Conditional Expectation

Le Wang

Motivation

The use of the entire conditional distribution to characterize the relationship between two (or even more) variables is purely nonparametric and hence model free.

However, what if one or more of the variables are **continuous**?

1. Dependent Variable, Y
2. Independent Variable, X

Motivation

For continuous **dependent** variables

Instead, we can focus on parts of the distribution

1. **Moments**: $\mathbb{E}[Y|X]$ (Conditional Expectation (Mean))
2. **Quantiles**: $Q_{Y|X}(y|x)$ is defined as usual, but for the conditional distribution (Conditional Quantiles)

$$\inf\{y : F_{Y|X=x}(y) \geq \tau\}$$

In other words, we will obtain the mean or quantiles of Y for the subpopulation when $X = x$. Business as usual.

Definition

Material:

1. Greene, Appendix B.8.
2. Hansen (2018), Chapter 2.
3. Wooldridge, Chapter 2.

Definition

Conditional Expectation is defined as follows

1. **Continuous Case**

$$\mathbb{E}[Y|X] = \int y f_{Y|X}(y|X) dy$$

2. **Discrete Case**

$$\mathbb{E}[Y|X] = \sum y_i p(y_i|X)$$

Properties of Conditional Expectation

Contrary to the **unconditional expectation** $\mathbb{Y}$, the conditional expectation is a **random variable** (since $f_{Y|X}(y|X)$ is a random function)!

Properties of Conditional Expectation

Contrary to the **unconditional expectation** $\mathbb{Y}$, the conditional expectation is a **random variable** (since $f_{Y|X}(y|X)$ is a random function)!

We can think of Y as a function of X:

$$\mathbb{E}[Y|X] = m(X) \quad \text{for some function } m(\cdot)$$

where $m(\cdot)$ is a function determined by the joint (and hence conditional) distribution of (X, Y).

Our goal: To figure out what is $m(X)$!

Note that for the sake of simplicity, we denote the predictor(s) by X, but it can be of any dimension.

$$\mathbb{E}[Y \mid X_1]$$
$$\mathbb{E}[Y \mid X_1, X_2]$$
$$\mathbb{E}[Y \mid X_1, X_2, X_3]$$

And they represent different quantities!

Use of Conditional Expectation

1. **Prediction**: We know that mean is the best prediction based on certain criterion. This result carries through here as well.

Later, we will also discuss the prediction property of the CEF.

Use of Conditional Expectation

1. **Prediction**: We know that mean is the best prediction based on certain criterion. This result carries through here as well.

Later, we will also discuss the prediction property of the CEF.

2. **Marginal Impact**: For example, in the one-dimension case of X

$$\mathbb{E}[Y|X=1]-\mathbb{E}[Y|X=0]$$

Or, if X is continuous, then the marginal effects of X is given by

$$\frac{d}{dX}m(X)$$

When we have more Xs, the interpretation of partial effects are conditioned on the variables that we control for in our model (**holding everything else constant**). Whatever is conditioned on or controlled for is held constant in our estimations.

$$\frac{\partial}{\partial X_1}\mathbb{E}[Y \mid X_1]$$
$$\frac{\partial}{\partial X_1}\mathbb{E}[Y \mid X_1, X_2, X_3]$$
$$\frac{\partial}{\partial X_1}\mathbb{E}[Y \mid X_1, X_2, X_3, \ldots]$$

Special Cases:

1. **Bivariate normal**
2. **Truncated Normal**

Special Conditional Distribution (I)

Greene, Appendix B. on Multivariate Normal Distribution

Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \right)$$

where Σ_{XX} is positive definite. Then,

Special Conditional Distribution (I)

Greene, Appendix B. on Multivariate Normal Distribution

Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \right)$$

where Σ_{XX} is positive definite. Then,

$$Y|X \sim \mathcal{N}(\mu_{Y|X}, \Sigma_{Y|X})$$

Special Conditional Distribution (I)

Greene, Appendix B. on Multivariate Normal Distribution

Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\right)$$

where Σ_{XX} is positive definite. Then,

$$Y|X \sim \mathcal{N}(\mu_{Y|X}, \Sigma_{Y|X})$$

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

Special Conditional Distribution (I)

Greene, Appendix B. on Multivariate Normal Distribution

Let

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\right)$$

where Σ_{XX} is positive definite. Then,

$$Y|X \sim \mathcal{N}(\mu_{Y|X}, \Sigma_{Y|X})$$

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

$$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

Special conditional Distributions (I)

$$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

Note that the conditional variance does not depend on X, while the conditional mean does!

