CHAPTER 2 GENERAL REGRESSION ANALYSIS AND MODEL SPECIFICATION

Key words: Conditional distribution, Conditional mean, consumption function, Linear regression model, Marginal propensity to consume, Model specification, Regression

Abstract: This chapter introduces *regression analysis*, the most popular statistical tool to explore the dependence of one variable (say Y) on others (say X). The variable Y is called the dependent variable, and X is called the independent variable or explanatory variable. The regression relationship between X and Y can be used to study the effect of X on Y or to predict Y using X. We motivate the importance of the regression function from both the economic and statististical perspectives, and characterize the condition for correct specification of a linear model for the regression function, which is shown to be crucial for a valid economic interpretation of model parameters.

2.1 Conditional Probability Distribution

Notational Convention: Throughout this book, capital letters ($e.g., Y$) denote random variables or random vectors, lower case letters ($e.g., y$) denote realizations of random variables.

We assume that $Z = (Y, X')'$ is a random vector with $E(Y^2) < \infty$, where Y is a scalar, X is a $(k+1) \times 1$ vector of economic variables with its first component being a constant, and X' denotes the transpose of X. Given this assumption, the conditional mean $E(Y|X)$ exists and is well-defined.

Statistically speaking, the relationship between two random variables or vectors X (e.g., oil price change) and Y (e.g., economic growth) can be characterized by their joint distribution function. Suppose $(X', Y)'$ are continuous random vectors, and the joint probability density function (*pdf*) of $(X', Y)'$ is $f(x, y)$. Then the marginal *pdf* of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy,$$

and the conditional *pdf* of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)},$$

provided $f_X(x) > 0$. The conditional pdf $f_{Y|X}(y|x)$ completely describes how Y depends on X. With this conditional *pdf* $f_{Y|X}(y|x)$, we can compute the following quantities:

- The conditional mean

$$\begin{aligned} E(Y|x) &\equiv E(Y|X=x) \\ &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy; \end{aligned}$$

- the conditional variance

$$\begin{aligned} \text{var}(Y|x) &\equiv \text{var}(Y|X=x) \\ &= \int_{-\infty}^{\infty} [y - E(Y|x)]^2 f_{Y|X}(y|x)dy \\ &= E(Y^2|x) - [E(Y|x)]^2; \end{aligned}$$

- the conditional skewness

$$S(Y|x) = \frac{E[(Y - E(Y|x))^3|x]}{[\text{var}(Y|x)]^{3/2}};$$

- the conditional kurtosis

$$K(Y|x) = \frac{E[(Y - E(Y|x))^4|x]}{[\text{var}(Y|x)]^2};$$

- the α-conditional quantile $Q(x, \alpha)$:

$$P\left[Y \leq Q(X, \alpha)|X = x\right] = \alpha \in (0, 1).$$

  Note that when $\alpha = 0.5$, $Q(x, 0.5)$ is the conditional median, which is the cutoff point or threshold that divides the population into two equal halves, conditional on $X = x$.

The class of conditional moments is a summary characterization of the conditional distribution $f_{Y|X}(y|x)$. A mathematical model (i.e., an assumed functional form with a finite number of unknown parameters) for a conditional moment is called an econometric model for that conditional moment.

Question: Which moment to model and use in practice?

It depends on economic applications. For some applications, we only need to model the first conditional moment, namely the conditional mean. For example, asset pricing aims at explaining excess asset returns by systematic risk factors. An asset pricing model is essentially a model for the conditional mean of asset returns on risk factors. For others, we may have to model higher order conditional moments and even the entire conditional distribution. In econometric practice, the most popular models are the first two conditional moments, namely the conditional mean

and conditional variance. There is no need to model the entire conditional distribution of Y given X when only certain conditional moments are needed. For example, when the conditional mean is of concern, there is no need to model the conditional variance or impose restrictive conditions on it.

The conditional moments, and more generally the conditional probability distribution of Y given X, are not the causal relationship from X to Y. They are a predictive relationship. That is, one can use the information on X to predict the distribution of Y or its attributes. These probability concepts cannot tell whether the change in Y is caused by the change in X. Such causal interpretation has to reply on economic theory. Economic theory usually hypothesizes that a change in Y is caused by a change in X, i.e., there exists a causal relationship from X to Y. If such an economic causal relationship exists, we will find a predictive relationship from X to Y. On the other hand, a documented predictive relationship from X to Y may not caused by an economic causal relationship from X to Y. For example, it is possible that both X and Y are positively correlated due to their dependence on a common factor. As a result, we will find a predictive relationship from X to Y, although they do not have any causal relationship. In fact, it is well-known in econometrics that some economic variables that trend consistently upwards over time are highly correlated even in the absence of any causal relationship between them. Such strong correlations are called spurious relationships.

2.2 Regression Analysis

We now focus on the first conditional moment, $E(Y|X)$, which is called the regression function of Y on X, where Y is called the regressand, and X is called the regressor vector. The term "regression" is used to signify a predictive relationship between Y and X.

Definition [Regression Function]: The conditional mean $E(Y|X)$ is called a regression function of Y on X.

Many economic theories can be characeterized by the conditional mean $E(Y|X)$ of Y given X, provided X and Y are suitably defined. Most, though not all, of dynamic economic theories and/or dynamic optimization models, such as rational expectations, efficient markets hypothesis, expectations hypothesis, and optimal dynamic asset pricing, have important implications on (and only on) the conditional mean of underlying economic variables given the information available to economic agents (e.g., Cochrane 2001, Sargent and Ljungqvist 2002). For example, the classical efficient market hypothesis states that the expected asset return given the information available, is zero, or at most, is constant over time; the optimal dynamic asset pricing theory implies that the expectation of the pricing error given the information available is zero for each asset (Cochrane 2001). Although economic theory may suggest a nonlinear relationship, it does not

give a completely specified functional form for the conditional mean of economic variables. It is therefore important to model the conditional mean properly.

Before modeling $E(Y|X)$, we first discuss some probabilistic properties of $E(Y|X)$.

Lemma: $E[E(Y|X)] = E(Y)$.

Proof: The result follows immediately from applying the law of iterated expectations below.

Lemma [Law of Iterated Expectations (LIE)]: *For any measurable function* $G(X, Y)$,

$$E\left[G(X,Y)\right] = E\left\{E\left[G(X,Y)|X\right]\right\},$$

provided the expectation $E[G(X,Y)]$ *exists.*

Proof: We consider the case of the continuous distribution of $(Y, X')'$ only. By the multiplication rule that the joint *pdf* $f(x,y) = f_{Y|X}(y|x)f_X(x)$, we have

$$\begin{aligned}
E[G(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)f_{XY}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)f_{Y|X}(y|x)f_X(x)dxdy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} G(x,y)f_{Y|X}(y|x)dy\right] f_X(x)dx \\
&= \int_{-\infty}^{\infty} E[G(X,Y)|X=x]f_X(x)dx \\
&= E\{E[G(X,Y)|X]\},
\end{aligned}$$

where the operator $E(\cdot|X)$ is the expectation with respect to $f_{Y|X}(\cdot|X)$, and the operator $E(\cdot)$ is the expectation with respect to $f_X(\cdot)$. This completes the proof.

Interpretation of $E(Y|X)$ **and LIE**:

Example 1: Suppose Y is wage, and X is a gender dummy variable, taking value 1 if an employee is female and value 0 if an employee is male. Then

$$\begin{aligned}
E(Y|X &= 1) = \text{ average wage of a female worker,} \\
E(Y|X &= 0) = \text{ average wage of a male worker,}
\end{aligned}$$

and the overall average wage

$$\begin{aligned} E(Y) &= E[E(Y|X)] \\ &= P(X=1)E(Y|X=1) + P(X=0)E(Y|X=0), \end{aligned}$$

where $P(X=1)$ is the proportion of female employees in the labor force, and $P(X=0)$ is the proportion of the male employees in the labor force. The use of LIE here thus provides some insight into the income distribution between genders.

