# L2. Introduction to General Regression Analysis

Yonghui Zhang

School of Economics, RUC

October 12, 2020

## Reading

- Chapter 2 *Econometrics* (Hansen, 2020)
- (If necessary) Chapters 1-6 *Introduction to Econometrics* (Hansen, 2020)

# Outline

# Basic notation

- Let $Z = (Y, X')'$ denote a $(k+1) \times 1$ random vector (rc) where $Y$ is a scalar random variable (rv) and $X$ is a $k \times 1$ r.c..
- Let $f(x, y)$ denote the joint *probability density function* (pdf) of $X$ and $Y$.
- Let $f_X(x)$ denote the *marginal pdf* of $X$ and $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$ denote the *conditional pdf* of $Y$ given $X = x$
- Suppress the subscripts and write $f_X(x)$ and $f_{Y|X}(y|x)$ as $f(x)$ and $f(y|x)$, respectively.
- We follow the tradition and study $f(y|x)$ through the study of its corresponding *moments*.

# Conditional mean and variance

- Conditional mean (before 1982):

$$E(Y|x) = E(Y|X=x) = \int yf(y|x)\,dx \text{ or } \sum_{i=1}^{\infty} y_i f(y_i|x)$$

- Conditional variance (Engle, 1982):

$$\begin{aligned} \mathtt{Var}(Y|x) &= \mathtt{Var}(Y|X=x) \\ &= E\left\{[Y - E(Y|x)]^2 |x\right\} \\ &= \int [y - E(Y|x)]^2 f(y|x)\,dy \\ &= E\left(Y^2|x\right) - [E(Y|x)]^2 \end{aligned}$$

# Conditional skewness and kurtosis

- Conditional skewness

$$s(Y|x) = \frac{E\left\{[Y - E(Y|x)]^3 \,|x\right\}}{[\mathtt{Var}(Y|x)]^{3/2}}$$

- Conditional kurtosis

$$\kappa(Y|x) = \frac{E\left\{[Y - E(Y|x)]^4 \,|x\right\}}{[\mathtt{Var}(Y|x)]^2}$$

# Conditional skewness and kurtosis

- Conditional skewness and kurtosis are widely used in finance. For example,
    - Harvey, C.R and Siddique, A. 1999. Autoregressive conditional skewness. JFQA.
    - Harvey, C.R and Siddique, A. 2000. Conditional skewness in asset pricing tests. JOF.
    - Jondeau, E., and Rockinger, M. 2003. Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. JEDC.
    - Ghysels, E., 2014. Conditional skewness with quantile regression models: SoFiE Presidential Address and a Tribute to Hal White. JFEc.
    - Langlois, H, 2020. Measuring skewness premia. JFE.
    - Jondeau, E., Zhang, Q., and Zhu, X. 2019. Average skewness matters. JFE.

# Conditional skewness and kurtosis

- Stein, R., 2019. Why Does Skewness Matter? Ask Kurtosis. Working paper.

**Abstract** I investigate the relationship between measures of skewness and expected stock returns. Forcing the data to fit a linear model, past research finds only a negative relationship between these variables. Using a novel methodology that endogenously estimates breakpoints in the relationship between two variables, I find three distinct zone. Expected returns are decreasing in skewness, but only for a region of relatively low absolute values of skewness. For distributions which are highly left- or right-skewed, the relationship is actually positive. Moreover, I find that kurtosis plays a major role in mediating this relationship. Adding measures of the fourth moment to all models tested turns all skewness coefficients negative, and most statistically insignificant. Relying on probability theory, I provide a theoretical framework that supports all empirical findings.

# Unconditional Moments

- **Remarks.** Recall the *unconditional* moments

$$\begin{aligned}
\mu &= E(Y), \\
\sigma^2 &= \mathtt{Var}(Y) = E[(Y-\mu)^2] \\
skewness &= \frac{E[(Y-\mu)^3]}{\sigma^3}, \\
kurtosis &= \frac{E[(Y-\mu)^4]}{\sigma^4}.
\end{aligned}$$

# Conditional Quantile

- Since Koenker and Basset (1978), conditional quantile function plays an important role in econometrics.
- The $\tau$th *conditional quantile function (CQF)* of $Y$ given $X = x$ is defined as

$$\begin{aligned} Q_\tau(x) &= \inf\{y : F(y|x) \geq \tau\} \\ &= F^{-1}(\tau|x) \text{ when the inverse function exists.} \end{aligned}$$

- Clearly, $P(Y \leq Q_\tau(x) | X = x) = \tau$.
- When $\tau = 1/2$, $Q_{1/2}(x)$ is the *conditional median*.

