi y_t) = \phi^2 y_t,$$

$$\vdots$$

$$\xRightarrow{\text{h steps ahead forecasts}} \mathbb{E}_t y_{t+h} = \mathbb{E}_t(\phi y_{t+h-1}) = \phi \mathbb{E}_t y_{t+h-1} = \phi^h y_t;$$

$$\xRightarrow{\text{forecast error variances}} \begin{cases} \text{var}_t y_{t+1} = \mathbb{E}_t[(y_{t+1} - \mathbb{E}_t y_{t+1})^2] = \mathbb{E}_t[(y_{t+1} - \phi y_t)^2] = \sigma_\epsilon^2, \\ \text{var}_t y_{t+2} = \mathbb{E}_t[(y_{t+2} - \phi^2 y_t)^2] = \mathbb{E}_t[(\phi\epsilon_{t+1} + \epsilon_{t+2})^2] = (1+\phi^2)\sigma_\epsilon^2, \\ \vdots \\ \text{var}_t y_{t+h} = \mathbb{E}_t[(y_{t+h} - \phi^h y_t)^2] = (1 + \phi^2 + \phi^4 + \cdots + \phi^{2(h-1)})\sigma_\epsilon^2. \end{cases}$$

2) Consider forecasts of AR(1) with an initial condition

$$y_t = \phi y_{t-1} + \epsilon_t = \phi^t y_0 + \sum_{i=0}^{t-1} \phi^i \epsilon_{t-i}, \quad |\phi| < 1, \; y_0 \text{ given.}$$

$$\xRightarrow{\text{move fowward}} \begin{cases} y_{t+1} = \phi y_t + \epsilon_{t+1}, \\ y_{t+2} = \phi y_{t+1} + \epsilon_{t+2} = \phi(\phi y_t + \epsilon_{t+1}) + \epsilon_{t+2} = \phi^2 y_t + \phi\epsilon_{t+1} + \epsilon_{t+2}, \\ \vdots \\ y_{t+h} = \phi y_{t+h-1} + \epsilon_{t+h}; \end{cases}$$

$$\xRightarrow{\text{1 step ahead forecast}} \mathbb{E}_t y_{t+1} = \mathbb{E}_t(\phi y_t) + \mathbb{E}_t \epsilon_{t+1} = \phi \mathbb{E}_t y_t + 0 = \phi y_t,$$

$$\xRightarrow{\text{2 steps ahead forecasts}} \mathbb{E}_t y_{t+2} = \mathbb{E}_t(\phi y_{t+1}) = \phi \mathbb{E}_t y_{t+1} = \phi(\phi y_t) = \phi^2 y_t,$$

$$\vdots$$

$$\xRightarrow{\text{h steps ahead forecasts}} \mathbb{E}_t y_{t+h} = \mathbb{E}_t(\phi y_{t+h-1}) = \phi \mathbb{E}_t y_{t+h-1} = \phi^h y_t.$$

Q1: What's differences about forecasts with and without an initial condition?

$$\mathbb{E}_t \qquad \Leftrightarrow \qquad \mathbb{E}[\cdot|\mathbf{x}_t]$$

Q2: What's differences about forecasts between conditional expectations and projections?

nonlinear vs. linear

2 Linear Projection vs. Conditional Expectation

$$y^*_{t+1|t} = c + \phi_1 y_t + \phi_2 y_{t-1} + \cdots + \phi_T y_{t-T+1} = \begin{bmatrix} c & \phi_1 & \phi_2 & \cdots & \phi_T \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ y_{t-1} \\ \vdots \\ y_{t-T+1} \end{bmatrix} \equiv \boldsymbol{\phi}'\mathbf{x}_t = \mathbb{P}(y_{t+1}|1, \mathbf{x}_t) = \mathbb{E}(y_{t+1}|\mathbf{x}_t).$$

$$\mathbf{0}' = \mathbb{E}[(y_{t+1} - \mathbb{E}_t y_{t+1})\mathbf{x}_t'] \overset{\text{posit the forecase error}\perp\mathbf{x}_t}{\Leftrightarrow} \mathbb{E}[(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)\mathbf{x}_t'] = \mathbf{0}' \leftarrow \text{linear projection of } y_{t+1} \text{ on } \mathbf{x}_t.$$

The linear projection turns out to produce the smallest mean squared error among the class of linear forecasting rules.

