Contents

Chapter 1

Introduction

1.1 What is Econometrics?

The term "econometrics" is believed to have been crafted by Ragnar Frisch (1895-1973) of Norway, one of the three principal founders of the Econometric Society, first editor of the journal *Econometrica*, and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch's own words in the introduction to the first issue of *Econometrica* to describe the discipline.

> A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the [Econometric] Society, in Section I of the Constitution, which reads: "The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems...."
>
> But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a defininitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics. Experience has shown that each of these three view-points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.
>
> Ragnar Frisch, *Econometrica*, (1933), 1, pp. 1-2.

This definition remains valid today, although some terms have evolved somewhat in their usage. Today, we would say that econometrics is the unified study of economic models, mathematical statistics, and economic data.

Within the field of econometrics there are sub-divisions and specializations. **Econometric theory** concerns the development of tools and methods, and the study of the properties of econometric methods. **Applied econometrics** is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

1.2 The Probability Approach to Econometrics

The unifying methodology of modern econometrics was articulated by Trygve Haavelmo (1911-1999) of Norway, winner of the 1989 Nobel Memorial Prize in Economic Sciences, in his seminal

paper "The probability approach in econometrics", *Econometrica* (1944). Haavelmo argued that quantitative economic models must necessarily be *probability models* (by which today we would mean *stochastic*). Deterministic models are blatently inconsistent with observed economic quantities, and it is incoherent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics. The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

Haavelmo's probability approach was quickly embraced by the economics profession. Today no quantitative work in economics shuns its fundamental vision.

While all economists embrace the probability approach, there has been some evolution in its implementation.

The **structural approach** is the closest to Haavelmo's original idea. A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified. Researchers often describe this as "taking their model seriously." The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.

A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified. Rather, it is more accurate to view a model as a useful abstraction or approximation. In this case, how should we interpret structural econometric analysis? The **quasi-structural approach** to inference views a structural economic model as an approximation rather than the truth. This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

Closely related is the **semiparametric approach**. A probabilistic economic model is partially specified but some features are left unspecified. This approach typically leads to estimation methods such as least-squares and the Generalized Method of Moments. The semiparametric approach dominates contemporary econometrics, and is the main focus of this textbook.

Another branch of quantitative structural economics is the **calibration approach.** Similar to the quasi-structural approach, the calibration approach interprets structural models as approximations and hence inherently false. The difference is that the calibrationist literature rejects mathematical statistics (deeming classical theory as inappropriate for approximate models) and instead selects parameters by matching model and data moments using non-statistical *ad hoc*[1] methods.

1.3 Econometric Terms and Notation

In a typical application, an econometrician has a set of repeated measurements on a set of variables. For example, in a labor application the variables could include weekly earnings, educational attainment, age, and other descriptive characteristics. We call this information the **data, dataset**, or **sample**.

We use the term **observations** to refer to the distinct repeated measurements on the variables. An individual observation often corresponds to a specific economic unit, such as a person, household, corporation, firm, organization, country, state, city or other geographical region. An individual observation could also be a measurement at a point in time, such as quarterly GDP or a daily interest rate.

[1] *Ad hoc* means "for this purpose" – a method designed for a specific problem – and not based on a generalizable principle.

Economists typically denote variables by the italicized roman characters y, x, and/or z. The convention in econometrics is to use the character y to denote the variable to be explained, while the characters x and z are used to denote the conditioning (explaining) variables.

Following mathematical convention, real numbers (elements of the real line $\mathbb{R}$, also called **scalars**) are written using lower case italics such as y, and vectors (elements of $\mathbb{R}^k$) by lower case bold italics such as $\boldsymbol{x}$, e.g.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}.$$

Upper case bold italics such as $\boldsymbol{X}$ are used for matrices.

We denote the number of observations by the natural number n, and subscript the variables by the index i to denote the individual observation, e.g. y_i, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. In some contexts we use indices other than i, such as in time-series applications where the index t is common and T is used to denote the number of observations. In panel studies we typically use the double index it to refer to individual i at a time period t.

