Boston University Questrom School of Business
MF793 – Fall 2021

Eric Jacquier

Optimal Conditional Prediction

- *This is the exact equivalent to what we learnt on the mean, …*

  "The mean is optimal with respect to MSE"

  … now we do the conditional mean
- We predict Y|X using its **Conditional Expectation Function**, E(Y|X), the **CEF**.
- We derive the fundamental properties of the CEF $\boldsymbol{E(Y|X) \equiv m_x}$, a possibly complicated function of X
- There is **no estimation** in this lecture note

A. Mean of the conditional prediction error
B. Variance of the conditional prediction error
C. CEF is the best (the same way mean was best)
D. So how does this take us to OLS, Ordinary Least Squares?

Readings: if desired, Hansen Chapter 2 $\rightarrow$ 2.11 included, not needed

A. CEF error The CEF $m_x \equiv E(Y|X)$ is unbiased, … by definition!

- Predicting Y|X with E(Y|X), we make a **random prediction error**: $\varepsilon = Y - E(Y|X)$
- Conditional mean of the error **E(ε|X)**

$$\begin{aligned} E(\varepsilon|X) &= E[(Y - m_X) \mid X] \\ &= E[Y|X] - E[m_X \mid X] \\ &= m_X - m_X = 0 \end{aligned}$$

- Unconditional mean of the error $E(\varepsilon) = E_X(\, E(\varepsilon|X)\,) = E_X(0) = 0$ by Iter. Exp.

- Covariance between prediction error and <u>any function of</u> X:

$$Cov(g(X), \varepsilon) = E(g(X)\,\varepsilon) - \;?\; E(g(x))\,E(\varepsilon) = 0$$

$$= E_X[\, E(g(X)\,\varepsilon \mid X)\,] = E_X[\, g(X)\; E(\varepsilon \mid X)\,] = 0$$

by IE

The CEF error is uncorrelated <u>with any function g of x, this includes X and m_X</u>

$$\boxed{E(\varepsilon|x) = 0 \qquad E(\varepsilon) = 0 \qquad Cov[E(y|x), \varepsilon] = 0}$$

B. Variance of the CEF error: $y = E(y|x) + \varepsilon = m_x + \varepsilon$

- Conditional variance of the y | X is the conditional variance of the prediction error ε

  So: **Var(ε | X)** = Var[(y - E(y|X)) | X] **= Var(y|x) + ?** $\cancel{?}$ Obviously, it's like a definition!

  Var(ε|x) can be a complicated function of x (heteroskedasticity, non linearity, etc)

- Unconditional variance of the prediction error: var(ε) = var(y-E(y|x))

  $$\text{Var}(y) = \text{Var}(E(y|x) + \varepsilon) = \text{Var}(E(y|x)) + \text{Var}(\varepsilon) + 2\,\text{Cov}(E(y|x), \varepsilon)$$

  $$= \text{Var}(E(y|x)) + \text{Var}(\varepsilon)$$

  `

  $$\geq \text{Var}(\varepsilon)$$
  $$\geq \text{Var}(y - E(y|x))$$ Draw a picture … worth a thousand words !

var(y) ≥ var(y-E(y|x)) = var(ε)
The variance of y around its conditional mean
is always smaller than the total variance of y (i.e., around its unconditional mean)

- Corollary: can show that $\text{var}(y-E(y|X_1)) \geq \text{var}(y-E(y|X_1,X_2))$ No proof

Increasing information always decreases the
variance of the unexplained portion of a random variable y

C. E(Y|X) is the best Predictor of y conditional on x

- We call **best predictor** a predictor that minimizes the MSE of prediction

- Remember: The mean E(Y) minimizes MSE of prediction of Y: $\mu = \text{Argmin } E(y-\theta)^2$

  We now prove the conditional version: we condition y on a given value of x,

- What predictor function g(x) minimizes the MSE of conditional prediction: $E(y - g(x) \mid x)^2$?