Table 1: Proof

Linear Projection	Conditional Expectation
Let $\mathbf{a}'\mathbf{x}_t$ is any arbitrary linear forecasting rule	Let $y^*_{t+1\|t} = f(\mathbf{x}_t)$ is any function other than conditional expectation
$\text{MSE} = \mathbb{E}(y_{t+1} - \mathbf{a}'\mathbf{x}_t)^2$	$\text{MSE} = \mathbb{E}[y_{t+1} - f(\mathbf{x}_t)]^2$
$= \mathbb{E}(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t + \boldsymbol{\phi}'\mathbf{x}_t - \mathbf{a}'\mathbf{x}_t)^2$	$= \mathbb{E}[y_{t+1} - \mathbb{E}_t y_{t+1} + \mathbb{E}_t y_{t+1} - f(\mathbf{x}_t)]^2$
$= \mathbb{E}(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)^2 + 2\mathbb{E}[(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)(\boldsymbol{\phi}'\mathbf{x}_t - \mathbf{a}'\mathbf{x}_t)]$ $+\mathbb{E}(\boldsymbol{\phi}'\mathbf{x}_t - \mathbf{a}'\mathbf{x}_t)^2$	$= \mathbb{E}(y_{t+1} - \mathbb{E}_t y_{t+1})^2 + 2\mathbb{E}[(y_{t+1} - \mathbb{E}_t y_{t+1})(\mathbb{E}_t y_{t+1} - f(\mathbf{x}_t))]$ $+\mathbb{E}[\mathbb{E}_t y_{t+1} - f(\mathbf{x}_t)]^2$
$= \mathbb{E}(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)^2 + 2\underbrace{\mathbb{E}[(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)\mathbf{x}_t'}_{0}(\boldsymbol{\phi} - \mathbf{a})]$ $+\mathbb{E}(\boldsymbol{\phi}'\mathbf{x}_t - \mathbf{a}'\mathbf{x}_t)^2$	$= \mathbb{E}(y_{t+1} - \mathbb{E}_t y_{t+1})^2 + \underbrace{2\mathbb{E}[(y_{t+1} - \mathbb{E}_t y_{t+1})(\mathbb{E}_t y_{t+1} - f(\mathbf{x}_t))]}_{0 \text{ (cf. Hamilton, 1994, p.73)}}$ $+\mathbb{E}[\mathbb{E}_t y_{t+1} - f(\mathbf{x}_t)]^2$
$\Rightarrow \mathbb{E}(y_{t+1} - \mathbf{a}'\mathbf{x}_t)^2 = \mathbb{E}(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)^2 + \mathbb{E}(\boldsymbol{\phi}'\mathbf{x}_t - \mathbf{a}'\mathbf{x}_t)^2$	$\Rightarrow \mathbb{E}[y_{t+1} - f(\mathbf{x}_t)]^2 = \mathbb{E}(y_{t+1} - \mathbb{E}_t y_{t+1})^2 + \mathbb{E}[\mathbb{E}_t y_{t+1} - f(\mathbf{x}_t)]^2$
$\xrightarrow{\text{min MSE}} \mathbb{E}(y_{t+1} - \mathbf{a}'\mathbf{x}_t)^2 = \mathbb{E}(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)^2$	$\xrightarrow{\text{min MSE}} \mathbb{E}[y_{t+1} - f(\mathbf{x}_t)]^2 = \mathbb{E}(y_{t+1} - \mathbb{E}_t y_{t+1})^2$
$\Rightarrow \mathbb{P}(y_{t+1}\|\mathbf{x}_t) \equiv \hat{y}_{t+1\|t} = \boldsymbol{\phi}'\mathbf{x}_t$	

$$\underbrace{\text{MSE}[\mathbb{P}(y_{t+1}|\mathbf{x}_t)] = \min \text{MSE} \geq \text{MSE}[\mathbb{E}(y_{t+1}|\mathbf{x}_t)] = \min \text{MSE}}_{\text{The conditional expectation offers the best possible forecast.}}$$

3 Linear Projection vs. OLS Regression

key words: $\boldsymbol{\phi}$ and observations.

