Econ 2120: Section 2

Part I - Linear Predictor Loose Ends

Ashesh Rambachan

Fall 2018

Outline

Outline

Orthogonality Conditions

Goal of first half of the class: Get you to think in terms of projections.

> Best Linear Predictor: Projection onto space of linear functions of X
>
> Conditional expectation: Projection onto space of all linear functions of X

Key: There is a "one-to-one" map between the projection problem and the orthogonality condition.

Orthogonality Conditions

We can always (trivially) write

$$Y = g(X) + U, \quad U = Y - g(X).$$

To understand what $g(X)$ describes, you **just** need to know the orthogonality condition that U satisfies.

Two cases:

(1): $g(X) = \beta_0 + \beta_1 X$ and $E[U] = E[U \cdot X] = 0$

$\implies$ $g(X)$ is the best linear predictor.

(2): $E[U \cdot h(X)] = 0$ for all $h(\cdot)$.

$\implies$ $g(X) = E[Y|X]$.

Orthogonality Conditions

Why is this useful?

If you can transform your minimization problem e.g.

$$\min_{\beta_0,\beta_1} E[(Y - \beta_0 - \beta_1 X)^2]$$

into a projection problem,

$$\min_{\beta_0,\beta_1} \|Y - \beta_0 - \beta_1 X\|^2,$$

you have a whole new set of (very) useful tools.

By the projection theorem, we can characterize the solution by a set of orthogonality conditions AND we have uniqueness.

Best Linear Predictor

Why are we emphasizing the best linear predictor $E^*[Y|X]$ so much?

Best Linear Predictor

Why are we emphasizing the best linear predictor $E^*[Y|X]$ so much?

In some sense, it can "always be computed."

- Requires very few assumptions aside from the existence of second moments of the joint distribution.
- You can always go to Stata, type `reg y x` and get something that has the interpretation of a best linear predictor.
  - Regression output = estimated coefficients of best linear predictor.

Best Linear Predictor

Why are we emphasizing the best linear predictor $E^*[Y|X]$ so much?

In some sense, it can "always be computed."

- Requires very few assumptions aside from the existence of second moments of the joint distribution.
- You can always go to Stata, type `reg y x` and get something that has the interpretation of a best linear predictor.
  - Regression output = estimated coefficients of best linear predictor.

BUT: It is a completely separate question if this is an "interesting" interpretation.

- To say more, you need to assume more.

Outline

Linear Predictor

$X = (X_1, \ldots, X_K)'$ is a $K \times 1$ vector, $\beta = (\beta_1, \ldots, \beta_K)'$ is the $K \times 1$ vector of coefficients of the best linear predictor. Coefficients characterized by K orthogonality conditions

$$E[(Y - X'\beta)X_j] = E[X_j(Y - X'\beta)] = 0 \quad \text{for} \quad j = 1, \ldots, K.$$

Linear Predictor

$X = (X_1, \ldots, X_K)'$ is a $K \times 1$ vector, $\beta = (\beta_1, \ldots, \beta_K)'$ is the $K \times 1$ vector of coefficients of the best linear predictor. Coefficients characterized by K orthogonality conditions

$$E[(Y - X'\beta)X_j] = E[X_j(Y - X'\beta)] = 0 \quad \text{for} \quad j = 1, \ldots, K.$$

Re-write in vector form to get

$$E[\underset{K\times 1}{X}(Y \underset{1\times 1}{-} X'\beta)] = 0.$$

This is a $K \times 1$ system of linear equations.

Linear Predictor

$X = (X_1, \ldots, X_K)'$ is a $K \times 1$ vector, $\beta = (\beta_1, \ldots, \beta_K)'$ is the $K \times 1$ vector of coefficients of the best linear predictor. Coefficients characterized by K orthogonality conditions

$$E[(Y - X'\beta)X_j] = E[X_j(Y - X'\beta)] = 0 \quad \text{for} \quad j = 1, \ldots, K.$$

Re-write in vector form to get

$$E[\underset{K\times 1}{X}(Y \underset{1\times 1}{-} X'\beta)] = 0.$$

This is a $K \times 1$ system of linear equations.