# Law of Iterated Expectations

- The key object in econometrics is $E(Y|X)$. Why?

**Definition (Regression function)**

$E(Y|X)$ is called the regression function (conditional mean) of $Y$ given $X$.

Here is the **first important property** of the conditional mean:

**Lemma (Law of Iterated/Double Expectations, LIE)**

*If* $E|Y| < \infty$ *and* $E|g(X,Y)| < \infty$,

$$\begin{aligned} &\text{(i) } E[E(Y|X)] = E(Y); \\ &\text{(ii) } E\{E[g(X,Y)|X]\} = E[g(X,Y)]; \\ &\text{(iii) } E[E(Y|X,Z)|X] = E(Y|X). \end{aligned}$$

# Law of Iterated Expectations

**Proof of LIE (ii).**

Note: (i) is a special case of (ii) with $g(X,Y) = Y$. Assume that $(X,Y)$ has a joint pdf $f(x,y)$. Then

$$\begin{aligned} Eg(X,Y) &= \int\int g(x,y) f(x,y)\, dxdy \\ &= \int\int g(x,y) f(y|x) f(x) dxdy \\ &= \int \left[ \int g(x,y) f(y|x)\, dy \right] f(x) dx \\ &= \int E\left[ g(X,Y) | x \right] \cdot f(x) dx \\ &= E\left\{ E\left[ g(X,Y) | X \right] \right\} \end{aligned}$$

$\square$

# Law of Iterated Expectations

Proof of LIE (iii).

$$
\begin{aligned}
E[E(Y|X,Z)|X] &= \int E(Y|X,z)\cdot f(z|X)dz \\
&= \int\left[\int yf(y|X,z)dy\right]\cdot f(z|X)dz \\
&= \int\int y\frac{f(y,X,z)}{f(X,z)}\cdot\frac{f(X,z)}{f(X)}dydz \\
&= \int y\left[\frac{\int f(y,X,z)dz}{f(X)}\right]dy \\
&= \int y\frac{f(y,X)}{f(X)}dy = \int yf(y|X)dy = E(Y|X).
\end{aligned}
$$

□

# Conditioning Theorem

**Theorem (Conditioning Theorem)**

*If* $E|Y| < \infty$, *then*

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

*If in addition* $E|g(X)| < \infty$ *then*

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

# Conditional mean

Example (The return of education)

Let $Y = wage$, $X = education$, $G = gender$ (male or female). Then

$$\begin{aligned} & E\left(Y|X = High\ School\right) \\ = \ & E\left(Y|X = HS, G = M\right) \cdot \Pr\left(M|X = HS\right) \\ & + E\left(Y|X = HS, G = F\right) \cdot \Pr\left(F|X = HS\right) \\ = \ & E\left[E\left(Y|X = HS, G\right)\right] \end{aligned}$$

# Conditional mean is optimal in MSE

### Definition (Mean squared error, MSE)

Let $g(X)$ be a predictor of $Y$. The mean squared error of $g(X)$ is

$$\mathtt{MSE}\left(g\right)=E\left[Y-g\left(X\right)\right]^{2}.$$

### Theorem (The optimality of $E\left(Y|X\right)$)

$$E\left(Y|X\right)=\underset{g\in\mathcal{G}}{\operatorname{argmin}}\mathtt{MSE}\left(g\right)=\underset{g\in\mathcal{G}}{\operatorname{argmin}}E\left[Y-g\left(X\right)\right]^{2}$$

*where $\mathcal{G}$ is the space of all measurable and square-integrable functions:*

$$\mathcal{G}=\left\{g:\int g\left(x\right)^{2}f\left(x\right)dx<\infty\right\}.$$

# Conditional mean is optimal in MSE

Proof of the optimality of conditional mean.