The i^{th} **observation** is the set $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. The **sample** is the set $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i) : i = 1, ..., n\}$.

It is proper mathematical practice to use upper case X for random variables and lower case x for realizations or specific values. Since we use upper case to denote matrices, the distinction between random variables and their realizations is not rigorously followed in econometric notation. Thus the notation y_i will in some places refer to a random variable, and in other places a specific realization. This is an undesirable but there is little to be done about it without terrifically complicating the notation. Hopefully there will be no confusion as the use should be evident from the context.

We typically use Greek letters such as β, θ and σ^2 to denote unknown parameters of an econometric model, and will use boldface, e.g. $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$, when these are vector-valued. Estimates are typically denoted by putting a hat "^", tilde "~" or bar "-" over the corresponding letter, e.g. $\hat{\beta}$ and $\tilde{\beta}$ are estimates of β.

The covariance matrix of an econometric estimator will typically be written using the capital boldface $\boldsymbol{V}$, often with a subscript to denote the estimator, e.g. $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \text{var}\left(\widehat{\boldsymbol{\beta}}\right)$ as the covariance matrix for $\widehat{\boldsymbol{\beta}}$. Hopefully without causing confusion, we will use the notation $\boldsymbol{V}_{\boldsymbol{\beta}} = \text{avar}(\widehat{\boldsymbol{\beta}})$ to denote the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ (the variance of the asymptotic distribution). Estimates will be denoted by appending hats or tildes, e.g. $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{V}_{\boldsymbol{\beta}}$.

1.4 Observational Data

A common econometric question is to quantify the impact of one set of variables on another variable. For example, a concern in labor economics is the returns to schooling – the change in earnings induced by increasing a worker's education, holding other variables constant. Another issue of interest is the earnings gap between men and women.

Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children's wage path after they mature and enter the labor force. The differences between the groups would be direct measurements of the effects of different levels of education. However, experiments such as this would be widely

condemned as immoral! Consequently, in economics non-laboratory experimental data sets are typically narrow in scope.

Instead, most economic data is **observational**. To continue the above example, through data collection we can record the level of a person's education and their wage. With such data we can measure the joint distribution of these variables, and assess the joint dependence. But from observational data it is difficult to infer **causality**, as we are not able to manipulate one variable to see the direct effect on the other. For example, a person's level of education is (at least partially) determined by that person's choices. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability, which suggests a high relative wage. This is an alternative explanation for an observed positive correlation between educational levels and wages. High ability individuals do better in school, and therefore choose to attain higher levels of education, and their high ability is the fundamental reason for their high wages. The point is that multiple explanations are consistent with a positive correlation between schooling levels and education. Knowledge of the joint distibution alone may not be able to distinguish between these explanations.

> Most economic data sets are observational, not experimental. This means that all variables must be treated as random and possibly jointly determined.

This discussion means that it is difficult to infer causality from observational data alone. Causal inference requires identification, and this is based on strong assumptions. We will discuss these issues on occasion throughout the text.

1.5 Standard Data Structures

There are three major types of economic data sets: cross-sectional, time-series, and panel. They are distinguished by the dependence structure across observations.

Cross-sectional data sets have one observation per individual. Surveys are a typical source for cross-sectional data. In typical applications, the individuals surveyed are persons, households, firms or other economic agents. In many contemporary econometric cross-section studies the sample size n is quite large. It is conventional to assume that cross-sectional observations are mutually independent. Most of this text is devoted to the study of cross-section data.

Time-series data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence so the random sampling assumption is inappropriate. Most aggregate economic data is only available at a low frequency (annual, quarterly or perhaps monthly) so the sample size is typically much smaller than in cross-section studies. The exception is financial data where data are available at a high frequency (weekly, daily, hourly, or by transaction) so sample sizes can be quite large.

Panel data combines elements of cross-section and time-series. These data sets consist of a set of individuals (typically persons, households, or corporations) surveyed repeatedly over time. The common modeling assumption is that the individuals are mutually independent of one another, but a given individual's observations are mutually dependent. This is a modified random sampling environment.