Provided $E[XX']$ is non-singular,

$$\underset{K\times 1}{E[XY]} - \underset{K\times K}{E[XX']}\underset{K\times 1}{\beta} = 0 \implies \underset{K\times 1}{\beta} = \underset{K\times K}{E[XX']^{-1}}\underset{K\times 1}{E[XY]}.$$

Least Squares

Data are $(y_i, x_{i1}, \ldots, x_{iK})$ for $i = 1, \ldots, n$. $x_i' = (x_{i1}, \ldots, x_{iK})$ is the $1 \times K$ vector of covariates for the i-th observation and $x^j = (x_{1j}, \ldots, x_{nj})'$ be the $n \times 1$ vector of observations of the j-th covariate.

Least Squares

Data are $(y_i, x_{i1}, \ldots, x_{iK})$ for $i = 1, \ldots, n$. $x_i' = (x_{i1}, \ldots, x_{iK})$ is the $1 \times K$ vector of covariates for the i-th observation and $x^j = (x_{1j}, \ldots, x_{nj})'$ be the $n \times 1$ vector of observations of the j-th covariate.

Define

$$\underset{n\times 1}{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{K\times 1}{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}, \quad \underset{n\times K}{x} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x^1 & \ldots & x^K \end{pmatrix}.$$

$n \times K$ matrix x is sometimes referred to as the **design matrix**.

Least Squares

Least-squares coefficients b_j defined by K orthogonality conditions

$$\underset{1\times n}{(x^j)'}(\underset{n\times 1}{y - xb}) = 0 \quad \text{for} \quad j = 1, \ldots, K.$$

Least Squares

Least-squares coefficients b_j defined by K orthogonality conditions

$$\underset{1\times n}{(x^j)'}(\underset{n\times 1}{y - xb}) = 0 \quad \text{for} \quad j = 1, \dots, K.$$

Stack these orthogonality conditions to get

$$\underset{K\times n}{\begin{pmatrix} (x^1)' \\ \vdots \\ (x^K)' \end{pmatrix}}(\underset{n\times 1}{y - xb}) = \underset{K\times n}{x'}(\underset{n\times 1}{y - xb}) = 0$$

Produces $K \times 1$ system of linear equations.

Least Squares

Least-squares coefficients b_j defined by K orthogonality conditions

$$\underset{1\times n}{(x^j)'}(y - \underset{n\times 1}{xb}) = 0 \quad \text{for} \quad j = 1, \ldots, K.$$

Stack these orthogonality conditions to get

$$\underset{K\times n}{\begin{pmatrix} (x^1)' \\ \vdots \\ (x^K)' \end{pmatrix}}(y - \underset{n\times 1}{xb}) = \underset{K\times n}{x'}(y - \underset{n\times 1}{xb}) = 0$$

Produces $K \times 1$ system of linear equations.

Provided $\underset{K\times K}{x'x}$ non-singular,

$$\underset{K\times 1}{b} = \underset{K\times K}{(x'x)^{-1}}\,\underset{K\times 1}{x'y}.$$

Least Squares

Can show (just algebra)

$$x'x = \sum_{i=1}^{n} x_i x_i', \quad x'y = \sum_{i=1}^{n} x_i y_i.$$

Least Squares

Can show (just algebra)

$$x'x = \sum_{i=1}^{n} x_i x_i', \quad x'y = \sum_{i=1}^{n} x_i y_i.$$

Can also write the least-squares coefficients as

$$b = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right).$$

This formula is extremely useful when we discuss asymptotics.