$$
\begin{aligned}
\mathtt{MSE}(g) &= E\left[\underbrace{Y - E(Y|X)} + \underbrace{E(Y|X) - g(X)}\right]^2 \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \\
&\quad +2E\{[Y - E(Y|X)] \cdot [E(Y|X) - g(X)]\} \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \\
&\quad +2E\{E[(Y - E(Y|X))(E(Y|X) - g(X))|X]\} \text{ (LIE)} \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \\
&\quad +2E\left\{\underbrace{E[(Y - E(Y|X))|X]}_{=0} \cdot (E(Y|X) - g(X))\right\} \\
&\geq E[Y - E(Y|X)]^2
\end{aligned}
$$

# Conditional mean is optimal in MSE

- **Remarks:**
  1. The above theorem states the **optimality** of $E(Y|X)$, which is the *second* important property of conditional mean.
  2. The MSE criterion is one of the most popular criteria to evaluate how well a function ($g(X)$ above) can approximate/predict $Y$.
  3. Other criteria for prediction in econometrics exist. For example, we will introduce mean absolutely deviation (MAE):

$$\texttt{MAE}(g) = E\left|Y - g(X)\right|.$$

     The optimizer for $\texttt{MAE}(g)$ is the conditional median function ($Q_{0.5}(X)$, 0.5 conditional quantile).

# Variance decomposition formula

**Theorem (Variance decomposition formula)**

*Let $X$ and $Y$ be two rv's and $\mathtt{Var}(Y) < \infty$. Then*

$$\mathtt{Var}(Y) = \mathtt{Var}(E(Y|X)) + E[\mathtt{Var}(Y|X)].$$

- **Remarks:**
  1. The variance of $Y$ can be decomposed into two parts: "between group" variation ($\mathtt{Var}(E(Y|X))$) and "within-group" variation ($E[\mathtt{Var}(Y|X)]$). In some sense, the regressor $X$ is used to determine the "group", or we group the data according to variable $X$. You can image the case with only one discrete $X$.
  2. The first part comes from the randomness of $X$ ($E(Y|X) = m(X)$ is a rv), which is the variance component captured by $X$.

# Variance decomposition formula

Proof of variance decomposition formula.

$$
\begin{aligned}
\mathtt{Var}(Y) &= E(Y-EY)^2 \\
&= E\left[Y-E(Y|X)+E(Y|X)-EY\right]^2 \\
&= E\left\{E\left[(Y-E(Y|X)+E(Y|X)-EY)^2|X\right]\right\} \text{ (LIE)} \\
&= E\left\{E\left[(Y\text{-}E(Y|X))^2|X\right]\right\}+E\left\{E\left[(E(Y|X)\text{-}EY)^2|X\right]\right\} \\
&\quad +E\left\{E\left[(Y-E(Y|X))\cdot(E(Y|X)-EY)|X\right]\right\} \\
&= E\left[\mathtt{Var}(Y|X)\right]+\mathtt{Var}\left[E(Y|X)\right]+0
\end{aligned}
$$

because of $E\{E[(Y-E(Y|X))|X]\cdot(E(Y|X)-EY)\} = E\{0\cdot(E(Y|X)-EY)\}=0.$ $\square$

# Variance decomposition formula

One important implication of variance decomposition formula is the following theorem.

**Theorem**

*If* $E\left(Y^2\right) < \infty$, *then*

$$\mathtt{Var}\left(Y\right) \geq \mathtt{Var}\left[Y - E\left(Y|X_1\right)\right] \geq \mathtt{Var}\left[Y - E\left(Y|X_1, X_2\right)\right].$$

**Proof.**

Hint. (i) $\mathtt{Var}\left[Y - E\left(Y|X_1\right)\right] = E\left[Y - E\left(Y|X_1\right)\right]^2 =$
$E\left\{E\left[\left[Y - E\left(Y|X_1\right)\right]^2|X_1\right]\right\} = E\left[\mathtt{Var}\left(Y|X_1\right)\right]$. (ii) Let
$A = Y - E\left(Y|X_1\right)$, and check
$A - E\left(A|X_1, X_2\right) = Y - E\left(Y|X_1, X_2\right)$. □

# Variance decomposition formula

- **Remarks:**
  1. You can also prove the above claim from the right hand side (RHS) to the left hand side (LHS); it is left as an exercise.
  2. The ratio can be explained by $X$ is given by

$$\frac{\mathtt{Var}\left(E\left(Y|X\right)\right)}{\mathtt{Var}\left(Y\right)}.$$

     Clearly, it lies between 0 and 1.
  3. More regressors are used to explain $Y$, the variance explained by regressors should increase.