Example 2: Suppose Y is an asset return and we have two information sets: X and $\tilde{X}$, where $X \subset \tilde{X}$ so that all information in X is also in $\tilde{X}$ but $\tilde{X}$ contains some extra information. Then we have a conditional version of the law of iterated expectations says that

$$E(Y|X) = E[E(Y|\tilde{X})|X]$$

or equivalently

$$E\left\{\left[Y - E(Y|\tilde{X})|X\right]\right\} = 0.$$

where $Y - E(Y|\tilde{X})$ is the prediction error using the superior information set $\tilde{X}$. The conditional LIE says that one cannot use limited information X to predict the prediction error one would make if one had superior information $\tilde{X}$. See Campbell, Lo and MacKinlay (1997, p.23) for more discussion.

Question: Why is $E(Y|X)$ important from a statistical perspective?

Suppose we are interested in predicting Y using some function $g(X)$ of X, and we use a so-called Mean Squared Error (MSE) criterion to evaluate how well $g(X)$ approximates Y. Then the optimal predictor under the MSE criterion is the conditional mean, as will be shown below.

We first define the MSE criterion. Intuitively, MSE is the average of the squared deviations between the predictor $g(X)$ and the actual outcome Y.

Definition [MSE]: Suppose function $g(X)$ is used to predict Y. Then the mean squared error of function $g(X)$ is defined as

$$MSE(g) = E\left[Y - g(X)\right]^2,$$

provided the expectation exists.

The theorem below states that $E(Y|X)$ minimizes the MSE.

Theorem [Optimality of $E(Y|X)$]: The regression function $E(Y|X)$ *is the solution to the*

optimization problem

$$\begin{aligned} E(Y|X) &= \arg\min_{g\in\mathbb{F}} MSE(g) \\ &= \arg\min_{g\in\mathbb{F}} E[Y-g(X)]^2, \end{aligned}$$

where $\mathbb{F}$ *is the space of all measurable and square-integrable functions*

$$\mathbb{F} = \{g(\cdot)\colon \int_{-\infty}^{\infty} g^2(x) f_X(x) dx < \infty\}.$$

Proof: We will use the variance and squared-bias decomposition technique. Put

$$g_o(X) \equiv E(Y|X).$$

Then

$$\begin{aligned} MSE(g) &= E[Y-g(X)]^2 \\ &= E\left[Y-g_o(X)+g_o(X)-g(X)\right]^2 \\ &= E\left[Y-g_o(X)\right]^2 + E\left[g_o(X)-g(X)\right]^2 \\ &\quad +2E\{\left[Y-g_o(x)\right]\left[g_o(X)-g(X)\right]\} \\ &= E\left[Y-g_o(X)\right]^2 + E\left[g_o(X)-g(X)\right]^2, \end{aligned}$$

where the cross-product term

$$E\{\left[Y-g_o(x)\right]\left[g_o(X)-g(X)\right]\} = 0$$

by LIE and the fact that $E\{[Y-g_o(X)]|X\} = 0$ a.s.

In the above MSE decomposition, the first term $E[Y-g_o(X)]^2$ is the quadratic variation of the prediction error of the regression function $g_o(X)$. This does not depend on the choice of function $g(X)$. The second term $E[g_o(X)-g(X)]^2$ is the is the quadratic variation of the approximation error of $g(X)$ for $g_o(X)$. This term achieves its minimum of zero if and only if one chooses $g(X) = g_o(X)$ a.s. Because the first term $E[Y-g_o(X)]^2$ does not depend on $g(X)$, minimizing MSE(g) is equivalent to minimizing the second term $E[g_o(X)-g(X)]^2$. Thefore, the optimal solution for minimizing MSE(g) is given by $g^*(X) = g_o(X)$. This completes the proof.

Remarks:

MSE is a popular criterion for measuring precision of a predictior $g(X)$ for Y. It has at least

two advantages: first, it can be analyzed conveniently, and second, it has a nice decomposition of a variance component and a squared-bias component.

However, MSE is one of many possible criteria for measuring goodness of the predictor $g(X)$ for Y. In general, any increasing function of the absolute value $|Y - g(X)|$ can be used to measure the goodness of fit for the predictor $g(X)$. For example, the Mean Absolute Error

$$MAE(g) = E|Y - g(X)|$$

is also a reasonable criterion.

It should be emphasized that different criteria have different optimizers. For example, the optimizer for $MAE(g)$ is the conditional median, rather than the conditional mean. The conditional median, say $m(x)$, is defined as the solution to

$$\int_{-\infty}^{m} f_{Y|X}(y|x)dy = 0.5.$$

In other words, $m(x)$ divides the conditional population into two equal halves.

Example: Let the joint *pdf* $f_{XY}(x, y) = e^{-y}$ for $0 < x < y < \infty$. Find $E(Y|X)$ and $\text{var}(Y|X)$.

Solution: We first find the conditional pdf $f_{Y|X}(y|x)$. The marginal *pdf* of X

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dy \\ &= \int_{x}^{\infty} e^{-y}dy \\ &= e^{-x} \text{ for } 0 < x < \infty. \end{aligned}$$

Therefore,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x, y)}{f_X(x)} \\ &= e^{-(y-x)} \text{ for } 0 < x < y < \infty. \end{aligned}$$

Then

$$\begin{aligned} E(Y|x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \\ &= \int_{x}^{\infty} y e^{-(y-x)}dy \\ &= e^{x}\int_{x}^{\infty} y e^{-y}dy \\ &= -e^{x}\int_{x}^{\infty} y de^{-y} \\ &= 1+x. \end{aligned}$$

Thus, the regression function $E(Y|X)$ is linear in X.

To compute $\text{var}(Y|X)$, we will use the formula

$$\text{var}(Y|X) = E(Y^2|X) - [E(Y|X)]^2 .$$

Because

$$\begin{aligned} E(Y^2|x) &= \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)dy \\ &= \int_{x}^{\infty} y^2 e^{-(y-x)}dy \\ &= e^{x}\int_{x}^{\infty} y^2 e^{-y}dy \\ &= -e^{x}\int_{x}^{\infty} y^2 de^{-y} \text{ where } de^{-y} = -e^{-y}dy. \\ &= (-e^{x})\left[y^2 e^{-y}|_{x}^{\infty} - \int_{x}^{\infty} e^{-y}dy^2\right] \\ &= [-e^{x}]\left[0 - x^2 e^{-x} - 2\int_{x}^{\infty} y e^{-y}dy\right] \\ &= x^2 + 2e^{x}\int_{x}^{\infty} y e^{-y}dy \\ &= x^2 + 2\int_{x}^{\infty} y e^{-(y-x)}dy \\ &= x^2 + 2(1+x), \end{aligned}$$

we have

$$\begin{aligned} \text{var}(Y|x) &= E(Y^2|x) - [E(Y|x)]^2 \\ &= x^2 + 2(1+x) - (1+x)^2 \\ &= 1. \end{aligned}$$

The conditional variance of Y given X does not depend on X. That is, X has no effect on the conditional variance of Y.

The above example shows that while the conditional mean of Y given X is a linear function of X, the conditional variance of Y may not depend on X. This is essentially the assumption made in the classical linear regression model (see Chapter 3 below). Another example for which we have a linear regression function with constant conditional variance is when X and Y are jointly normally distributed (see Exercise 2 at the end of this chapter).