OVB formula (again)

Short linear predictor of Y:

$$E^*[Y|X_1, \ldots, X_{K-1}] = \alpha_1 X_1 + \ldots + \alpha_{K-1} X_{K-1},$$

OVB formula (again)

Short linear predictor of Y:

$$E^*[Y|X_1, \ldots, X_{K-1}] = \alpha_1 X_1 + \ldots + \alpha_{K-1} X_{K-1},$$

Long linear predictor of Y

$$E^*[Y|X_1, \ldots, X_K] = \beta_1 X_1 + \ldots + \beta_K X_K$$

OVB formula (again)

Short linear predictor of Y:

$$E^*[Y|X_1, \ldots, X_{K-1}] = \alpha_1 X_1 + \ldots + \alpha_{K-1} X_{K-1},$$

Long linear predictor of Y

$$E^*[Y|X_1, \ldots, X_K] = \beta_1 X_1 + \ldots + \beta_K X_K$$

Auxiliary linear predictor of X_K

$$E^*[X_K|X_1, \ldots, X_{K-1}] = \gamma_1 X_1 + \ldots + \gamma_{K-1} X_{K-1}.$$

OVB formula (again)

Let

$$\tilde{X}_K = X_K - \gamma_1 X_1 - \ldots - \gamma_{K-1} X_{K-1}.$$

OVB formula (again)

Let

$$\tilde{X}_K = X_K - \gamma_1 X_1 - \ldots - \gamma_{K-1} X_{K-1}.$$

So

$$X_K = \gamma_1 X_1 + \ldots + \gamma_{K-1} X_{K-1} + \tilde{X}_K,$$

$$Y = \tilde{\beta}_1 X_1 + \ldots + \tilde{\beta}_{K-1} X_{K-1} + \beta_K \tilde{X}_K + U$$

where $U = Y - E^*[Y|X_1, \ldots, X_K]$ and $\tilde{\beta}_j = \beta_j + \gamma_j \beta_K$ for $j = 1, \ldots, K-1$.

OVB formula (again)

Note that

$$\langle Y - \tilde{\beta}_1 X_1 - \ldots - \tilde{\beta}_{K-1} X_{K-1}, X_j \rangle = \langle \beta_K \tilde{X}_K + U, X_j \rangle = 0$$

for $j = 1, \ldots, K-1$.

OVB formula (again)

Note that

$$\langle Y - \tilde{\beta}_1 X_1 - \ldots - \tilde{\beta}_{K-1} X_{K-1}, X_j \rangle = \langle \beta_K \tilde{X}_K + U, X_j \rangle = 0$$

for $j = 1, \ldots, K-1$.

These are the same orthogonality conditions of the short linear predictor. So

$$\alpha_j = \beta_j + \gamma_j \beta_K$$

for $j = 1, \ldots, K-1$.

Outline

Residual Regression

Very useful trick that is used all the time.

Consider best linear predictor with K covariates

$$E^*[Y|X_1, \ldots, X_K] = \beta_1 X_1 + \ldots + \beta_K X_K.$$

Focus on β_K. Is there simple closed-form expression for β_K?

Residual Regression

Very useful trick that is used all the time.

Consider best linear predictor with K covariates

$$E^*[Y|X_1, \ldots, X_K] = \beta_1 X_1 + \ldots + \beta_K X_K.$$

Focus on β_K. Is there simple closed-form expression for β_K?

Yes! Use **residual regression**.

Residual Regression

Auxiliary linear predictor of X_K given $X_1, \ldots, X_{K-1}$. Denote this as

$$E^*[X_K | X_1, \ldots, X_{K-1}] = \gamma_1 X_1 + \ldots + \gamma_{K-1} X_{K-1}$$

Associated residual is

$$\tilde{X}_K = X_K - \gamma_1 X_1 - \ldots \gamma_{K-1} X_{K-1}.$$

Residual Regression

Auxiliary linear predictor of X_K given $X_1, \ldots, X_{K-1}$. Denote this as

$$E^*[X_K | X_1, \ldots, X_{K-1}] = \gamma_1 X_1 + \ldots + \gamma_{K-1} X_{K-1}$$

Associated residual is

$$\tilde{X}_K = X_K - \gamma_1 X_1 - \ldots \gamma_{K-1} X_{K-1}.$$

Theorem (Residual Regression)