# Conditional Mean Regression Equation

**Definition (Conditional mean regression function)**

Let

$$m(X) = E(Y|X) \text{ and } \varepsilon = Y - m(X).$$

Then we obtain the following regression equation

$$Y = m(X) + \varepsilon$$

where $\varepsilon$ is called the regression disturbance/error term or conditional mean error, which satisfies $E(\varepsilon|X) = 0$ because of

$$E(\varepsilon|X) = E(Y - m(X)|X) = E(Y|X) - m(X) = 0.$$

# Conditional Mean Regression Equation

- **Remarks:**
  - $E(Y|X)$ can be used to *predict* the mean value of $Y$ using the information of $X$.
  - $E(\varepsilon|X) = 0$: there is no systematic bias in the above predictor. It further implies that
    - $E(\varepsilon) = 0$ by the LIE.
    - $E(g(X)\varepsilon) = 0$ and $E(X\varepsilon) = 0$.
  - Nothing about higher order conditional moments of $Y$ or $\varepsilon$ given $X$.

$$\texttt{Var}(Y|X) = \texttt{Var}(\varepsilon|X) = \sigma^2 \text{ or } \sigma^2(X).$$

Both *conditional homoskedasticity* or *heteroskedasticity* are allowed.

# Conditional Mean Regression Equation

Let us look at the following example where the conditional variance is of the main interest (degenerated first order characteristics).

Example (Autoregressive conditional heteroskedasticity, ARCH)

Suppose $\varepsilon = Z\sqrt{\alpha_0 + \alpha_1 X^2}$, where $Z$ and $X$ are independent, $E(Z) = 0$ and $\mathtt{Var}(Z) = 1$. Then

$$\begin{aligned} E(\varepsilon|X) &= E\left(Z\sqrt{\alpha_0 + \alpha_1 X^2}|X\right) \\ &= E(Z|X)\sqrt{\alpha_0 + \alpha_1 X^2} = E(Z)\sqrt{\alpha_0 + \alpha_1 X^2} = 0 \end{aligned}$$

and

$$\begin{aligned} \mathtt{Var}(\varepsilon|X) &= E\left(\varepsilon^2|X\right) - \left[E(\varepsilon|X)\right]^2 = E\left[Z^2\left(\alpha_0 + \alpha_1 X^2\right)|X\right] - 0 \\ &= E\left(Z^2\right)\left(\alpha_0 + \alpha_1 X^2\right) = \alpha_0 + \alpha_1 X^2. \end{aligned}$$

# Linear Regression and Best Linear Predictor

- In general, economic theory cannot predict the functional form of $m(X) = E(Y|X)$;
- In the classical econometrics, it is frequently assumed that $E(Y|X)$ is a function of $X$ and a vector of unknown parameters $\beta$:

$$m(X, \beta) - \text{a nonlinear function of } \beta$$

In most case, $m$ is specified as an *affine* function of $X$:

$$m(x, \beta) = x'\beta = \sum_{l=1}^{k} x_l \beta_l.$$

- Actually, $m(\cdot)$ can also be estimated *nonparametrically* (kernel method, sieve method, $k$-nearest-neighbor\KNN, neural network, deep learning,...).