Data Structures

- Cross-section
- Time-series
- Panel

Many contemporary econometric applications combine elements of cross-section, time-series, and panel data modeling. These include models of spatial correlation and clustering.

As we mentioned above, most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations. By mutual independence we mean that the i^{th} observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ is independent of the j^{th} observation $(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j)$ for $i \neq j$. (Sometimes the label "independent" is misconstrued. It is a statement about the relationship between observations i and j, not a statement about the relationship between y_i and $\boldsymbol{x}_i$ and/or $\boldsymbol{z}_i$.)

Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a random draw from the same probability distribution. In this case we say that the data are **independent and identically distributed** or **iid**. We call this a **random sample**. For most of this text we will assume that our observations come from a random sample.

Definition 1.5.1 *The observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are a **random sample** if they are mutually independent and identically distributed (**iid**) across $i = 1, ..., n$.*

In the random sampling framework, we think of an individual observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ as a realization from a joint probability distribution $F(y, \boldsymbol{x}, \boldsymbol{z})$ which we can call the **population**. This "population" is infinitely large. This abstraction can be a source of confusion as it does not correspond to a physical population in the real world. It is an abstraction since the distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample. The *assumption* of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

The random sampling framework was a major intellectural breakthrough of the late 19th century, allowing the application of mathematical statistics to the social sciences. Before this conceptual development, methods from mathematical statistics had not been applied to economic data as the latter was viewed as non-random. The random sampling framework enabled economic samples to be treated as random, a necessary precondition for the application of statistical methods.

1.6 Sources for Economic Data

Fortunately for economists, the internet provides a convenient forum for dissemination of economic data. Many large-scale economic datasets are available without charge from governmental agencies. An excellent starting point is the Resources for Economists Data Links, available at `rfe.org`. From this site you can find almost every publically available economic data set. Some specific data sources of interest include

- Bureau of Labor Statistics
- US Census

- Current Population Survey
- Survey of Income and Program Participation
- Panel Study of Income Dynamics
- Federal Reserve System (Board of Governors and regional banks)
- National Bureau of Economic Research
- U.S. Bureau of Economic Analysis
- CompuStat
- International Financial Statistics

Another good source of data is from authors of published empirical studies. Most journals in economics require authors of published papers to make their datasets generally available. For example, in its instructions for submission, *Econometrica* states:

> *Econometrica* has the policy that all empirical, experimental and simulation results must be replicable. Therefore, authors of accepted papers must submit data sets, programs, and information on empirical analysis, experiments and simulations that are needed for replication and some limited sensitivity analysis.

The *American Economic Review* states:

> All data used in analysis must be made available to any researcher for purposes of replication.

The *Journal of Political Economy* states:

> It is the policy of the *Journal of Political Economy* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication.

If you are interested in using the data from a published paper, first check the journal's website, as many journals archive data and replication programs online. Second, check the website(s) of the paper's author(s). Most academic economists maintain webpages, and some make available replication files complete with data and programs. If these investigations fail, email the author(s), politely requesting the data. You may need to be persistent.

As a matter of professional etiquette, all authors absolutely have the obligation to make their data and programs available. Unfortunately, many fail to do so, and typically for poor reasons. The irony of the situation is that it is typically in the best interests of a scholar to make as much of their work (including all data and programs) freely available, as this only increases the likelihood of their work being cited and having an impact.

Keep this in mind as you start your own empirical project. Remember that as part of your end product, you will need (and want) to provide all data and programs to the community of scholars. The greatest form of flattery is to learn that another scholar has read your paper, wants to extend your work, or wants to use your empirical methods. In addition, public openness provides a healthy incentive for transparency and integrity in empirical analysis.

1.7 Econometric Software

Economists use a variety of econometric, statistical, and programming software.

STATA (www.stata.com) is a powerful statistical program with a broad set of pre-programmed econometric and statistical tools. It is quite popular among economists, and is continuously being updated with new methods. It is an excellent package for most econometric analysis, but is limited when you want to use new or less-common econometric methods which have not yet been programed.