Special conditional Distributions (I)

Consider a bivariate normal distribution or any cases where Y is a scalar. Below is one assumption regarding the error term in the sample selection models:

$$\begin{pmatrix} \epsilon_{\mathbf{1}} \\ \epsilon_{\mathbf{2}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

From our results above

$$\mathbb{E}[Y \mid X] = \mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

Special conditional Distributions (I)

Consider a bivariate normal distribution or any cases where Y is a scalar. Below is one assumption regarding the error term in the sample selection models:

$$\begin{pmatrix} \epsilon_\mathbf{1} \\ \epsilon_\mathbf{2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

From our results above

$$\begin{aligned} \mathbb{E}[Y \mid X] &= \mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) \\ \mathbb{E}[\epsilon_2 \mid \epsilon_1] &= \mu_{\epsilon_2} + \Sigma_{\epsilon_2\epsilon_1}\Sigma_{\epsilon_1}^{-1}(\epsilon_1 - \mu_{\epsilon_1}) \end{aligned}$$

Special conditional Distributions (I)

Consider a bivariate normal distribution or any cases where Y is a scalar. Below is one assumption regarding the error term in the sample selection models:

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

From our results above

$$\begin{aligned} \mathbb{E}[Y \mid X] &= \mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) \\ \mathbb{E}[\epsilon_2 \mid \epsilon_1] &= \mu_{\epsilon_2} + \Sigma_{\epsilon_2\epsilon_1}\Sigma_{\epsilon_1}^{-1}(\epsilon_1 - \mu_{\epsilon_1}) \\ &= 0 + \frac{\sigma_{12}}{\sigma_1}(\epsilon_1 - 0) \end{aligned}$$

Special conditional Distributions (I)

Consider a bivariate normal distribution or any cases where Y is a scalar. Below is one assumption regarding the error term in the sample selection models:

$$\begin{pmatrix} \epsilon_{\mathbf{1}} \\ \epsilon_{\mathbf{2}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$$

From our results above

$$\begin{aligned} \mathbb{E}[Y \mid X] &= \mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) \\ \mathbb{E}[\epsilon_2 \mid \epsilon_1] &= \mu_{\epsilon_2} + \Sigma_{\epsilon_2\epsilon_1}\Sigma_{\epsilon_1}^{-1}(\epsilon_1 - \mu_{\epsilon_1}) \\ &= 0 + \frac{\sigma_{12}}{\sigma_1}(\epsilon_1 - 0) \\ &= \sigma_{12}\epsilon_1 \end{aligned}$$

We will be using the result above in, e.g., **Sample Selection Models**. When we introduce actual models with error terms, this result can be immediately applied.

A special truncated distribution: **Truncated Normal**

Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then the following results hold

$$\Pr[Y > L] = 1 - \Phi(\frac{L - \mu}{\sigma}) = 1 - \Phi(\alpha)$$

A special truncated distribution: **Truncated Normal**

Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then the following results hold

$$\begin{aligned} \Pr[Y > L] &= 1 - \Phi(\frac{L-\mu}{\sigma}) = 1 - \Phi(\alpha) \\ f(y|Y > L) &= \frac{f(y)}{1 - \Phi(\alpha)} \\ &= \frac{\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})}{1 - \Phi(\alpha)} \end{aligned}$$

where $\alpha = \frac{L-\mu}{\sigma}$, $\phi(\cdot)$ ($\Phi(\cdot)$) is the **standard** normal density (distribution) function.