# Linear Regression and Best Linear Predictor

**Definition (Affine function)**

Let $x = (x_1, \cdots, x_k)'$ and $\beta = (\beta_1, \cdots, \beta_k)'$. The class of affine functions is

$$\mathcal{A} = \{m : m(x) = x'\beta = \sum_{l=1}^{k} x_l \beta_l, \beta_l \in \mathbb{R}\}.$$

**Theorem (Best linear least square (LS) predictor)**

*Let $Y$ and $X$ be rv's such that $E(Y^2) < \infty$ and $E(XX')$ is non-singular. Then the best linear least squares (LS) predictor that solves the minimization problem*

$$\min_{m \in \mathcal{A}} E[Y - m(X)]^2 = \min_{\beta \in \mathbb{R}^k} E[Y - X'\beta]^2$$

*is the linear function $m^*(x) = x'\beta^*$ with $\beta^* = [E(XX')]^{-1} E(XY)$ being*

# Linear Regression and Best Linear Predictor

**Proof of best linear projection.**

$$S(\beta) = E\left(Y - X'\beta\right)^2 = E\left(Y^2\right) + \beta' E\left(XX'\right)\beta - 2E\left(YX'\right)\beta$$

which is quadratic form of $\beta$. Take the FOC (first order condition) and get

$$\frac{\partial S(\beta)}{\partial \beta} = 2E\left(XX'\right)\beta - 2E(XY) = 0.$$

It leads to $E\left(XX'\right)\beta^* - E(XY) = 0$ and $\beta^* = \left[E\left(XX'\right)\right]^{-1} E(XY)$.
Note the SOC (second order condition) is

$$\frac{\partial S(\beta)}{\partial \beta \partial \beta'} = 2E\left(XX'\right) > 0 \text{ (p.d.)}$$

Then $\beta^*$ is a well-defined global minimizer. $\square$

# Linear Regression and Best Linear Predictor

- **Remarks**

1. The conditions that $E\left(Y^2\right) < \infty$ and $E(XX')$ is non-singular can ensure $E\left(XY\right)$ is well defined by the Cauchy-Schwarz (CS) inequality. When $X$ is a scalar rv, we have

$$E\left(XY\right) \leq \left\{E\left(X^2\right)E\left(Y^2\right)\right\}^{1/2}.$$

2. In almost all regressions, an intercept is included in the regression. That is $X = \left(1, X_2, \cdots, X_k\right)^T$.
3. In general, $m^*\left(X\right) = X'\beta^* \neq E\left(Y|X\right)$ unless $E\left(Y|X\right)$ is indeed linear in $X$.

# Linear Model Specification

## Definition (Linear regression model)

The specification

$$Y = X'\beta + U,\ \beta \in \mathbb{R}^k$$

is called a linear regression model (LRM), where $U$ is called the disturbance or error term of the model.

## Theorem

*Let $Y$ and $X$ be rv's such that $E\left(Y^2\right) < \infty$ and $E(XX')$ is non-singular. Let*

$$Y = X'\beta + U,\ \beta \in \mathbb{R}^k.$$

*Let $\beta^*$ be the best linear LS approximation coefficients. Then*

$$\beta = \beta^* \text{ if and only if } E\left(XU\right) = 0.$$

# Linear Model Specification

Proof.

(Sufficiency) First, we prove the "if" part. If $E(XU) = 0$, then

$$E(XU) = E\left[X\left(Y - X'\beta\right)\right] = E[XY] - E\left[XX'\right]\beta = 0.$$

It follows that $\beta = \left[E\left(XX'\right)\right]^{-1} E(XY) = \beta^*$. (Necessity). Exercise. □

- The above theorem holds no matter whether $E(Y|X)$ is linear or nonlinear in $X$. This is important because even if $E(Y|X)$ is nonlinear in $X$ we can always write

$$Y = X'\beta + U$$

  for some $\beta \in \mathbb{R}^k$ such that the orthogonality condition $E(XU) = 0$ holds. (Note that $E(XU) = 0$ is weaker than $E(U|X) = 0$)
- **Reminder**. The LRM is an *artificial specification*. Nothing in economics ensures that $E(Y|X) = X'\beta_0$ for some $\beta_0 \in \mathbb{R}^k$.

# Linear Model Specification

**Definition (Correct specification in conditional mean)**

The LRM

$$Y = X'\beta + U, \, \beta \in \mathbb{R}^k$$

is correctly specified for $E(Y|X)$ if

$$E(Y|X) = X'\beta_0$$

for some $\beta_0 \in \mathbb{R}^k$.

- If $E(Y|X) \neq X'\beta$ for all $\beta \in \mathbb{R}^k$ then we say that the above LRM is *misspecified* for $E(Y|X)$. (or $E(Y|X) \notin \mathcal{A}$ : the class of affine functions)
- When the LRM is correctly specified, $\beta_0$ exists and is called the true parameter of interest.