  0

$$2(m_x - g(x))\, E(y - m_x \mid x)$$

$$E([y - m_x + m_x - g(x)] \mid x)^2 = E((y - m_x) \mid x)^2 + E(m_x - g(x) \mid x)^2 + 2\, E[(m_x - g(x))(y - m_x) \mid x]$$

$$= \quad 0 \quad + (m_x - g(x))^2 \quad + \quad 0 \qquad ?$$

Conditional mean E(y|x) is the best predictor of y | x,
it minimizes the MSE of prediction

- There was **nothing** about estimation in these pages.
  It is all about the properties of the true, possibly unknown conditional mean of y, $E(y|X) \equiv m_x$.

D. Justification for OLS

What does this have to do with OLS ? **Ordinary Least Squares**

….. least squares of **what?**

1. m_X is the best predictor of $y|X$ ... but we don't know the true m_X .

2. Say we approximate m_x by a linear model (linear in X) like **$E(Y|X) = \beta X$** [1]

3. OK, but then how do we choose β?

We must choose β to makes the approximation as good as possible

… that minimizes MSE of prediction of the (surely wrong !) linear model.

Find β to minimize approximate $MSE = E(y-X\beta)^2$

$$= E(y^2) + \beta^2 E(X^2) - 2\beta E(X y)$$

$$\partial MSE/\partial\beta = 2\beta E(X^2) - 2 E(Xy) = 0 \text{ at the optimum} \quad [2]$$

$$\beta = \frac{E(Xy)}{E(X^2)} \quad [3]$$

Conditions needed: $E(X^2)<\infty$, $E(Xy)<\infty$

- Result: The linear model with β chosen as in [2] satisfies $E(x\varepsilon) = 0$

$$E(X\varepsilon) = E\big(X\,(y - X\beta)\big) = E(Xy) - \beta E(X^2) = \qquad\qquad = 0$$

The β in [3] that minimizes the MSE results in a prediction error uncorrelated with X

- β X with β as in [3] is the **best linear** conditional prediction of y | X

Careful: without further assumption, nothing says that $y=x\beta$ is the best model.

- Econometric models, and regressions in particular, are attempts to characterize the unknown CEF of a dependent variable (y) which we want to predict by (x).

The models may or may not be good attempts !

All models are wrong, some are useful (George Box)

- We have already seen that a possible justification for an exact linear model: the joint normality of x and y. Then the linear model on P.5 would be correct

E. Some models ..

$X\beta$

- **Mean:** $y = \mu + \varepsilon$ $E(\varepsilon) = 0$ mean plus noise

  Predict y using the unconditional mean μ

- **Conditional relationship:**

  $y = \beta x_1 + \varepsilon$ $E(\varepsilon|x) = 0$

  $y = \alpha + \beta x_1 + \varepsilon$ $E(\varepsilon|x) = 0$

  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ $E(\varepsilon|x) = 0$

  are linear models of the **conditional** expectation $E(y \mid x)$

All these models attempt to represent the conditional expectation of a Y given some other variable(s) X.

y: dependent variable, **Left hand side (LHS)** variable
x: explanatory variable, independent variable, regressor, **Right hand side (RHS)** variable

- **Linear or Non-linear model ?**

Are these model linear? $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2$

X_3 X_4

$y = \alpha + \beta \log(X_1) + \gamma e^{X_2}$

$y = \alpha x^{\beta}$ —— $\log y = \log a + \beta \log X + \epsilon$

$\sigma_t = \sigma_0 S_t^{\alpha}$ CEV model for volatility:

These: $y = \alpha + x^{\beta}$

$C = BS(S|K,\tau,\sigma,r)$ Black Scholes model

Some models are easily "linearized" by transformation of the data,
Some models ... are not.

- Many models do not suggest the form ofthe prediction error.

... Because they come from a theoretical world where they are ..... perfect!

Example: Black-Scholes arises from no-arbitrage....
So: $C = BS(S|K,\tau,\sigma,r) + \varepsilon$? or $\log C = \log BS(S|K,\tau,\sigma,r) + \varepsilon$?

Better

Theory does not say!