Theorem [Regression Identity]: *Suppose $E(Y|X)$ exists. Then we can always write*

$$Y = E(Y|X) + \varepsilon,$$

where ε is called the regression disturbance and has the property that

$$E(\varepsilon|X) = 0.$$

Proof: Put $\varepsilon = Y - E(Y|X)$. Then

$$Y = E(Y|X) + \varepsilon,$$

where

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

Remarks:

The regression function $E(Y|X)$ can be used to predict the expected value of Y using the information of X. In regression analysis, an important issue is the direction of causation between Y and X.In practice, one often hope to check whether Y "depends" on or can be "explained" by X, with help of economic theory. For this reason, Y is called the dependent variable, and X is

called the explantory variable or vector. However, it should be emphasized that the regression function $E(Y|X)$ itself does not tell any causal relationship between Y and X.

The random variable ε represents the part of Y that is not captured by $E(Y|X)$. It is usually called a *noise* or a *disburbance*, because it "disturbs" an otherwise stable relationship between Y and X. On the other hand, the regression function $E(Y|X)$ is called a *signal*.

The property that $E(\varepsilon|X) = 0$ implies that the regression disturbance ε contains no systematic information of X that can be used to predict the expected value of Y. In other words, all information of X that can be used to predict the expectation of Y has been completely summarized by $E(Y|X)$. The condition $E(\varepsilon|X) = 0$ is crucial for the validity of economic interpretation of model parameters, as will be seen shortly.

$E(\varepsilon|X) = 0$ implies that the unconditional mean of ε is zero:

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

and that ε is orthogonal to X :

$$\begin{aligned} E(X\varepsilon) &= E\left[E(X\varepsilon|X)\right] \\ &= E\left[XE(\varepsilon|X)\right] \\ &= E(X \cdot 0) \\ &= 0. \end{aligned}$$

Since $E(\varepsilon) = 0$, we have $E(X\varepsilon) = \text{cov}(X, \varepsilon)$. Thus, orthogonality ($E(X\varepsilon) = 0$) means that X and ε are uncorrelated.

In fact, ε is orthogonal to any measurable function of X, i.e., $E[\varepsilon h(X)] = 0$ for any measurable function $h(\cdot)$. This implies that we cannot predict the mean of ε by using any possible model $h(X)$, no matter it is linear or nonlinear.

Question: Is $E(\varepsilon|X) = 0$ equivalent to $E[\varepsilon h(X)] = 0$ for all measurable $h(\cdot)$?
Answer: Yes. How to show it? See Exercise 11 at the end of this chapter for more discussion.

It is possible that $E(\varepsilon|X) = 0$ but $\text{var}(\varepsilon|X)$ is a function of X. If $\text{var}(\varepsilon|X) = \sigma^2 > 0$, we say that there exists **conditional homoskedasticity** for ε. In this case, X cannot be used to predict the (quadratic) variation of Y. On the other hand, if $\text{var}(\varepsilon|X) \neq \sigma^2$ for any constant $\sigma^2 > 0$, we say that there exists **conditional heteroskedasticity**. Econometric procedures of regression analysis are usually different, depending on whether there exists conditional heteroskedasticity. For example, the so-called conventional t-test and F-test are invalid under conditional heteroskedasticity (see Chapter 3 for the introduction of the t-test and F-test). This will be discussed in detail in subsequent chapters.

Example: Suppose

$$\varepsilon = \eta\sqrt{\beta_0 + \beta_1 X^2},$$

where random variables X and η are independent, and $E(\eta) = 0$,$\text{var}(\eta) = 1$. Find $E(\varepsilon|X)$ and $\text{var}(\varepsilon|X)$.

Solution:

$$\begin{aligned} E(\varepsilon|X) &= E\left[\eta\sqrt{\beta_0 + \beta_1 X^2}\,|X\right] \\ &= \sqrt{\beta_0 + \beta_1 X^2}E(\eta|X) \\ &= \sqrt{\beta_0 + \beta_1 X^2}E(\eta) \\ &= \sqrt{\beta_0 + \beta_1 X^2}\cdot 0 \\ &= 0. \end{aligned}$$

Next,

$$\begin{aligned} \text{var}(\varepsilon|X) &= E\left\{[\varepsilon - E(\varepsilon|X)]^2|X\right\} \\ &= E(\varepsilon^2|X) \\ &= E[\eta^2(\beta_0 + \beta_1 X^2)|X] \\ &= (\beta_0 + \beta_1 X^2)E(\eta^2|X) \\ &= (\beta_0 + \beta_1 X^2)\cdot 1 \\ &= \beta_0 + \beta_1 X^2. \end{aligned}$$

Although the conditional mean ε given X is identically zero, the conditional variance of ε given X depends on X.

The regression analysis (conditional mean analysis) is the most popular statistical method in econometrics. It has been applied widely to economics. For example, it can be used to

- estimate the relationship between economic variables.
- test economic hypotheses.
- forecast future values of Y.

Example 1: Let Y=consumption, X=disposable income. Then the regression function $E(Y|X) = C(X)$ is the so-called consumption function, and the marginal propensity to consume (MPC) is the derivative

$$MPC = C'(X) = \frac{d}{dX}E(Y|X).$$

MPC is an important concept in the "multiplier effect" analysis. The magnitude of MPC is important in macroeconomic policy analysis and forecasting. On the other hand, when Y is consumption on food only, then Engle's law implies that MPC must be a decreasing function of X. Therefore, we can test Engle's law by testing whether $C'(X) = \frac{d}{dX}E(Y|X)$ is a decreasing function of X.

Example 2: Y=output, X=(labor, capital, and raw material)$'$, then the regression $E(Y|X) = F(X)$ is the so-called production function. This can be used to test the hypothesis of constant return to scale (CRS), which is defined as

$$\lambda F(X) = F(\lambda X) \text{ for all } \lambda > 0.$$

Example 3: Let Y be the cost of producing certain output X. Then the regression function $E(Y|X) = C(X)$ is the cost function. For a monopoly firm or industry, the marginal cost must be declining in output X. That is,

$$\begin{aligned} \frac{d}{dX}E(Y|X) &= C'(X) > 0, \\ \frac{d^2}{dX^2}E(Y|X) &= C''(X) < 0. \end{aligned}$$

These imply that the cost function of a monopoly is a nonlinear function of X.

Question: Why may there exist conditional heteroskedasticity?

Generally speaking, given that $E(Y|X)$ depends on X, it is conceivable that $\text{var}(Y|X)$ and other higher order conditional moments may also depend on X. In fact, conditional hereroskedasticity may arise from different sources. For example, a larger firm may have a large output variation. Granger and Machina (2006) explain why economic variables may display volatility clustering from an econometric structral perspective.

The following example shows that conditional heteroskedasticity may arise due to random coefficients in a data generating process.

Example 4 [Random Coefficient Process]: Suppose

$$Y = \beta_0 + (\beta_1 + \beta_2\eta)X + \eta,$$

where X and η are independent, and $E(\eta) = 0$,$\text{var}(\eta) = \sigma_\eta^2$. Find the conditional mean $E(Y|X)$ and conditional variance $\text{var}(Y|X)$.