Theorem Moments of the Truncated Normal Distribution:

$$\mu^L \equiv \mathbb{E}[Y \mid Y > L] = \int_L^\infty yf(y \mid Y > L)dy$$
$$Var[Y \mid Y > L] = \int_L^\infty (y - \mu^L)^2 f(y \mid Y > L)dy$$

Theorem Moments of the Truncated Normal Distribution

If $Y \sim \mathcal{N}(\mu, \sigma^2)$ and L is a constant, then

$$\mathbb{E}[Y|\text{truncated sample}] = \mu + \sigma\lambda(\alpha)$$
$$Var[Y|\text{truncated sample}] = \sigma^2(1 - \sigma(\alpha))$$

Theorem Moments of the Truncated Normal Distribution

If $Y \sim \mathcal{N}(\mu, \sigma^2)$ and L is a constant, then

$$\mathbb{E}[Y|\text{truncated sample}] = \mu + \sigma\lambda(\alpha)$$
$$Var[Y|\text{truncated sample}] = \sigma^2(1 - \sigma(\alpha))$$

where $\alpha = \frac{L-\mu}{\sigma}$, $\phi(\cdot)$ $(\Phi(\cdot))$ is the **standard** normal density (distribution) function and

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1-\Phi(\alpha)} \quad \text{if truncated sample is defined as } Y > L$$
$$\lambda(\alpha) = -\frac{\phi(\alpha)}{\Phi(\alpha)} \quad \text{if truncated sample is defined as } Y < L$$

and

$$\sigma(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$$

Note that

$$0 < \sigma(\alpha) < 1$$

1. CEF and Probability (Conditional PDF and CDF)
2. CEF and (Unconditional) Mean Independence
3. CEF and (Conditional) Mean Independence
4. Use of CEF
5. Law of Iterated Expectations (Simple and General)
6. Conditioning Theorem
7. Decomposition Property
8. Prediction Property
9. ANOVA Theorem

CEF and Probability (I): Conditional PDF

$$\mathbb{E}[Y|X] = \sum y_i p(y_i|X)$$

CEF and Probability (I): Conditional PDF

$$\mathbb{E}[Y|X] = \sum y_i p(y_i|X)$$

Suppose that $D = \{0, 1\}$, we can show that

$$\mathbb{E}[D \mid X] = \Pr[D = 1 \mid X] \equiv p(X)$$

$$\mathbb{E}[Y|X] = \sum y_i p(y_i|X)$$

$$\begin{aligned}\mathbb{E}[D \mid X] &= 1 \cdot \Pr[D = 1 \mid X] + 0 \cdot \Pr[D = 0 \mid X] \\ &= \Pr[D = 1 \mid X] \equiv p(X)\end{aligned}$$

CEF and Probability (II): Conditional CDF

Similarly, we can show that

$$\mathbb{E}[\mathbb{I}[Y \leq y] \mid X] = \Pr[Y \leq y \mid X] = F_{Y|X}(y \mid x)$$

This will be left as **homework**. This expression indeed opens the door for recent advances in **distributional regression**.

CEF and Unconditional (mean) Independence

$$Y \perp X$$

implies

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$$

Intuitively if the distribution of Y does not vary with X, the mean does not, either! The proof will be left as **homework**.

A typical example of **mean independence** later:

$$\mathbb{E}[\epsilon] = \mathbb{E}[\epsilon \mid X] = 0$$

CEF and Conditional (mean) Independence

Note that

$$Y \perp D \mid X$$

implies

$$\mathbb{E}[Y \mid D, X] = \mathbb{E}[Y \mid X]$$

Or,

CEF and Conditional (mean) Independence

Note that

$$Y \perp D \mid X$$

implies

$$\mathbb{E}[Y \mid D, X] = \mathbb{E}[Y \mid X]$$

Or,

$$\mathbb{E}[D \mid Y, X] = \mathbb{E}[D \mid X]$$

You can easily verify these results using the definitions of **conditional expectation** and **conditional independence** from previous slides. **Homework**

In the potential outcome framework, we will consider

$$(Y(1), Y(0)) \perp D \mid X$$

similarly implies

$$\mathbb{E}[D \mid Y(1), Y(0), X] = \mathbb{E}[D \mid X]$$

Mean Independence

Note that mean independence is a weaker independence than full independence, but stronger than correlation. In other words,

Mean independence implies zero correlation, but not the other way around.

In your homework, you will be asked to show a special case of this holds.