R (www.r-project.org), GAUSS (www.aptech.com), MATLAB (www.mathworks.com), and Ox (www.oxmetrics.net) are high-level matrix programming languages with a wide variety of built-in statistical functions. Many econometric methods have been programed in these languages and are available on the web. The advantage of these packages is that you are in complete control of your analysis, and it is easier to program new methods than in STATA. Some disadvantages are that you have to do much of the programming yourself, programming complicated procedures takes significant time, and programming errors are hard to prevent and difficult to detect and eliminate. Of these languages, Gauss used to be quite popular among econometricians, but currently Matlab is more popular. A smaller but growing group of econometricians are enthusiastic fans of R, which of these languages is uniquely open-source, user-contributed, and best of all, completely free!

For highly-intensive computational tasks, some economists write their programs in a standard programming language such as Fortran or C. This can lead to major gains in computational speed, at the cost of increased time in programming and debugging.

As these different packages have distinct advantages, many empirical economists end up using more than one package. As a student of econometrics, you will learn at least one of these packages, and probably more than one.

1.8 Reading the Manuscript

I have endeavored to use a unified notation and nomenclature. The development of the material is cumulative, with later chapters building on the earlier ones. Never-the-less, every attempt has been made to make each chapter self-contained, so readers can pick and choose topics according to their interests.

To fully understand econometric methods, it is necessary to have a mathematical understanding of its mechanics, and this includes the mathematical proofs of the main results. Consequently, this text is self-contained, with nearly all results proved with full mathematical rigor. The mathematical development and proofs aim at brevity and conciseness (sometimes described as mathematical elegance), but also at pedagogy. To understand a mathematical proof, it is not sufficient to simply *read* the proof, you need to follow it, and re-create it for yourself.

Never-the-less, many readers will not be interested in each mathematical detail, explanation, or proof. This is okay. To use a method it may not be necessary to understand the mathematical details. Accordingly I have placed the more technical mathematical proofs and details in chapter appendices. These appendices and other technical sections are marked with an asterisk (*). These sections can be skipped without any loss in exposition.

1.9 Common Symbols

y	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matrix
$\mathbb{R}$	real line
$\mathbb{R}^k$	Euclidean k space
$\mathbb{E}(y)$	mathematical expectation
$\text{var}(y)$	variance
$\text{cov}(x, y)$	covariance
$\text{var}(\boldsymbol{x})$	covariance matrix
$\text{corr}(x, y)$	correlation
Pr	probability
$\longrightarrow$	limit
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
$\text{plim}_{n\to\infty}$	probability limit
$\text{N}(\mu, \sigma^2)$	normal distribution
$\text{N}(0, 1)$	standard normal distribution
χ^2_k	chi-square distribution with k degrees of freedom
$\boldsymbol{I}_n$	identity matrix
$\text{tr}\,\boldsymbol{A}$	trace
$\boldsymbol{A}'$	matrix transpose
$\boldsymbol{A}^{-1}$	matrix inverse
$\boldsymbol{A} > 0$	positive definite
$\boldsymbol{A} \geq 0$	positive semi-definite
$\|\boldsymbol{a}\|$	Euclidean norm
$\|\boldsymbol{A}\|$	matrix (Frobinius) norm
$\approx$	approximate equality
$\stackrel{def}{=}$	definitional equality
$\sim$	is distributed as
log	natural logarithm

Chapter 2

Conditional Expectation and Projection

2.1 Introduction

The most commonly applied econometric tool is least-squares estimation, also known as **regression**. As we will see, least-squares is a tool to estimate an approximate conditional mean of one variable (the **dependent variable**) given another set of variables (the **regressors**, **conditioning variables**, or **covariates**).

In this chapter we abstract from estimation, and focus on the probabilistic foundation of the conditional expectation model and its projection approximation.