Solution: (i)

$$\begin{aligned} E(Y|X) &= \beta_0 + E[(\beta_1 + \beta_2\eta)X|X] + E(\eta|X) \\ &= \beta_0 + \beta_1 X + \beta_2 X E(\eta|X) + E(\eta|X) \\ &= \beta_0 + \beta_1 X + \beta_2 X E(\eta) + E(\eta) \\ &= \beta_0 + \beta_1 X + \beta_2 X \cdot 0 + 0 \\ &= \beta_0 + \beta_1 X. \end{aligned}$$

(ii)

$$\begin{aligned} \text{var}(Y|X) &= E\left[(Y - E(Y|X))^2|X\right] \\ &= E\left[(\beta_0 + (\beta_1 + \beta_2\eta)X + \eta - \beta_0 - \beta_1 X)^2|X\right] \\ &= E\left[(\beta_2 X\eta + \eta)^2|X\right] \\ &= E\left[(\beta_2 X + 1)^2\eta^2|X\right] \\ &= (1 + \beta_2 X)^2 E(\eta^2|X) \\ &= (1 + \beta_2 X)^2 E(\eta^2) \\ &= (1 + \beta_2 X)^2\sigma_\eta^2. \end{aligned}$$

The random coefficient process has been used to explain why the conditional variance may depend on the regressor X. We can write this process as

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where

$$\varepsilon = (1 + \beta_2 X)\eta.$$

Note that $E(\varepsilon|X) = 0$ but $\text{var}(\varepsilon|X) = (1 + \eta_2 X)^2\sigma_\eta^2$.

2.3 Linear Regression Modeling

As we have known above, the conditional mean $g_o(X) \equiv E(Y|X)$ is the solution to the MSE optimization problem

$$\min_{g\in\mathbb{F}} E[Y - g(X)]^2,$$

where $\mathbb{F}$ is a class of functions that includes all measurable and square-integrable functions, i.e.,

$$\mathbb{F} = \left\{ g(\cdot) : \mathbb{R}^{k+1} \to \mathbb{R} \mid \int g^2(x) f_X(x) dx < \infty \right\}.$$

In general, the regression function $E(Y|X)$ is an unknown functional form of X. Economic

theory usually suggests a qualitative relationship between X and Y (e.g., the cost of production is an increasing function of output X), but it never suggests a concrete functional form. One needs to use some mathematical model to approximate $g_o(X)$.

Question: How to model $g_o(X)$?

In econometrics, a most popular modeling strategy is the parametric approach, which assumes a known functional form for $g_o(X)$, up to some unknown parameters. In particular, one usually uses a class of linear functions to approximate $g_o(x)$, which is simple and easy to interpret. This is the approach we will take in most of this book.

We first introduce a class of affine functions.

Definition [Affine Functions]: Denote

$$X = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_k \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Then the class of affine functions is defined as

$$\begin{aligned} \mathbb{A} &= \{g : \mathbb{R}^{k+1} \to \mathbb{R} : g(X) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j, \beta_j \in \mathbb{R}\} \\ &= \{g : \mathbb{R}^{k+1} \to \mathbb{R} \mid g(X) = \beta' X\}. \end{aligned}$$

Here, there is no restriction on the values of parameter vector β. For this class of functions, the functional form is known to be linear in both explanatory varibles X and parameters β; the unknown is the $(k+1) \times 1$ vector β.

Remarks:

From an econometric point of view, the key feature of $\mathbb{A}$ is that $g(X) = X'\beta$ is linear in β, not in X. Later, we will generalize $\mathbb{A}$ so that $g(X) = X'\beta$ is linear in β but is possibly nonlinear in X. For example, when $k = 1$, we can generalize $\mathbb{A}$ to include

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2,$$

or

$$g(X) = \beta_0 + \beta_1 \ln X_1.$$

These possibilities are included in $\mathbb{A}$ if we properly redefine X as $X = (1, X_1, X_1^2)'$ or $X = (1, \ln X_1)'$. Therefore, our econometric theory to be developed in subsequent chapters are actually applicable to all regression models that are linear in β but not necessarily linear in X. Such models are called linear regression models. Conversely, a nonlinear regression model for $g_o(X)$ means a known parametric functional form $g(X, \beta)$ which is nonlinear in β. An example is the so-called logistic regression model

$$g(X, \beta) = \frac{1}{1 + \exp(-X'\beta)}.$$

Nonlinear regression models can be handled using the analytic tools developed in Chapter 8. See more discussions there.

We now solve the constrained minimization problem

$$\min_{g \in \mathbb{A}} E[Y - g(X)]^2 = \min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2.$$

The solution $g^*(X) = X'\beta^*$ is called the **Best Linear Least Squares** Predictor for Y, and β^* is called the best LS approximation coefficient vector.

Theorem [Best Linear LS Prediction]: *Suppose $E(Y^2) < \infty$ and the $(k+1) \times (k+1)$ matrix $E(XX')$ is nonsingular. Then the best linear LS predictor that solves*

$$\min_{g \in \mathbb{A}} E[Y - g(X)]^2 = \min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2$$

is the linear function

$$g^*(X) = X'\beta^*,$$

where the optimizing coefficient vector

$$\beta^* = [E(XX')]^{-1} E(XY).$$

Proof: First, noting that

$$\min_{g \in \mathbb{A}} E[Y - g(X)]^2 = \min_{\beta \in \mathbb{R}^{k+1}} E(Y - \beta'X)^2,$$

we first find the FOC:

$$\frac{d}{d\beta} E(Y - X'\beta)^2|_{\beta = \beta^*} = 0.$$

The left hand side

$$\begin{aligned}\frac{d}{d\beta}E(Y-X'\beta)^2 &= E\left[\frac{\partial}{\partial\beta}(Y-X'\beta)^2\right] \\ &= E\left[2(Y-X'\beta)\frac{\partial}{\partial\beta}(-X'\beta)\right] \\ &= -2E\left[(Y-X'\beta)\frac{\partial}{\partial\beta}(X'\beta)\right] \\ &= -2E[X(Y-X'\beta)].\end{aligned}$$

Therefore, FOC implies that

$$\begin{aligned}E[X(Y-X'\beta^*)] &= 0 \text{ or} \\ E(XY) &= E(XX')\beta^*.\end{aligned}$$

Multiplying the inverse of $E(XX')$, we obtain

$$\beta^* = [E(XX')]^{-1}E(XY).$$

It remains to check SOC: The Hessian matrix

$$\frac{d^2}{d\beta d\beta'}E(Y-\beta'X)^2 = 2E(XX')$$

is positive definite provided $E(XX')$ is nonsingular (why?). Therefore, β^* is a global minimizer. This completes the proof.

Remarks:

The moment condition $E(Y^2) < \infty$ ensures that $E(Y|X)$ exists and is well-defined. When the $(k+1)\times(k+1)$ matrix

$$E(XX') = \begin{bmatrix} 1 & E(X_1) & E(X_2) & \cdots & E(X_k) \\ E(X_1) & E(X_1^2) & E(X_1X_2) & \cdots & E(X_2X_k) \\ E(X_2) & E(X_2X_1) & E(X_2^2) & \cdots & \\ \vdots & \vdots & & & \\ E(X_k) & E(X_kX_1) & & & E(X_k^2) \end{bmatrix}$$

is nonsingular and $E(XY)$ exists, the best linear LS approximation coefficient β^* is always well-defined, no matter whether $E(Y|X)$ is linear or nonlinear in X.

To gain insight into the nature of β^*, we consider a simple case where $\beta = (\beta_0, \beta_1)'$ and

$X = (1, X_1)'$. Then the slope coefficient and the intercept coefficient are, respectively,

$$\begin{aligned} \beta_1^* &= \frac{\text{cov}(Y, X_1)}{\text{var}(X_1)}, \\ \beta_0^* &= E(Y) - \beta_1^* E(X_1). \end{aligned}$$

Thus, the best linear LS approximation coefficient β_1^* is proportional to $\text{cov}(Y, X_1)$. In other words, β_1^* captures the dependence between Y and X_1 that is measurable by $\text{cov}(Y, X_1)$. It will miss the dependence between Y and X_1 that cannot be measured by $\text{cov}(Y, X_1)$. Therefore, linear regression analysis is essentially *correlation analysis*.