An Interesting Application in IV Estimation

An external IV is often difficult to find and controversial. However, it is sometimes possible to consider an internal IV constructed simply from the variables available to us. Consider the following model

$$y = D \cdot \gamma + x'\beta + \epsilon$$

where ϵ is an error such that $\mathbb{E}[\epsilon|x] =$ and D is an endogenous binary variable and related to x through $h(x)$.

An Interesting Application in IV Estimation

Without other identification assumptions, this model cannot be consistently estimated. However, we can if an IV z satisfies the following conditions is available:

1. $\mathbb{E}[\epsilon \mid z] = 0$
2. $Cov(z, D) \neq 0$

$z = \mathbb{E}[D \mid X] = p(x)$ is a natural candidate!

Law of Iterated Expecations

1. **Simple Law of Iterated Expectations**
2. **Law of Iterated Expectations**

LIE

Simple Law of Iterated Expectations

If $\mathbb{E}|y| < \infty$, then for any random vector x,

$$\mathbb{E}[\mathbb{E}[y|x]] = \mathbb{E}[y]$$

LIE

Simple Law of Iterated Expectations

If $\mathbb{E}|y| < \infty$, then for any random vector x,

$$\mathbb{E}[\mathbb{E}[y|x]] = \mathbb{E}[y]$$

Note that when x is discrete

$$\mathbb{E}[\mathbb{E}[y|x]] = \sum_{j=1}^{\infty} \mathbb{E}[y|x = x_j] \Pr[x = x_j]$$

LIE

Simple Law of Iterated Expectations

If $\mathbb{E}|y| < \infty$, then for any random vector x,

$$\mathbb{E}[\mathbb{E}[y|x]] = \mathbb{E}[y]$$

Note that when x is discrete

$$\mathbb{E}[\mathbb{E}[y|x]] = \sum_{j=1}^{\infty} \mathbb{E}[y|x = x_j] \Pr[x = x_j]$$

When x is continuous

$$\mathbb{E}[\mathbb{E}[y|x]] = \int \mathbb{E}[y|x = x] f(x) dx$$

Example

$$\mathbb{E}[\mathbb{E}[y|x]] = \sum_{j=1}^{\infty} \mathbb{E}[y|x = x_j] \Pr[x = x_j]$$

$$\begin{aligned} \mathbb{E}[\log(wages)] &= \mathbb{E}[\log(wages)|sex = man] \Pr[sex = man] \\ &+ \mathbb{E}[\log(wages)|sex = woman] \Pr[sex = womman] \end{aligned}$$

You can think of this as **Law of Total Expectation**.

Simple Law of Iterated Expectations

Proof:

$$\mathbb{E}[\mathbb{E}[y|x]] = \int \mathbb{E}[y|x] f(x) dx \quad \text{Note that } \mathbb{E}[y|x] \text{is a function of } x$$

Simple Law of Iterated Expectations

Proof:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[y|x]] &= \int \mathbb{E}[y|x] f(x) dx \quad \text{Note that } \mathbb{E}[y|x] \text{is a function of } x \\ &= \int \int y f(y|x) dy f(x) dx \end{aligned}$$

Simple Law of Iterated Expectations

Proof:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[y|x]] &= \int \mathbb{E}[y|x]f(x)dx \quad \text{Note that } \mathbb{E}[y|x] \text{is a function of } x \\
&= \int \int yf(y|x)dyf(x)dx \\
&= \int \int yf(x,y)dydx
\end{aligned}$$

Simple Law of Iterated Expectations

Proof:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[y|x]] &= \int \mathbb{E}[y|x] f(x) dx \quad \text{Note that } \mathbb{E}[y|x] \text{is a function of } x \\
&= \int \int y f(y|x) dy f(x) dx \\
&= \int \int y f(x, y) dy dx \\
&= \int y \left\{ \int f(x, y) dx \right\} dy
\end{aligned}
$$

Simple Law of Iterated Expectations

Proof:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[y|x]] &= \int \mathbb{E}[y|x]f(x)dx \quad \text{Note that } \mathbb{E}[y|x] \text{is a function of } x \\
&= \int \int yf(y|x)dyf(x)dx \\
&= \int \int yf(x,y)dydx \\
&= \int y \left\{ \int f(x,y)dx \right\} dy \\
&= \int yf(y)dy
\end{aligned}
$$