# Linear Model Specification

**Theorem**

*If the LRM* $Y = X'\beta + U$ *is correctly specified for* $E(Y|X)$*, then*
*(i)* $Y = X'\beta_0 + \varepsilon$ *for some* $\beta_0 \in \mathbb{R}^k$ *where* $E(\varepsilon|X) = 0$;
*(ii)* $\beta_0 = [E(XX')]^{-1} E(XY) = \beta^*$.

# Linear Model Specification

Proof.

(i) The correct specification of the LRM $Y = X'\beta + U$ implies that

$$E(Y|X) = X'\beta_0 \text{ for some } \beta_0 \in \mathbb{R}^k.$$

Let $\varepsilon = Y - X'\beta_0$. Then

$$E(\varepsilon|X) = E\left(Y - X'\beta_0|X\right) = E(Y|X) - X'\beta_0 = 0.$$

(ii) $E(\varepsilon X) = E[E(\varepsilon X|X)] = E[E(\varepsilon|X)X] = 0$. Then

$$0 = E(\varepsilon X) = E\left[X\left(Y - X'\beta_0\right)\right] = E[XY] - E\left[XX'\right]\beta_0 = 0.$$

It follows that $\beta_0 = [E(XX')]^{-1} E(XY)$. $\square$

# Misspecification

- If the LRM $Y = X^T\beta + U$ is *misspecified*, we can have

$$E(XU) = 0 \text{ but } E(U|X) \neq 0.$$

  Then

$$E(Y|X) = X'\beta + E(U|X)$$

  and we usually cannot estimate $E(Y|X)$ through the estimation of $\beta$.
- If one is interested in the estimation of $E(Y|X)$, a test for the correct specification of the LRM can be based on checking whether $E(U|X) = 0$ or not.
- Keep in mind. $E(U|X) = 0$ is testable, but $E(UX) = 0$ is generally *not testable*.

# Misspecification

## Example

Consider the following data generating process (DGP):
$Y = 1 + X_2 + 0.5\left(X_2^2 - 1\right) + \varepsilon$ where $X_2$ and $\varepsilon$ are independent $N(0,1)$.
(i) Find $E(Y|X_2)$;
(ii) Suppose a LRM: $Y = \beta_1 + \beta_2 X_2 + U$ is specified, where $\beta = (\beta_1, \beta_2)'$. Find the best LS approximation coefficient $\beta^*$ and the linear LS predictor $m^*(X) = X'\beta^*$;
(iii) Let $U = Y - X'\beta^*$. Show that $E(XU) = 0$;
(iv) Check whether $\frac{d}{dX_2}E(Y|X_2) = \beta_2^*$, the second element in $\beta^*$.

# Omitted Variable Bias

- Consider the long regression as

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

and the short regression as

$$Y = \gamma_1 X_1 + \epsilon$$

where $U$ and $\epsilon$ are the projection errors, respectively. If $\beta_1$ in the long regression is the parameter of interest, omitting $X_2$ as in the short regression will render omitted variable bias (meaning $\gamma_1 \neq \beta_1$) unless $E\left(X_1 X_2\right) = 0$.

- Note that

$$\begin{aligned}
\gamma_1 &= \left[E\left(X_1^2\right)\right]^{-1} E\left(X_1 Y\right) \\
&= \left[E\left(X_1^2\right)\right]^{-1} E\left[X_1\left(\beta_1 X_1 + \beta_2 X_2 + U\right)\right] \\
&= \beta_1 + \beta_2 \left[E\left(X_1^2\right)\right]^{-1} E\left(X_1 X_2\right) \neq \beta_1
\end{aligned}$$

# Misspecification

**Final remarks on linear regression model:**

- Most economic theories may have some implications on and only on the conditional mean of the underlying economic variable $Y$. $E(Y|X)$ is important from a statistical perspective: it is the optimal predictor of $Y$ under the MSE criterion.
- On the other hand, even though economic theory may suggest a nonlinear/linear relationship, it doesn't give a completely specified function form for $E(Y|X)$.
- The commonly-used linear regression model (LRM) uses a linear function to approximate $E(Y|X)$.
  - If the LRM is correctly specified, $\frac{\partial}{\partial X_j} E(Y|X) = \beta_j$, the partial (marginal) effect of $X_j$ on $Y$, has a meaningful economic interpretation.
  - If the LRM is incorrectly specified, the above interpretation is invalid anymore. (only in average sense)
  - Therefore, we must be very cautious about the economic interpretations of the linear regression coefficients.