In general, the best linear LS predictor $g^*(X) \equiv X'\beta^* \neq E(Y|X)$. An important question is what happens if $g^*(X) = X'\beta^* \neq E(Y|X)$? In particular, what is the interpretation of β^*?

We now discuss the relationship between the best linear LS prediction and a linear regression model.

Definition [Linear Regression Model]: The specification

$$Y = X'\beta + u, \ \beta \in \mathbb{R}^{k+1},$$

is called a linear regression model, where u is the regression model disturbance or regression model error. If $k = 1$, it is called a bivariate linear regression model or a straight line regression model. If $k > 1$, it is called a multiple linear regression model.

The linear regression model is an artificial specification. Nothing ensures that the regression function is linear, namely $E(Y|X) = X'\beta^o$ for some β^o. In other words, the linear model may not contain the true regression function $g_o(X) \equiv E(Y|X)$. However, even if $g_o(X)$ is not a linear function of X, the linear regresson model $Y = X'\beta + u$ may still have some predictive ability although it is a misspecified model.

We first characterize the relationship between the best linear LS approximation and the linear regression model.

Theorem: *Suppose the conditions of the previous theorem hold. Let*

$$Y = X'\beta + u,$$

and let β^* be the best linear least square approximation coefficient. Then

$$\beta = \beta^*$$

if and only if the following orthogonality condition holds:

$$E(Xu) = 0.$$

Proof: From the linear regression model $Y = X'\beta + u$, we have $u = Y - X'\beta$, and so

$$E(Xu) = E(XY) - E(XX')\beta.$$

(a) Necessarity: If $\beta = \beta^*$, then

$$\begin{aligned} E(Xu) &= E(XY) - E(XX')\beta^* \\ &= E(XY) - E(XX')[E(XX')]^{-1}E(XY) \\ &= 0. \end{aligned}$$

(b) Sufficiency: If $E(Xu) = 0$, then

$$\begin{aligned} E(Xu) &= E(XY) - E(XX')\beta \\ &= 0. \end{aligned}$$

From this and the fact that $E(XX')$ is nonsingular, we have

$$\beta = [E(XX')]^{-1}E(XY) \equiv \beta^*.$$

This completes the proof.

Remarks:

This theorem implies that no matter whether $E(Y|X)$ is linear or nonlinear in X, we can always write

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

for some $\beta = \beta^*$ such that the orthogonality condition $E(Xu) = 0$ holds, where $u = Y - X'\beta^*$.

The orthogonality condition $E(Xu) = 0$ is fundamentally linked with the best least squares optimizer. If β is the best linear LS coefficient β^*, then the disturbance u must be orthogonal to X. On the other hand, if X is orthogonal to u, then β must be the least squares minimizer β^*. Essentially the orthogonality between X and ε is the FOC of the best linear LS problem! In other words, the orthoganality condition $E(Xu) = 0$ will always hold as long as the MSE ceriterion is used to obtain the best linear prediction. Note that when X contains an intercept, the orthogonality condition $E(Xu) = 0$ implies that $E(u) = 0$. In this case, we have $E(Xu) = \text{cov}(X, u)$. In other words, the orthogonality condition is equivalent to uncorrelatedness between X and u. This implies that u does not contain any component that can be predicted by a linear function

of X.

The condition $E(Xu) = 0$ is fundamentally different from $E(u|X) = 0$. The latter implies the former but not vice versa. In other words, $E(u|X) = 0$ implies $E(Xu) = 0$ but it is possible that $E(Xu) = 0$ and $E(u|X) \neq 0$. This can be illustrated by the following example.

Example: Suppose $u = (X^2 - 1) + \varepsilon$, where X and ε are independent N(0,1) random variables. Then

$$\begin{aligned} E(u|X) &= X^2 - 1 \neq 0, \text{ but} \\ E(Xu) &= E[X(X^2 - 1)] + E(X\varepsilon) \\ &= E(X^3) - E(X) + E(X)E(\varepsilon) \\ &= 0. \end{aligned}$$

2.4 Correct Model Specification for Conditional Mean

Question: What is the characterization for correct model specification in conditional mean?

Definition [Correct Model Specification in Mean]: *The linear regression model*

$$Y = X'\beta + u, \ \beta \in \mathbb{R}^{k+1},$$

is said to be correctly specified for $E(Y|X)$ *if*

$$E(Y|X) = X'\beta^o \text{ for some } \beta^o \in \mathbb{R}^{k+1}.$$

On the other hand, if

$$E(Y|X) \neq X'\beta \text{ for all } \beta \in \mathbb{R}^{k+1},$$

then the linear model is said to be misspecified for $E(Y|X)$.

Remarks:

The class of linear regression models contains an infinite number of linear functions, each corresponding to a particular value of β. When the linear model is correctly specified, a linear function corresponding to some β^o will coincide with $g_o(X)$. The coefficient β^o is called the "true parameter", because now it has a meaningful economic interpretation as the marginal effect of X on Y :

$$\beta^o = \frac{d}{dX}E(Y|X).$$

For example, when Y is comsumption and X is income, β^o is the marginal propensity to consume (MPC).

When β^o is a vector, the component

$$\beta_j^o = \frac{\partial E(Y|X)}{\partial X_j}, \qquad 1 \leq j \leq k,$$

is the partial marginal effect of X_j on Y when holding all other explanatory variables in X fixed.

Question: What is the interpretation of the intercept coefficient β_0^o when a linear regression model is correctly specified for $g_o(X)$?

Answer: The intercept β_0^o corresponds to the variable $X_0 = 1$, which is always uncorrelated with any other random variables. It captures the "average effect" on Y from all possible factors rather than the explanatory variables in X_t. For example, consider the standard Capital Asset Pricing Model (CAPM)

$$E(Y|X) = \beta_0^o + \beta_1^o X_1,$$

where Y is the excess portfolio return (i.e., the difference between a portfolio return and a risk-free rate) and X_1 is the excess market portfolio return (i.e., the difference between the market portfolio return and a risk-free rate). Here, β_0^o represents the average pricing error. When CAPM holds, $\beta_0^o = 0$. Thus, if the data generating process has $\beta_0^o > 0$, there exists underpricing for CAPM underprices the portfolio. If $\beta_0^o < 0$, CAPM overprices the portfolio.

No economic theory ensures that the functional form of $E(Y|X)$ must be linear in X. Non-linear functional form in X is a generic possibility. Therefore, we must be very cautious about the economic interpretation of linear coefficients.

Theorem: *If the linear model*

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

is correctly specified for $E(Y|X)$, *then*

(a) $Y = X'\beta^o + \varepsilon$ *for some* β^o *and* ε, *where* $E(\varepsilon|X) = 0$;

(b) $\beta^* = \beta^o$.

Proof: (a) If the linear model is correctly specified for $E(Y|X)$, then $E(Y|X) = X'\beta^o$ for some β^o.

On the other hand, we always have the regression identity $Y = E(Y|X)+\varepsilon$, where $E(\varepsilon|X) = 0$. Combining these two equations gives result (a) immediately.

(b) From part (a) we have

$$\begin{aligned} E(X\varepsilon) &= E[XE(\varepsilon|X)] \\ &= E(X \cdot 0) \\ &= 0. \end{aligned}$$

It follows that the orthogonality condition holds for $Y = X'\beta^o + \varepsilon$. Therefore, we have $\beta^* = \beta^o$ by the previous theorem (which one?).