LP

The coefficient for a linear projection (LP) of y_{t+1} on $\mathbf{x}_t$ is the value of $\boldsymbol{\phi}$ that

$$\begin{aligned} \mathbf{0}' &= \mathbb{E}[(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)\mathbf{x}_t'], \\ \Rightarrow \mathbb{E}(y_{t+1}\mathbf{x}_t') &= \boldsymbol{\phi}'\mathbb{E}(\mathbf{x}_t\mathbf{x}_t') \qquad \text{assume } \mathbb{E}(\mathbf{x}_t\mathbf{x}_t') \text{ is nonsingular} \\ \Rightarrow \qquad \boldsymbol{\phi}' &= \mathbb{E}(y_{t+1}\mathbf{x}_t')[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1} \qquad \text{when } \mathbb{E}(\mathbf{x}_t\mathbf{x}_t') \text{ is singular, see Hamilton, p.75, footnote2.} \end{aligned}$$

The MSE/forecast error associated with a linear projection given a sample of T observations on $\mathbf{x}_t$ is given by:

$$\begin{aligned} \mathbb{E}(y_{t+1} - \boldsymbol{\phi}'\mathbf{x}_t)^2 &= \mathbb{E}y_{t+1}^2 - 2\mathbb{E}(\boldsymbol{\phi}'\mathbf{x}_t y_{t+1}) + \mathbb{E}(\boldsymbol{\phi}'\mathbf{x}_t\mathbf{x}_t'\boldsymbol{\phi}) \\ &= \mathbb{E}y_{t+1}^2 - 2\boldsymbol{\phi}'\mathbb{E}(\mathbf{x}_t y_{t+1}) + \boldsymbol{\phi}'\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')\boldsymbol{\phi} \\ &= \mathbb{E}y_{t+1}^2 - 2\mathbb{E}(y_{t+1}\mathbf{x}_t')[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\mathbb{E}(\mathbf{x}_t y_{t+1}) + \mathbb{E}(y_{t+1}\mathbf{x}_t')\{[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')\}[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\mathbb{E}(\mathbf{x}_t y_{t+1}) \\ &= \mathbb{E}y_{t+1}^2 - 2\mathbb{E}(y_{t+1}\mathbf{x}_t')[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\mathbb{E}(\mathbf{x}_t y_{t+1}) + \mathbb{E}(y_{t+1}\mathbf{x}_t')[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\mathbb{E}(\mathbf{x}_t y_{t+1}) \\ &= \mathbb{E}y_{t+1}^2 - \mathbb{E}(y_{t+1}\mathbf{x}_t')[\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\mathbb{E}(\mathbf{x}_t y_{t+1}). \end{aligned}$$

LP is closely related to ordinary least squares (OLS) regression, but still have diffirences between them.

OLS
A linear regression model (given a sample of T observations on y_{t+1}):

$$\left.\begin{aligned} Y_2 &= \phi Y_1 + \epsilon_1 \\ Y_3 &= \phi Y_2 + \epsilon_2 \\ Y_4 &= \phi Y_3 + \epsilon_3 \\ &\vdots \\ Y_T &= \phi Y_{T-1} + \epsilon_{T-1} \\ Y_{T+1} &= \phi Y_T + \epsilon_T \end{aligned}\right\} \xRightarrow{\text{OLS}} \sum_{i=1}^{T} \epsilon_t^2 = \epsilon_1 + \epsilon_2 + \cdots + \epsilon_T = Y_2 - \phi_1 Y_1 + Y_3 - \phi_2 Y_2 + \cdots + Y_{T+1} - \phi_T Y_T = \sum_{t=1}^{T} (Y_{t+1} - \phi Y_t)^2$$

$$\begin{cases} Y_{t+1} = c + \phi_1 Y_t + \phi_2 Y_{t-1} + \cdots + \phi_T Y_{t-T+1} + \epsilon_t & \Leftrightarrow \quad Y_{t+1} = \boldsymbol{\phi}' \mathbf{X}_t + \epsilon_t; \\ Y_{T+1} = c + \phi_1 Y_1 + \phi_2 Y_1 + \cdots + \phi_T Y_T + \epsilon_T & \Leftrightarrow \quad Y_{T+1} = \boldsymbol{\phi}' \mathbf{X}_T + \epsilon_T. \end{cases}$$

Given a sample of T observations on y_{t+1} and $\mathbf{x}_t$ (i.e., 2T observations), the sample sum of squared residuals (SSR)