# Multivariate Normality

**Definition (Multivariate normality)**

Note that an $n \times 1$ rv

$$X \sim N(\mu, \Sigma)$$

if its pdf is given by

$$f(x) = (2\pi)^{-n/2} |\det(\Sigma)|^{-1/2} \exp\left\{-\frac{(x-\mu)' \Sigma^{-1} (x-\mu)}{2}\right\}.$$

- Partition $X$ as $X = (X_1', X_2')'$, where $X_1$ and $X_2$ are $n_1$- and $n_2$-dimensional, respectively. Partition the mean and variance of $X$ conformably as:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}$$

# Multivariate Normality

Let $A$ and $b$ be respectively a nonrandom matrix and nonrandom vector, each conformable with $X$.
The following theorem shows some important properties of multivariate normality

### Theorem

*If*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}\right]$$

*then*
$(i)$ $AX + b \sim N\left(A\mu + b, A\Sigma A'\right)$;
$(ii)$ $(X-\mu)' \Sigma^{-1} (X-\mu) \sim \chi^2(n)$;
$(iii)$ $X_1 \perp X_2$ *if and only if* $\Sigma_{12} = 0$;
$(iv)$ $X_1 | X_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'\right)$.

# Multivariate Normality

Proof.

Proof of part $(iv)$. Consider a random vector:

$$\begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

which is normal as a linear transformation of $X$. It is easy to verify that the two sub-vectors $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ and $X_2$ are uncorrelated and thus independent. Now, write

$$X_1 = \left(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2\right) + \Sigma_{12}\Sigma_{22}^{-1}X_2$$

where the first term is independent of $X_2$ and its conditional distribution given $X_2$ is the unconditional distribution, which is normal with mean $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$ and variance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. The second term is a constant when $X_2$ is given, which only shifts the mean of the conditional distribution of $X_1$ given $X_2$. Thus $(iv)$ follows. $\square$

# Multivariate Normality

- For the bivariate normal example,

$$
\begin{aligned}
f(x) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
&\times \exp\left\{-\frac{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2}{2\left(1-\rho^2\right)}\right\}
\end{aligned}
$$

where $E(X) = (\mu_1, \mu_2)'$ and $\mathtt{Var}(X) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

- Then

$$
\begin{aligned}
E(X_1|X_2 = x) &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) \\
\mathtt{Var}(X_1|X_2 = x) &= \sigma_1^2\left(1-\rho^2\right).
\end{aligned}
$$

# Causal Effects

- A variable $X_1$ can be said to have a causal effect on the response variable $Y$ if the latter changes when all other inputs are held constant. To make this precise, we need a mathematical formulation. Write a full model for the response variable $Y$ as

$$Y = h\left(X_1, X_2 . u\right)$$

  where $X_1$ and $X_2$ are the observed variables, $u$ is an $l \times 1$ unobserved random factor, and $h$ is a functional relationship.
- This framework is called the **potential outcomes** framework.
- We define the **causal effect** of $X_1$ within this model as the change in $Y$ due to a change in $X_1$ holding the other variables $X_2$ and $u$ constant.

# Causal Effects

**Definition**

The causal effect of $X_1$ on $Y$ is

$$C(X_1, X_2, u) = \frac{\partial h(X_1, X_2, u)}{\partial X_1},$$

the change in $Y$ due to a change in $X_1$, holding $X_2$ and $u$ constant.