Remarks:

Theorem (a) implies $E(Y|X) = X'\beta^o$ under correct model specification for $E(Y|X)$. This, together with Theorem (b), implies that when a linear regression model is correctly specified, the conditional mean $E(Y|X)$ will coincide with the best linear least squares predictor $g^*(X) = X'\beta^*$.

Under correct model specification, the best linear LS approximation coefficient β^* is equal to the true marginal effect parameter β^o. In other words, β^* can be interpreted as the true parameter β^o when (and only when) the linear regression model is correctly specified.

Question: What happens if the linear regression model

$$Y = X'\beta + u,$$

where $E(Xu) = 0$, is misspecified for $E(Y|X)$? In other words, what happens if $E(Xu) = 0$ but $E(u|X) \neq 0$?

Answer: The regression function

$$\begin{aligned} E(Y|X) &= X'\beta + E(u|X) \\ &\neq X'\beta. \end{aligned}$$

There exists some neglected structure in u that can be exploited to improve the prediction of Y using X. A misspecified model always yields suboptimal predictions. A correctly specified model yields optimal predictions in terms of MSE.

Example 3: Consider the following data generating process (DGP)

$$Y = 1 + \frac{1}{2}X_1 + \frac{1}{4}(X_1^2 - 1) + \varepsilon,$$

where X_1 and ε are mutually independent $N(0, 1)$.

(a) Find the conditional mean $E(Y|X_1)$ and $\frac{d}{dX_1}E(Y|X_1)$, the marginal effect of X_1 on Y.

Suppose now a linear regression model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + u \\ &= X'\beta + u, \end{aligned}$$

where $X = (X_0, X_1)' = (1, X_1)'$, is specified to approximate this DGP.

(b) Find the best LS approximation coefficient β^* and the best linear LS predictor $g^*_{\mathbb{A}}(X) = X'\beta^*$.

(c) Let $u = Y - X'\beta^*$. Show $E(Xu) = 0$.

(d) Check if the true marginal effect $\frac{d}{dX_1}E(Y|X_1)$ is equal to β^*_1, the model-implied marginal effect.

Solution: (a) Given that X_1 and u are independent, we obtain

$$\begin{aligned} E(Y|X_1) &= 1 + \frac{1}{2}X_1 + \frac{1}{4}(X_1^2 - 1), \\ \frac{d}{dX_1}E(Y|X_1) &= \frac{1}{2} + \frac{1}{2}X_1. \end{aligned}$$

(b) Using the best LS approximation formula, we have

$$\begin{aligned} \beta^* &= [E(XX')]^{-1} E(XY) \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}. \end{aligned}$$

Hence, we have

$$g^*(X) = X'\beta^* = 1 + \frac{1}{2}X_1.$$

(c) By definition and part (b), we have

$$\begin{aligned} u &= Y - X'\beta^* \\ &= Y - (\beta_0^* + \beta_1^* X_1) \\ &= \frac{1}{4}(X_1^2 - 1) + \varepsilon. \end{aligned}$$

It follows that

$$\begin{aligned} E(Xu) &= E\begin{bmatrix} 1 \cdot (\frac{1}{4}(X_1^2 - 1) + \varepsilon) \\ X_1 \cdot (\frac{1}{4}(X_1^2 - 1) + \varepsilon) \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \end{aligned}$$

although

$$E(u|X_1) = \frac{1}{4}(X_1^2 - 1) \neq 0.$$

(d) No, because

$$\frac{d}{dX_1}E(Y|X_1) = \frac{1}{2} + \frac{1}{2}X_1 \neq \beta_1^* = \frac{1}{2}.$$

The marginal effect depends on the level of X_1, rather than only on a constant. Therefore, the condition $E(Xu) = 0$ is not sufficient for the validity of the economic interpretation for β_1^* as the marginal effect.

Any parametric regression model is subject to potential model misspecification. This can occur due to the use of a misspecified functional form, as well as the existence of omitted variables which are correlated the existing regressors, among other things. In econometrics, there exists a modelling strategy which is free of model misspecification when a data set is sufficiently large. This modelling strategy is called a nonparametric approach, which does not assume any functional form for $E(Y|X)$ but let data speak for the true relation. We now introduce the basic idea of a nonparametric approach.

Nonparametric modelling is a statistical method that can model the unknown function arbitrarily well without having to know the functional form of $E(Y|X)$. To illustrate the basic idea of nonparametric modelling, suppose $g_o(x)$ is a smooth function of x. Then we can expand $g_o(x)$ using a set of orthonormal "basis" functions $\{\psi_j(x)\}_{j=0}^{\infty}$:

$$g_o(x) = \sum_{j=0}^{\infty} \beta_j \psi_j(x) \text{ for } x \in \text{support}(X),$$

where the Fourier coefficient

$$\beta_j = \int_{-\infty}^{\infty} g_o(x)\psi_j(x)dx$$

and

$$\int_{-\infty}^{\infty} \psi_i(x)\psi_j(x)dx = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The function δ_{ij} is called the Kronecker delta.

Example 1: Suppose $g_o(x) = x^2$ where $x \in [-\pi, \pi]$. Then

$$\begin{aligned} g_o(x) &= \frac{\pi^2}{3} - 4\left[\cos(x) - \frac{\cos(2x)}{2^2} + \frac{\cos(3x)}{3^2} - \cdots\right] \\ &= \frac{\pi^2}{3} - 4\sum_{j=1}^{\infty}(-1)^{j-1}\frac{\cos(jx)}{j^2}. \end{aligned}$$

Example 2: Suppose

$$g_o(x) = \begin{cases} -1 & \text{if } -\pi < x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } 0 < x < \pi. \end{cases}$$

Then

$$\begin{aligned} g_o(x) &= \frac{4}{\pi}\left[\sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \cdots\right] \\ &= \frac{4}{\pi}\sum_{j=0}^{\infty}\frac{\sin[(2j+1)x]}{(2j+1)}. \end{aligned}$$

Generally, suppose $g_o(x)$ is square-integrable. We have

$$\begin{aligned} \int_{-\pi}^{\pi} g_o^2(x)dx &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\beta_j\beta_k\int_{-\pi}^{\pi}\psi_j(x)\psi_k(x)dx \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\beta_j\beta_k\delta_{jk} \text{ by orthonormality of } \{\psi_j(\cdot)\} \\ &= \sum_{j=0}^{\infty}\beta_j^2 < \infty, \end{aligned}$$

Therefore, $\beta_j \to 0$ as $j \to \infty$. That is, the Fourier coefficient β_j will eventually vanish to zero as the order j goes to infinity. This motivates us to use the following truncated approximation:

$$g_p(x) = \sum_{j=0}^{p}\beta_j\psi_j(x),$$

where p is the order of bases. The approximation bias of $g_p(x)$ for $g_o(x)$ is

$$\begin{aligned} B_p(x) &= g_o(x) - g_p(x) \\ &= \sum_{j=p+1}^{\infty}\beta_j\psi_j(x) \\ &= \text{Bias}. \end{aligned}$$

The coefficients $\{\beta_j\}$ are unknown in practice, so we have to estimate them from data $\{Y_t, X_t\}_{t=1}^n$, where n is the sample size. We consider a linear regression

$$Y_t = \sum_{j=0}^{p}\beta_j\psi_j(X_t) + u_t, \qquad t = 1, ..., n.$$

Obviously, we need to let $p = p(n) \to \infty$ as $n \to \infty$ to ensure that the bias $B_p(x)$ vanishes to zero as $n \to \infty$. However, we should not let p grow to infinity too fast, because otherwise there will be too much sampling variation in parameter estimators (due to too many unknown parameters). This requires $p/n \to 0$ as $n \to \infty$.