Simple Law of Iterated Expectations

Proof:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[y|x]] &= \int \mathbb{E}[y|x]f(x)dx \quad \text{Note that } \mathbb{E}[y|x] \text{is a function of } x \\
&= \int \int yf(y|x)dyf(x)dx \\
&= \int \int yf(x,y)dydx \\
&= \int y \left\{ \int f(x,y)dx \right\} dy \\
&= \int yf(y)dy \\
&= \mathbb{E}[y]
\end{aligned}
$$

LIE (General Case)

Law of Iterated Expectation If $\mathbb{E}|y| < \infty$ then for any random vectors x_1, x_2

$$\mathbb{E}[\mathbb{E}[y|x_1, x_2]|x_1] = \mathbb{E}[y|x_1]$$

Example

$$\begin{aligned} &\mathbb{E}[\log(wages)|sex = man] \\ = &\ \mathbb{E}[\log(wages)|sex = man, race = white]\Pr[race = white|sex = man] \\ &+ \mathbb{E}[\log(wages)|sex = man, race = black]\Pr[race = black|sex = man] \\ &+ \mathbb{E}[\log(wages)|sex = man, race = others]\Pr[race = others|sex = man] \end{aligned}$$

LIE (even more general)

Law of Iterated Expectation If $\mathbb{E}|y| < \infty$ then for a random vector, $\mathbf{w} = \mathbf{f}(\mathbf{x})$ (any random vector $\mathbf{x}$)

$$\mathbb{E}[\mathbb{E}[y|\mathbf{x}]|\mathbf{w}] = \mathbb{E}[y|\mathbf{w}]$$

Application: Identification results for propensity scores, Partial Linear Model or Semiparametric Model (Wooldridge, 2e, p29, Exercise 2.7)

A special case with propensity scores $D = 0, 1$ and $\mathbb{E}[D \mid X] = \Pr[D = 1 \mid X] \equiv p(X)$

$$\mathbb{E}[D \mid p(X)] = \mathbb{E}[\mathbb{E}[D \mid X, p(X)] \mid p(X)]$$

A special case with propensity scores $D = 0, 1$ and $\mathbb{E}[D \mid X] = \Pr[D = 1 \mid X] \equiv p(X)$

$$\begin{aligned} \mathbb{E}[D \mid p(X)] &= \mathbb{E}[\mathbb{E}[D \mid X, p(X)] \mid p(X)] \\ &= \mathbb{E}[\mathbb{E}[D \mid X] \mid p(X)] \end{aligned}$$

A special case with propensity scores $D = 0, 1$ and $\mathbb{E}[D \mid X] = \Pr[D = 1 \mid X] \equiv p(X)$

$$\begin{aligned}
\mathbb{E}[D \mid p(X)] &= \mathbb{E}[\mathbb{E}[D \mid X, p(X)] \mid p(X)] \\
&= \mathbb{E}[\mathbb{E}[D \mid X] \mid p(X)] \\
&= \mathbb{E}[p(X) \mid p(X)]
\end{aligned}$$

A special case with propensity scores $D = 0, 1$ and $\mathbb{E}[D \mid X] = \Pr[D = 1 \mid X] \equiv p(X)$

$$
\begin{aligned}
\mathbb{E}[D \mid p(X)] &= \mathbb{E}[\mathbb{E}[D \mid X, p(X)] \mid p(X)] \\
&= \mathbb{E}[\mathbb{E}[D \mid X] \mid p(X)] \\
&= \mathbb{E}[p(X) \mid p(X)] \\
&= p(X)
\end{aligned}
$$

Properties of Conditional Expectation

A property of conditional expectations is that when you condition on a random vector x you can effectively treat it as if it is constant.

$$\mathbb{E}[x|x] = x$$

and

$$\mathbb{E}[g(x)|x] = g(x)$$

This is known as **conditioning theorem**

Conditioning Theorem

If $\mathbb{E}|y| < \infty$ and $|g(x)y| < \infty$, then

$$\mathbb{E}[g(x)y|x] = g(x)\mathbb{E}(y|x)$$

and

$$\mathbb{E}[g(x)y] = \mathbb{E}[g(x)\mathbb{E}(y|x)]$$

Conditioning Set = Information Set

Three More Properties of Conditional Expectation

Why is the conditional expectation so important?