$$\left.\begin{aligned} Y_2 &= c + \phi_1 Y_1 + \phi_2 Y_0 + \cdots + \phi_{-T+2} Y_{-T+2} + \epsilon_1 = \boldsymbol{\phi}' \mathbf{X}_1 + \epsilon_1 \\ Y_3 &= c + \phi_1 Y_2 + \phi_2 Y_1 + \cdots + \phi_{-T+2} Y_{-T+3} + \epsilon_2 = \boldsymbol{\phi}' \mathbf{X}_2 + \epsilon_2 \\ Y_4 &= c + \phi_1 Y_3 + \phi_2 Y_2 + \cdots + \phi_{-T} Y_{-T+4} + \epsilon_3 = \boldsymbol{\phi}' \mathbf{X}_3 + \epsilon_3 \\ &\vdots \\ Y_T &= c + \phi_1 Y_{T-1} + \phi_2 Y_{T-2} + \cdots + \phi_{-T+2} Y_1 + \epsilon_{T-1} = \boldsymbol{\phi}' \mathbf{X}_{T-1} + \epsilon_{T-1} \\ Y_{T+1} &= c + \phi_1 Y_T + \phi_2 Y_{T-1} + \cdots + \phi_{-T+2} Y_1 + \epsilon_T = \boldsymbol{\phi}' \mathbf{X}_T + \epsilon_T \end{aligned}\right\} \xRightarrow{\text{OLS}} \sum_{i=1}^{T} \epsilon_t^2 = \sum_{t=1}^{T} (Y_{t+1} - \boldsymbol{\phi}' \mathbf{X}_t)^2$$

$$\min \text{MSE} \Rightarrow \begin{cases} \mathbf{0}' = \mathbb{E}[(y_{t+1} - \boldsymbol{\phi}' \mathbf{x}_t) \mathbf{x}_t'] & \Rightarrow \\ \mathbb{E}(y_{t+1} \mathbf{x}_t') = \boldsymbol{\phi}' \mathbb{E}(\mathbf{x}_t \mathbf{x}_t') & \Rightarrow \\ \boldsymbol{\phi}' = \mathbb{E}(y_{t+1} \mathbf{x}_t') [\mathbb{E}(\mathbf{x}_t \mathbf{x}_t')]^{-1}. \end{cases}$$

$$\begin{aligned} \min_{\boldsymbol{\phi}} \text{SSR} &= \sum_{t=1}^{T} (Y_{t+1} - \boldsymbol{\phi}' \mathbf{X}_t)^2 \\ &= \sum_{t=1}^{T} [Y_{t+1}^2 - 2 Y_{t+1} \boldsymbol{\phi}' \mathbf{X}_t + (\boldsymbol{\phi}' \mathbf{X}_t)^2] \\ &= \sum_{t=1}^{T} (Y_{t+1}^2 - 2 Y_{t+1} \boldsymbol{\phi}' \mathbf{X}_t + \boldsymbol{\phi}' \mathbf{X}_t \mathbf{X}_t' \boldsymbol{\phi}) \\ &= \sum_{t=1}^{T} (Y_{t+1}^2 - 2 Y_{t+1} \mathbf{X}_t' \boldsymbol{\phi} + \boldsymbol{\phi}' \mathbf{X}_t \mathbf{X}_t' \boldsymbol{\phi}) \\ \Rightarrow \quad \mathbf{0} &= \sum_{t=1}^{T} (-2 \mathbf{X}_t Y_{t+1} + 2 \mathbf{X}_t \mathbf{X}_t' \hat{\boldsymbol{\phi}}) \\ \Rightarrow \quad \hat{\boldsymbol{\phi}} &= \left[\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{X}_t Y_{t+1} \right] \\ &= \left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \right]^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t Y_{t+1} \right]. \end{aligned}$$

Obviously, $\boldsymbol{\phi}'$ in linear projection is constructed from population moments while it in OLS regression is constructed from sample moments. In other words, OLS regression is a summary of the particular sample observations $\{\mathbf{X}_t\}_{t=0}^{T}$ and $\{Y_t\}_{t=1}^{T+1}$, whereas linear projection is a discription of the population characteristics of the stochastic process $\{\mathbf{x}_t, y_{t+1}\}_{t=-\infty}^{\infty}$. However, there is a formal mathematical sense in which the two operations are the same.

Parallel between OLS and LP (cf. Hamilton, 1994, appendix 4.A, pp.113-114)

$$\begin{cases} \text{OLS} & \rightarrow \quad \text{the particular sample moments}, (X_1, X_2, \ldots, X_T) \text{ and } (Y_2, Y_3, \ldots, Y_{T+1}); \\ \text{LP} & \rightarrow \quad \text{the population moments}, \{\mathbf{x}_t, y_{t+1}\}_{t=-\infty}^{\infty}. \end{cases}$$