The nonparametric approach just described is called **nonparametric series regression** (see, e.g., Andrews 1991, Hong and White 1995). There are many nonparametric methods available in the literature. Another popular nonparametric method is called **kernel method**, which is based on the idea of the Taylor series expansion in a local region. See Hardle (1990), *Applied Nonparametric Regression*, for more discussion on kernel smoothing. The key feature of nonparametric modelling is that it does not specify a concrete functional form or model but rather estimate the unknown true function from data. As can be seen above, nonparametric series regression is easy to use and understand, because it is a natural extension of linear regression with the number of resgressors increasing with the sample size n.

The nonparametric approach is flexible and powerful, but it generally requires a large data set for precise estimation because there is a large number of unknown parameters. Moreover, there is little economic interpretation for it (for example, it is difficult to give economic interpretation for the coefficients $\{\beta_j\}$). Nonparametric analysis is usually treated in a separate, more advanced econometric course (see more discussion in Chapter 10).

2.5 Summary and Conclusion

Most economic theories (e.g., rational expectations theory) have implications on and only on the conditional mean of the underlying economic variable given some suitable information set. The conditional mean $E(Y|X)$ is called the regression function of Y on X. In this chapter, we have shown that the regression function $E(Y|X)$ is the optimal solution to the MSE minimization problem

$$\min_{g \in \mathbb{F}} E[Y - g(X)]^2,$$

where $\mathbb{F}$ is the space of measurable and square-integrable functions.

The regression function $E(Y|X)$ is generally unknown, because economic theory usually does not tell a concrete functional form. In practice, one usually uses a parametric model for $E(Y|X)$ that has a known functional form but with a finite number of unknown parameters. When we restrict $g(X)$ to $\mathbb{A} = \{g : \mathbb{R}^K \to \mathbb{R} \mid g(x) = x'\beta\}$, a class of affine functions, the optimal predictor that solves

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

is $g^*(X) = X'\beta^*$, where

$$\beta^* = [E(XX')]^{-1} E(XY)$$

is called the best least squares approximation coefficient. The best linear least squares predictor $g_A^*(X) = X'\beta^*$ is always well-defined, no matter whether $E(Y|X)$ is linear in X.

Suppose we write

$$Y = X'\beta + u.$$

Then $\beta = \beta^*$ if and only if

$$E(Xu) = 0.$$

This orthogonality condition is actually the first order condition for the best linear least squares minimization problem. It does not guarantee correct specification of a linear regression model. A linear regression model is correctly specified for $E(Y|X)$ if $E(Y|X) = X'\beta^o$ for some β^o, which is equivalent to the condition that

$$E(u|X) = 0,$$

where $u = Y - X'\beta^o$. That is, correct model specification for $E(Y|X)$ holds if and only if the conditional mean of the linear regression model error is zero when evaluated at some parameter β^o. Note that $E(u|X) = 0$ is equivalent to the condition that $E[uh(X)] = 0$ for all measurable functions $h(\cdot)$. When $E(Y|X) = X'\beta^o$ for some β^o, we have $\beta^* = \beta^o$. That is, the best linear least squares approximation coefficient β^* will coincide with the true model parameter β^o and can be interpreted as the marginal effect of X on Y. The condition $E(u|X) = 0$ fundamentally differs from $E(Xu) = 0$. The former is crucial for validity of economic interpretation of the coefficient β^* as the true coefficient β^o. The orthogonality condition $E(Xu) = 0$ does not guarantee this interpretation. Correct model specification is important for economic interpretation of model coefficient and for optimal predictions.

An econometric model aims to provide a concise and reasonably accurate reflection of the data generating process. By disregarding less relevant aspects of the data, the model helps to obtain a better understanding of the main aspects of the DGP. This implies that an econometric model will never provide a completely accurate description of the DGP. Therefore, the concept of a "true model" does not make much practical sense. It reflects an idealized situation that allows us to obtain mathematically exact results. The idea is that similar results hold approximately true if the model is a reasonably accurate approximation of the DGP.

The main purpose of this chapter is to provide a general idea of regression analysis and to shed some light on the nature and limitation of linear regression models, which have been popularly used in econometrics and will be the subject of study in Chapters 3 to 7.

References

Andrews, D.W.K. (1991), *Econometrica*.

Cochrance, J. (2001), Asset Pricing. Princeton University Press: Princeton.

Härdle, W. (1990), Applied Nonparametric Regression. Cambridge University Press.
Hong, Y. and H. White (1995), Econometrica.
Sargent, T. and L. (2003),

EXERCISES

2.1. Put $\varepsilon = Y - E(Y|X)$. Show $\text{var}(Y|X) = \text{var}(\varepsilon|X)$.

2.2. Show $\text{var}(Y) = \text{var}[E(Y|X)] + \text{var}[Y - E(Y|X)]$.

2.3. Suppose (X, Y) follows a bivariate normal distribution with joint pdf

$$\begin{aligned} & f_{XY}(x,y) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}, \end{aligned}$$

where $-1 < \rho < 1, -\infty < \mu_1, \mu_2 < \infty, 0 < \sigma_1, \sigma_2 < \infty$. Find

(a) $E(Y|X)$.

(b) $\text{var}(Y|X)$. (Hint: Use the change of variable method for integration and the fact that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx = 1$.)

2.4. Suppose $Z \equiv (Y, X')'$ is a stochastic process such that the conditional mean $g_o(X) \equiv E(Y|X)$ exists, where X is a $(k+1) \times 1$ random vector. Suppose one uses a model (or a function) $g(X)$ to predict Y. A popular evaluation criterion for model $g(X)$ is the mean squared error $MSE(g) \equiv E[Y - g(X)]^2$.

(a) Show that the optimal predictor $g^*(X)$ for Y that minimizes $MSE(g)$ is the conditional mean $g_o(X)$; namely, $g^*(X) = g_o(X)$.

(b) Put $\varepsilon \equiv Y - g_o(X)$, which is called the true regression disturbance. Show that $E(\varepsilon|X) = 0$ and interpret this result.

2.5. The choices of model $g(X)$ in Exercise 2.4 are very general. Suppose that we now restrict our choice of $g(X)$ to a linear (or affine) models $\{g_{\mathbb{A}}(X) = X'\beta\}$, where β is a $(k+1) \times 1$ parameter. One can choose a linear function $g_{\mathbb{A}}(X)$ by choosing a value for parameter β. Different values of β give different linear functions $g_{\mathbb{A}}(X)$. The best linear predictor g_L^* that minimizes the mean squared error criterion is defined as $g_{\mathbb{A}}^*(X) \equiv X'\beta^*$, where

$$\beta^* \equiv \arg\min_{\beta \in \mathbb{R}^{k+1}} E(Y - X'\beta)^2$$

is called the optimal linear coefficient.

(a) Show that

$$\beta^* = [E(XX')]^{-1}E(XY).$$

(b) Define $u^* \equiv Y - X'\beta^*$. Show that $E(Xu^*) = 0$, where 0 is a $(k+1) \times 1$ zero vector.

(c) Suppose the conditional mean $g_o(X) = X'\beta^o$ for some given β^o. Then we say that the linear model $g_{\mathbb{A}}(X)$ is correctly specified for conditional mean $g_o(X)$, and β^o is the true parameter of the data generating process. Show that $\beta^* = \beta^o$ and $E(u^*|X) = 0$.

(d) Suppose the conditional mean $g_o(X) \neq X'\beta$ for any value of β. Then we say that the linear model $g_{\mathbb{A}}(X)$ is misspecified for conditional mean $g_o(X)$. Check if $E(u^*|X) = 0$ and discuss its implication.