- Sometimes it is useful to write this relationship as a potential outcome function

$$Y(X_1) = h(X_1, X_2.u)$$

where the notation implies that $Y(X_1)$ is holding $X_2$ and $u$ constant

# Causal Effects

Treatment effect of a binary regressor $X_1$

- $X_1 = 1$-treatment and $X_1 = 0$-non-treatment. Then

$$\begin{aligned} Y(0) &= h(0, X_2, u) \\ Y(1) &= h(1, X_2, u) \end{aligned}$$

- The causal effect of treatment for the individual is the change in $Y$ due to treatment while we hold both $X_2$ and $u$ constant:

$$C(X_2, u) = Y(1) - Y(0),$$

which is random (a function of $X_2$ and $u$) as the potential outcomes $Y(0)$ and $Y(1)$ differ across individuals.

- Problems: (i) we can **only observe the realized value** $Y = Y(0)$ if $X_1 = 0$, $= Y(1)$ if $X_1 = 1$; (ii) As the causal effect **varies across individuals** and is not observable it cannot be measured on the individual level.

# Causal Effects

Average causal effects

**Definition**

The average causal effect of $X_1$ on $Y$ conditional on $X_2 = x_2$ is

$$\begin{aligned} \texttt{ACE}\left(x_1, x_2\right) &= E\left[C\left(X_1, X_2, u\right) \middle| X_1 = x_1, X_2 = x_2\right] \\ &= \int_{\mathbb{R}^l} \frac{\partial h(x_1, x_2, u)}{\partial x_1} f\left(u | x_1, x_2\right) du \end{aligned}$$

where $f\left(u | x_1, x_2\right)$ is the conditional density of $u$ given $x_1$ and $x_2$.

# Causal Effects

**Question**

What is the relationship between the average causal effect $\texttt{ACE}\left(x_1, x_2\right)$ and the regression derivative $\frac{\partial m\left(x_1, x_2\right)}{\partial x_1}$?

- First, note that

$$\begin{aligned} m\left(x_1, x_2\right) &= E\left(h\left(x_1, x_2, u\right) \middle| x_1, x_2\right) \\ &= \int_{\mathbb{R}^l} h(x_1, x_2, u) f\left(u \middle| x_1, x_2\right) du \end{aligned}$$

- Second, we have

$$\begin{aligned} \frac{\partial m(x_1, x_2)}{\partial x_1} &= \int_{\mathbb{R}^l} \frac{\partial h(x_1, x_2, u)}{\partial x_1} f\left(u \middle| x_1, x_2\right) du \\ &\quad + \int_{\mathbb{R}^l} h(x_1, x_2, u) \frac{\partial f\left(x_1, x_2, u\right)}{\partial x_1} du \\ &= \texttt{ACE}\left(x_1, x_2\right) + \int_{\mathbb{R}^l} h(x_1, x_2, u) \frac{\partial f\left(x_1, x_2, u\right)}{\partial x_1} du \end{aligned}$$

# Causal Effects

When $\frac{\partial f(x_1,x_2,u)}{\partial x_1} = 0$, namely, the conditional density of $u$ given $(X_1, X_2)$ does not depend on $X_1$, the regression derivative equals the ACE.

**Definition (Conditional Independence Assumption (CIA))**

Conditional on $X_2$, the rv $X_1$ and $u$ are statistically independent.

- Under this assumption, $f(u|x_1, x_2) = f(u|x_2)$ and thus $\frac{\partial f(x_1,x_2,u)}{\partial x_1} = 0$.

**Theorem**

*The Conditional Independence Assumption implies*

$$\frac{\partial m(x_1, x_2)}{\partial x_1} = \mathtt{ACE}(x_1, x_2)$$

*and the regression derivative equals the average causal effect for $X_1$ on $Y$ conditional on $X_2$.*

# Exercises

**Due: October 19, 18:00PM (No LATE WILL BE ACCEPTED!)**

1. Let $X$ be a uniform random variable on $[0,1]$, i.e., its pdf is $f(x) = 1_{\{0 \leq x \leq 1\}}$, where $1_A$ is the usual indicator function. Suppose $Y$ is binary variables such that

$$P(Y = 1|X = x) = x \text{ and } P(Y = 0|X = x) = 1 - x.$$

   Calculate $E(Y)$ and $\mathtt{Var}(Y)$ by the LIE and variance decomposition formula.

2. Hansen *Econometrics*. Ex 2.1, 2.2, 2.5, 2.6, 2.10–2.14, 2.16, 2.21.