Consider an artificial discrete-valued random variable x that can take on only one of sample of size T, each with probability $\frac{1}{T}$:

$$\text{prob}\{x = X_1 = x_1\} = \frac{1}{T},$$
$$\text{prob}\{x = X_2 = x_2\} = \frac{1}{T},$$
$$\vdots$$
$$\text{prob}\{x = X_T = x_T\} = \frac{1}{T}.$$

Denote $\mathbf{x} = (x_1, x_2, \ldots, x_t)'$ as the explanatory vector and $\mathbf{X}_t = (X_1, X_2, \ldots, X_t)'$ as the real value of explanatory vector.

$$\underbrace{\mathbb{E}\mathbf{x}}_{\text{the population mean}} = \mathbb{E}\{x_t\}_{t=1}^{T} = \sum_{t=1}^{T} (X_t \cdot \text{prob}\{x_t = X_t\}) = \underbrace{\frac{1}{T}\sum_{t=1}^{T} X_t}_{\text{the sample mean}},$$

$$\begin{aligned}
\underbrace{\mathbb{E}(\mathbf{x}\mathbf{x}')}_{\text{the population second moment}} &= \left\{ \begin{bmatrix} X_1 \cdot (X_1 \cdot \text{prob}\{x = X_1\}) & X_1 \cdot (X_2 \cdot \text{prob}\{x = X_2\}) & \cdots & X_1 \cdot (X_T \cdot \text{prob}\{x = X_T\}) \\ X_2 \cdot (X_1 \cdot \text{prob}\{x = X_1\}) & X_2 \cdot (X_2 \cdot \text{prob}\{x = X_2\}) & \cdots & X_2 \cdot (X_T \cdot \text{prob}\{x = X_T\}) \\ \vdots & \vdots & \cdots & \vdots \\ X_T \cdot (X_1 \cdot \text{prob}\{x = X_1\}) & X_T \cdot (X_2 \cdot \text{prob}\{x = X_2\}) & \cdots & X_T \cdot (X_T \cdot \text{prob}\{x = X_T\}) \end{bmatrix} + \cdots \right\} \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t\mathbf{X}_t'}_{\text{the sample second moment}} . \\
&= \frac{1}{T}(\mathbf{X}_1\mathbf{X}_1' + \mathbf{X}_2\mathbf{X}_2' + \cdots + \mathbf{X}_T\mathbf{X}_T') \\
&= \frac{1}{T}\left\{ \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots & X_T \end{bmatrix} + \cdots \right\} \\
&= \frac{1}{T}\left\{ \begin{bmatrix} X_1^2 & X_1X_2 & \cdots & X_1X_T \\ X_2X_1 & X_2^2 & \cdots & X_2X_T \\ \vdots & \vdots & \cdots & \vdots \\ X_TX_1 & X_TX_2 & \cdots & X_T^2 \end{bmatrix} + \cdots \right\}
\end{aligned}$$

$$\text{var}x = \mathbb{E}[(x - \mathbb{E}x)^2] = \sum_{t=1}^{T} \left[(X_t - \mathbb{E}X_t)^2 \cdot \text{prob}\{x_t = X_t\}\right] = \frac{1}{T}\sum_{t=1}^{T}(X_t - \mathbb{E}X_t)^2.$$

We can construct a second artificial variable y that can take on one of the discrete values $(y_2, y_3, \ldots, y_{T+1})$. Notice that these are not observations on y.

Posit that the joint distribution of $\mathbf{x}$ and y is given by

$$\text{prob}\{\mathbf{x} = \mathbf{X}_t, y = Y_{t+1}\} = \frac{1}{T}, \quad \text{for } t = 1, 2, \ldots, T.$$

Then

$$\mathbb{E}(\mathbf{x}y) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t Y_{t+1}.$$

The coefficient for a linear projection of y on x is the value of $\boldsymbol{\phi}$ that minimizes

$$\min \text{MSE} \equiv \mathbb{E}(y - \boldsymbol{\phi}'\mathbf{x})^2 = \frac{1}{T}\sum_{t=1}^{T}(Y_{t+1} - \boldsymbol{\phi}'\mathbf{X}_t)^2 = \min \text{SSR}$$

Thus, the formulas for an OLS regression can be viewed as a special case of formulas for a linear projection.

Notice that is the stochastic process $\{\mathbf{x}_t, y_{t+1}\}_{t=-\infty}^{\infty}$ is covariance-stationary and ergodic[1] for second moments, then the sample moments will converge to the population moments as the sample size T goes to infinity, i.e.,

$$\hat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi} \qquad \leftarrow \quad \text{a consistent estimator.}$$

[1]cf. Miao 2014 ch.4, p.115