2.6. Suppose $Y = \beta_0^* + \beta_1^* X_1 + u$, where Y and X_1 are scalars, and $\beta^* = (\beta_0^*, \beta_1^*)'$ is the best linear least squares approximation coefficient.

(a) Show that $\beta_1^* = \text{cov}(Y, X_1)/\sigma_{X_1}^2$ and $\beta_0^* = E(Y) - \beta_1^* E(X)$, and the mean squared error

$$E[Y - (\beta_0^* + \beta_1^* X_1)]^2 = \sigma_Y^2(1 - \rho_{X_1 Y}^2),$$

where $\sigma_Y^2 = \text{var}(Y)$ and $\rho_{X_1 Y}$ is the correlation coefficient between Y and X_1.

(b) Suppose in addition Y and X_1 follow a bivariate normal distribution. Show $E(Y|X_1) = \beta_0^* + \beta_1^* X_1$ and $\text{var}(Y|X_1) = \sigma_Y^2(1 - \rho_{X_1 Y}^2)$. That is, the conditional mean of Y given X_1 coincides with the best linear least squares predictor and the conditional variance of Y given X_1 is equal to the mean squared error of the best linear least squares predictor.

2.7. Suppose

$$Y = \beta_0 + \beta_1 X_1 + |X_1|\varepsilon,$$

where $E(X_1) = 0$, $\text{var}(X_1) = \sigma_{X_1}^2 > 0$, $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma_\varepsilon^2 > 0$, and ε and X_1 are independent. Both β_0 and β_1 are scalar constants.

(a) Find $E(Y|X_1)$.

(b) Find $\text{var}(Y|X_1)$.

(c) Show that $\beta_1 = 0$ if and only if $\text{cov}(X_1, Y) = 0$.

2.8. Suppose an aggregate consumption function is given by

$$Y = 1 + 0.5X_1 + \frac{1}{4}(X_1^2 - 1) + \varepsilon,$$

where $X_1 \sim N(0, 1), \varepsilon \sim N(0, 1)$, and X_1 is independent of ε.

(a) Find the conditional mean $g_o(X) \equiv E(Y|X)$, where $X \equiv (1, X_1)'$.

(b) Find the marginal propensity to consume (MPC) $\frac{d}{dX_1} g_o(X)$.

(c) Suppose we use a linear model

$$Y = X'\beta + u = \beta_0 + \beta_1 X_1 + u$$

where $\beta \equiv (\beta_0, \beta_1)'$ to predict Y. Find the optimal linear coefficient β^* and the optimal linear predictor $g_{\mathbb{A}}^*(X) \equiv X'\beta^*$.

(d) Compute the partial derivative of the linear model $\frac{d}{dX_1}g^*_{\mathbb{A}}(X)$, and compare it with the MPC in part (b). Discuss the results you obtain.

2.9. Put $g_o(X) = E(Y|X)$,where $X = (1, X_1)'$. Then we have

$$Y = g_o(X) + \varepsilon,$$

where $E(\varepsilon|X) = 0$.

Consider a first order Taylor series expansion of $g_o(X)$ around $\mu_1 = E(X_1)$:

$$\begin{aligned} g_o(X) &\approx g_o(\mu_1) + g_o'(\mu_1)(X_1 - \mu_1) \\ &= [g_o(\mu_1) - \mu g_o'(\mu_1)] + g_o'(\mu_1)X_1. \end{aligned}$$

Suppose $\beta^* = (\beta_0^*, \beta_1^*)'$ is the best linear least squares approximation coefficient. Is it true that $\beta_1^* = g_o'(\mu_1)$? Provide your reasoning.

2.10. Suppose a data generating process is given by

$$Y = 0.8X_1X_2 + \varepsilon,$$

where $X_1 \sim N(0,1), X_2 \sim N(0,1), \varepsilon \sim N(0,1)$, and X_1, X_2 and ε are mutually independent. Put $X = (1, X_1, X_2)'$.

(a) Is Y predictable in mean using information X?

(b) Suppose we use a linear model

$$\begin{aligned} g_{\mathbb{A}}(X) &= X'\beta + u \\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 + u \end{aligned}$$

to predict Y. Does this linear model has any predicting power? Explain.

2.11. Show that $E(u|X) = 0$ if and only if $E[h(X)u] = 0$ for any measurable functions $h(\cdot)$.

2.13. Suppose $E(u|X)$ exists, X is a bounded random variable, and $h(X)$ is an arbitrary measurable function. Put $g(X) = E(\varepsilon|X)$ and assume that $E[g^2(X)] < \infty$.

(a) Show that if $g(X) = 0$, then $E[\varepsilon h(X)] = 0$.

(b) Show that if $E[\varepsilon h(X)] = 0$, then $E(\varepsilon|X) = 0$. [Hint: Consider $h(X) = e^{tX}$ for t in a small neighborhood containing 0. Given that X is bounded, we can expand

$$g(X) = \sum_{j=0}^{\infty} \beta_j X^j$$

where $\beta_j = \int_{-\infty}^{\infty} g(x)x^j f_X(x)dx$ is the Fourier coefficient. Then

$$\begin{aligned} E(\varepsilon e^{tX}) &= E\left[E(\varepsilon|X)e^{tX}\right] \\ &= E\left[g(X)e^{tX}\right] \\ &= \sum_{j=0}^{\infty} \frac{t^j}{j!} E\left[g(X)X^j\right] \\ &= \sum_{j=0}^{\infty} \frac{t^j}{j!}\beta_j \end{aligned}$$

for all t in a small neighborhood containing 0.]

2.14. Consider the following nonlinear least squares problem

$$\min_{\beta \in \mathbf{R}^{k+1}} E\left[Y - g(X,\beta)\right]^2,$$

where $g(X,\beta)$ is possibly a nonlinear function of β. [An example is a logistic regression model where $g(X,\beta) = \frac{1}{1+\exp(-X'\beta)}$.] Suppose $E\left[\frac{\partial}{\partial\beta}g(X,\beta)\frac{\partial}{\partial\beta'}g(X,\beta)\right]$ is a $(k+1)\times(k+1)$ bounded and nonsingular matrix for all $\beta \in \mathbf{R}^{k+1}$, where $\frac{\partial}{\partial\beta'}g(X,\beta)$ is the transpose of the $(k+1)\times 1$ column vector $\frac{\partial}{\partial\beta}g(X,\beta)$.

(a) Derive the first order condition for the best nonlinear least squares approximation coefficient β^* (say).

(b) Put $Y = g(X,\beta) + u$. Show that $\beta = \beta^*$ if and only if $E[u\frac{\partial}{\partial\beta}g(X,\beta^*)] = 0$. Do we have $E(Xu) = 0$ when $g(X,\beta)$ is nonlinear in β?

(c) The nonlinear regression model $g(X,\beta)$ is said to be correctly specified for $E(Y|X)$ if there exists some unknown β^o such that $E(Y|X) = g(X,\beta^o)$ almost surely. Here, β^o can be interpreted as a true model parameter. Show that $\beta^* = \beta^o$ if and only if the model $g(X,\beta)$ is correctly specified for $E(Y|X)$.

(d) Do we have $E(u|X) = 0$, where $u = Y - g(X,\beta^o)$, for some β^o, when the model $g(X,\beta)$ is correctly specified?

(e) If $E(u|X) = 0$, where $u = Y - g(X,\beta^o)$ for some β^o, is $g(X,\beta)$ correctly specified for $E(Y|X)$?

2.15. Comment on the following statement: "All econometric models are approximations of the economic system of interest and are therefore misspecified. Therefore, there is no need to check correct model specification in practice."