1. Decomposition Property **A powerful result of LIE**
2. Prediction Property
3. ANOVA Theorem

Decomposition Property

Decomposition Property any random variable y can be expressed as

$$y = \mathbb{E}[y|x] + \epsilon$$

where ϵ is a random varaible satisfying the following conditions

1. $\mathbb{E}[\epsilon|x] = 0$, which implies $\mathbb{E}[\epsilon] = 0$
2. $\mathbb{E}[h(x)\epsilon] = 0$, where $h(\cdot)$ is any function of x!

Decomposition Property

Intuition: any variable can be decomposed in two parts:

1. the conditional expectation and
2. an orthogonoal 'error' term. We are not claiming $\mathbb{E}[y|x]$ is linear.

In other words,

1. A part that can be "explained by X"
2. A part that has nothing to do with x

Orthogonal CEF Error

$\mathbb{E}[y|x]$ is called conditional expectation function (CEF), and ϵ CEF error.

All the properties of the CEF error are **NOT** assumptions, but the implications of the Law of Iterated Expectations

Orthogonal CEF Error

Proof

$$\mathbb{E}[\epsilon|x] = \mathbb{E}\left[y - \mathbb{E}[y|x]|x\right]$$

Orthogonal CEF Error

Proof

$$\begin{aligned}\mathbb{E}[\epsilon|x] &= \mathbb{E}\left[y - \mathbb{E}[y|x]|x\right] \\ &= \mathbb{E}[y|x] - \mathbb{E}[\mathbb{E}[y|x]|x]\end{aligned}$$

Orthogonal CEF Error

Proof

$$\begin{aligned}\mathbb{E}[\epsilon|x] &= \mathbb{E}\left[y - \mathbb{E}[y|x]|x\right] \\ &= \mathbb{E}[y|x] - \mathbb{E}[\mathbb{E}[y|x]|x] \\ &= \mathbb{E}[y|x] - \mathbb{E}[y|x]\end{aligned}$$

Orthogonal CEF Error

Proof

$$\begin{aligned}
\mathbb{E}[\epsilon|x] &= \mathbb{E}\left[y - \mathbb{E}[y|x]|x\right] \\
&= \mathbb{E}[y|x] - \mathbb{E}[\mathbb{E}[y|x]|x] \\
&= \mathbb{E}[y|x] - \mathbb{E}[y|x] \\
&= 0
\end{aligned}$$

$$\mathbb{E}[\epsilon] = \mathbb{E}[\mathbb{E}[\epsilon|x]] = \mathbb{E}[0] = 0$$

The last property will be left as an exercise in your homework.

Orthogonal CER Error

The fact that ϵ is uncorrelated with any functions of x esnures that we have fully accounted for the effects of x_1 and x_2 on **the expected value** of y.

Or,

We have the function form of $\mathbb{E}[y|x]$ properly specified.

Orthogonal CEF Error

$\mathbb{E}[\epsilon|x] = \mathbb{E}[\epsilon] = 0$

This implies that ϵ is **mean** independent of x.

Mean Independence if the conditional mean of two variables y, x satisfies the following condition,

$$\mathbb{E}[y|x] = \mathbb{E}[y]$$

then y is mean independent of x.

Prediction Property

The CEF is a good summary of the relationship between x and y for a number of reasons.

Just like unconditional mean, conditional mean provides a representative value (or **"best" prediction**) for a random variable, y, for a sub-population with characteristics, x.

This property is called **prediction property**

Prediction Property

Let $m(x)$ be any function of x. Then

$$\mathbb{E}[y|x] = \underset{m(x)}{\arg\min} \mathbb{E}[(y - m(x))^2]$$

Intuition: the conditional expectation is the best prediction in the sense that it minimizes mean squared error.