β_K can be written as

$$\beta_K = \frac{E[Y \tilde{X}_K]}{E[\tilde{X}_K^2]}.$$

Residual Regression: proof

By definition,

$$X_K = \gamma_1 X_1 + \ldots + \gamma_{K-1} X_{K-1} + \tilde{X}_K.$$

Residual Regression: proof

By definition,

$$X_K = \gamma_1 X_1 + \ldots + \gamma_{K-1} X_{K-1} + \tilde{X}_K.$$

Substitute into $E^*[Y|X_1, \ldots, X_k]$ to get

$$E^*[Y|X_1, \ldots, X_K] = \tilde{\beta}_1 X_1 + \ldots + \tilde{\beta}_{K-1} X_{K-1} + \beta_K \tilde{X}_K$$

where

$$\tilde{\beta}_j = \beta_j + \gamma_j \beta_K \quad \text{for} \quad j = 1, \ldots, K-1.$$

Residual Regression: proof

$\tilde{X}_K$ is a linear combination of $X_1, \ldots, X_K$. So, it is orthogonal to the residual $Y - E^*[Y|X_1, \ldots, X_K]$.

$$\Rightarrow \langle Y - \tilde{\beta}_1 X_1 - \ldots - \tilde{\beta}_{K-1} X_{K-1} - \beta_K \tilde{X}_K, \tilde{X}_K \rangle = 0.$$

Residual Regression: proof

$\tilde{X}_K$ is a linear combination of $X_1, \ldots, X_K$. So, it is orthogonal to the residual $Y - E^*[Y|X_1, \ldots, X_K]$.

$$\Rightarrow \langle Y - \tilde{\beta}_1 X_1 - \ldots - \tilde{\beta}_{K-1} X_{K-1} - \beta_K \tilde{X}_K, \tilde{X}_K \rangle = 0.$$

$\tilde{X}_K$ is orthogonal to $X_1, \ldots, X_{K-1}$. Above simplifies to

$$\langle Y - \beta_K \tilde{X}_K, \tilde{X}_K \rangle = 0$$

and so,

$$\beta_K = \frac{E[Y \tilde{X}_K]}{E[\tilde{X}_K^2]}.$$

Residual Regression: intuition

Coefficient β_K on X_K is the coefficient of the best linear predictor of Y given the residuals of X_K.

> If the conditional expectation is linear, β_K is the "partial effect" of X_K on Y holding all else constant.

Exercise 1

Consider the long and short predictors in the population:

$$E^*[Y|1, X_1, X_2, X_3] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$E^*[Y|1, X_1, X_2] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$$

(1) Provide a formula relating α_2 and β_2.

Exercise 1

Consider the long and short predictors in the population:

$$E^*[Y|1, X_1, X_2, X_3] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$E^*[Y|1, X_1, X_2] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$$

(1) Provide a formula relating α_2 and β_2.

(2) Suppose that X_3 is uncorrelated to X_1 and X_2. Does $\alpha_2 = \beta_2$?

Exercise 1

Consider the long and short predictors in the population:

$$E^*[Y|1, X_1, X_2, X_3] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$E^*[Y|1, X_1, X_2] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$$

(1) Provide a formula relating α_2 and β_2.

(2) Suppose that X_3 is uncorrelated to X_1 and X_2. Does $\alpha_2 = \beta_2$?

(3): Suppose that

$$Cov(X_2, X_3) = 0, \quad Cov(X_1, X_3) \neq 0, \quad Cov(X_1, X_2) \neq 0.$$

Does $\alpha_2 = \beta_2$?

Exercise 2

The partial covariance between the random variables Y, X_K given $X_1, \ldots, X_{K-1}$ is

$$Cov^*(Y, X_K | X_1, \ldots, X_{K-1}) = E[\tilde{Y}\tilde{X}_K]$$

where

$$\tilde{Y} = Y - E^*[Y|X_1, \ldots, X_{K-1}]$$
$$\tilde{X}_K = X_K - E^*[X_K|X_1, \ldots, X_{K-1}].$$

The partial variance of Y is

$$V^*(Y|X_1, \ldots, X_{K-1}) = E[\tilde{Y}^2]$$

. The partial correlation between Y, X_K is

$$Cor^*(Y, X_K|X_1, \ldots, X_{K-1}) = \frac{E[\tilde{Y}\tilde{X}_K]}{(E[\tilde{Y}]E[\tilde{X}_K])^{1/2}}.$$

Exercise 2 (continued)

(1): Consider the linear predictor

$$E^*[Y|1, X_1, \ldots, X_K] = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K.$$

Use our residual regression results to show that

$$\beta_K = \frac{Cov^*(Y, X_K|1, X_1, \ldots, X_{K-1})}{V^*(X_K|1, X_1, \ldots, X_{K-1})}.$$

Residual Regression: sample analog

Consider the least-squares fit using K covariates

$$\hat{y}_i = y_i | x_1, \ldots, x_K = b_1 x_{i1} + \ldots + b_K x_{iK}.$$

Following a similar argument, can show that b_K is the coefficient of the least-squares fit of y on $\tilde{x}_k$, the vector of residuals given by

$$\tilde{x}_{iK} = x_{iK} - x_{iK} | x_1, \ldots, x_{K-1}.$$

That is, b_K can be written as

$$b_K = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i \tilde{x}_{iK}}{\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{iK}^2}.$$

Frisch-Waugh-Lovell Theorem

Residual regression is typically known as the **Frisch-Waugh-Lovell Theorem**.

Frisch-Waugh-Lovell Theorem

Residual regression is typically known as the **Frisch-Waugh-Lovell Theorem**.

Interested in regression

$$Y = X_1\beta_1 + X_2\beta_2 + u,$$

where Y is an $n \times 1$ vector, X_1 is an $n \times K_1$ matrix, X_2 is an $n \times K_2$ matrix and u is an $n \times 1$ vector of residuals.

$\Rightarrow$ How can we write the least squares coefficients in β_2?

Frisch-Waugh-Lovell Theorem

Same as the estimate in the modified regression

$$M_{X_1} Y = M_{X_1} X_2 \beta_2 + M_{X_1} u$$

where M_{X_1} is the orthogonal complement of the projection matrix $X_1(X_1'X_1)^{-1}X_1$.

$$M_{X_1} = I - X_1(X_1'X_1)^{-1}X_1'$$

It projects onto the orthogonal complement of the space spanned by the columns of X_1.

Frisch-Waugh-Lovell Theorem

Same as the estimate in the modified regression

$$M_{X_1} Y = M_{X_1} X_2 \beta_2 + M_{X_1} u$$

where M_{X_1} is the orthogonal complement of the projection matrix $X_1(X_1' X_1)^{-1} X_1$.

$$M_{X_1} = I - X_1(X_1' X_1)^{-1} X_1'$$

It projects onto the orthogonal complement of the space spanned by the columns of X_1.

$\Rightarrow$ β_2 can be written as the coefficient in the regression of residuals of Y on residuals of X_2.

Outline

Useful Trick - Orthogonal Decomposition

Notice we used the same trick in OVB and residual regression results.

> Decomposed variable into two pieces - one piece that lives in "simple" space and another that is orthogonal to this "simple" space.

Useful Trick - Orthogonal Decomposition

Example: random variable Y, decompose it into

$$Y = E[Y|X] + \underbrace{(Y - E[Y|X])}_{U}.$$

$E[Y|X]$ lives on the space of functions of X only and U is orthogonal to any function of X.

Useful Trick - Orthogonal Decomposition

Example: random variable Y, decompose it into

$$Y = E[Y|X] + \underbrace{(Y - E[Y|X])}_{U}.$$

$E[Y|X]$ lives on the space of functions of X only and U is orthogonal to any function of X.

Example: random variable Y, decompose it into

$$Y = E * [Y|X_1, \ldots, X_K] + \underbrace{(Y - E^*[Y|X_1, \ldots, X_K])}_{V}.$$

$E^*[Y|X_1, \ldots, X_K]$ lives on the space of linear functions of $X_1, \ldots, X_K$ only and V is orthogonal to any linear function of $X_1, \ldots, X_K$.