Prediction Property (Proof)

Proof:

$$\mathbb{E}[(y - m(x))^2]$$

Prediction Property (Proof)

Proof:

$$\mathbb{E}[(y - m(x))^2] = \mathbb{E}[(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - m(x))^2]$$

Prediction Property (Proof)

Proof:

$$
\begin{aligned}
\mathbb{E}[(y - m(x))^2] &= \mathbb{E}[(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - m(x))^2] \\
&= \mathbb{E}[\underbrace{(y - \mathbb{E}[y|x])^2}_{(1)} + \underbrace{(\mathbb{E}[y|x] - m(x))^2}_{(2)}
\end{aligned}
$$

Prediction Property (Proof)

Proof:

$$\begin{aligned}\mathbb{E}[(y - m(x))^2] &= \mathbb{E}[(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - m(x))^2] \\ &= \mathbb{E}[\underbrace{(y - \mathbb{E}[y|x])^2}_{(1)} + \underbrace{(\mathbb{E}[y|x] - m(x))^2}_{(2)} \\ &\quad + \underbrace{2(y - \mathbb{E}[y|x])(\mathbb{E}(y|x) - m(x))}_{(3)}]\end{aligned}$$

Let's examine these terms

1. Term (1) is fixed and not a choice
2. Term (3) is always equal to zero by the **decomposition property**

In other words,

$$\underset{m(x)}{\arg\min}\mathbb{E}[(y - m(x))^2] = \underset{m(x)}{\arg\min}\mathbb{E}[(\mathbb{E}[y|x] - m(x))^2]$$

Prediction Property

Question:

Why $\mathbb{E}[2(y - \mathbb{E}[y|x])(\mathbb{E}(y|x) - m(x))] = 0$?

Hint: This is a result of the decomposition property

Prediction Property

Trick:

$$
\begin{aligned}
&\mathbb{E}[2 \cdot (y - \mathbb{E}[y|x])(\mathbb{E}(y|x) - m(x))] \\
&= \mathbb{E}[2 \cdot \epsilon \cdot (\mathbb{E}(y|x) - m(x))] \\
&= 2 \cdot \mathbb{E}[\epsilon \cdot h(x)] \\
&= 0
\end{aligned}
$$

where $h(x) \equiv \mathbb{E}(y|x) - m(x))$

We will use this trick again in your homework to prove the following theorem.

ANOVA Theorem

Conditional Variance is defined as follows

$$Var(y|x) = \mathbb{E}[(y - \mathbb{E}[y \mid x])^2 \mid x]$$

ANOVA Theorem

A final property of the CEF, closely related to both the decomposition and prediction properties, is the analysis of variance (ANOVA) theorem.

$$Var(y) = Var(\mathbb{E}[y|x]) + \mathbb{E}[Var(y|x)]$$

where $Var(y|x)$ is conditional variance of y.

Intuition: The variance of y decomposes into

1. the variance of the conditional mean function
2. the variance of the error term (i.e., the expected variance around the conditional mean)

ANOVA Theorem

You will see such analysis a lot, e.g., from your regression output.

ANOVA Theorem: Application

Research on inequality, people decompose changes in the income distribution into parts that can be accounted for by changes in worker characteristics and changes in what's left over after accounting for these factors (see, e.g., Autor et al. 2005)

Regression Variance

An important measure of the dispersion about the CEF is the unconditional variance of the CEF error ϵ

$$\sigma^2 = var(\epsilon) = \mathbb{E}[(\epsilon - \mathbb{E}[\epsilon])^2] = \mathbb{E}[\epsilon^2]$$

Intuition: The amount of variation in y which is not "explained" or accounted for in the conditional mean $\mathbb{E}[y \mid x]$

Regression Variance

A simple but useful result:

$$\text{If } \mathbb{E}[y^2] < \infty, \text{then } \sigma^2 < \infty$$

Regression Variance

What will happen to regression variance or the variation in y that is not explained by the CEF when we condition on more x?

$$y = \mathbb{E}[y \mid x_1] + \epsilon_1$$
$$y = \mathbb{E}[y \mid x_1, x_2] + \epsilon_2$$

In other words, what is the relationship between

$$Var(y), Var(\epsilon_1), Var(\epsilon_2)$$

Regression Variance

Theorem

If $\mathbb{E}[y^2] < \infty$, then

$$Var(y) \geq Var(\epsilon_1) \geq Var(\epsilon_2)$$

Intuition the variance of the unexplained portion of y decreases (weakly) whenver we include more to the conditioning set.