2.2 The Distribution of Wages

Suppose that we are interested in wage rates in the United States. Since wage rates vary across workers, we cannot describe wage rates by a single number. Instead, we can describe wages using a probability distribution. Formally, we view the wage of an individual worker as a random variable *wage* with the **probability distribution**

$$F(u) = \Pr(wage \leq u).$$

When we say that a person's wage is random we mean that we do not know their wage before it is measured, and we treat observed wage rates as realizations from the distribution F. Treating unobserved wages as random variables and observed wages as realizations is a powerful mathematical abstraction which allows us to use the tools of mathematical probability.

A useful thought experiment is to imagine dialing a telephone number selected at random, and then asking the person who responds to tell us their wage rate. (Assume for simplicity that all workers have equal access to telephones, and that the person who answers your call will respond honestly.) In this thought experiment, the wage of the person you have called is a single draw from the distribution F of wages in the population. By making many such phone calls we can learn the distribution F of the entire population.

When a distribution function F is differentiable we define the probability density function

$$f(u) = \frac{d}{du}F(u).$$

The density contains the same information as the distribution function, but the density is typically easier to visually interpret.

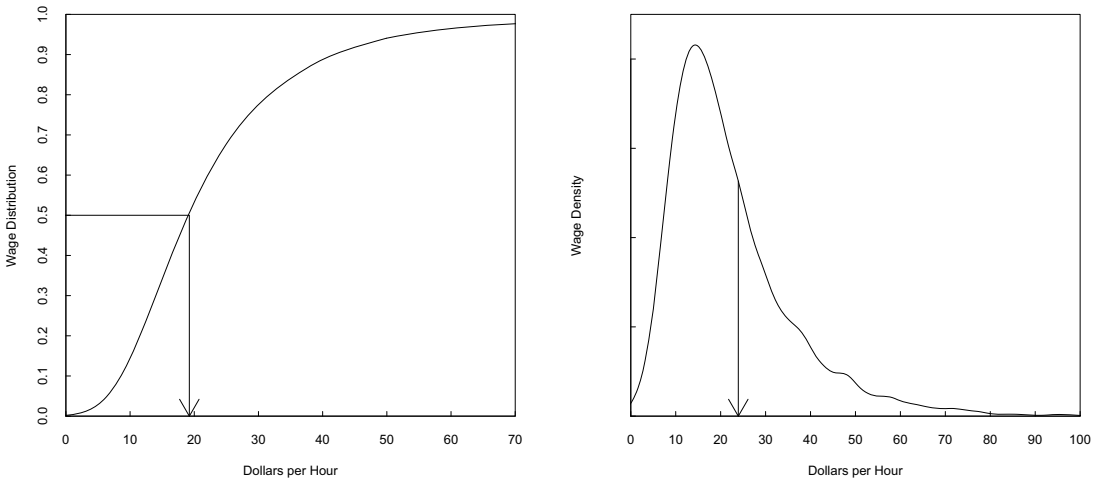


Figure 2.1: Wage Distribution and Density. All full-time U.S. workers

In Figure 2.1 we display estimates[1] of the probability distribution function (on the left) and density function (on the right) of U.S. wage rates in 2009. We see that the density is peaked around \$15, and most of the probability mass appears to lie between \$10 and \$40. These are ranges for typical wage rates in the U.S. population.

Important measures of central tendency are the median and the mean. The **median** m of a continuous[2] distribution F is the unique solution to

$$F(m) = \frac{1}{2}.$$

The median U.S. wage (\$19.23) is indicated in the left panel of Figure 2.1 by the arrow. The median is a robust[3] measure of central tendency, but it is tricky to use for many calculations as it is not a linear operator.

The **expectation** or **mean** of a random variable y with density f is

$$\mu = \mathbb{E}(y) = \int_{-\infty}^{\infty} u f(u) du.$$

Here we have used the common and convenient convention of using the single character y to denote a random variable, rather than the more cumbersome label *wage*. A general definition of the mean is presented in Section 2.31. The mean U.S. wage (\$23.90) is indicated in the right panel of Figure 2.1 by the arrow.

We sometimes use the notation the notation $\mathbb{E}y$ instead of $\mathbb{E}(y)$ when the variable whose expectation is being taken is clear from the context. There is no distinction in meaning.

The mean is a convenient measure of central tendency because it is a linear operator and arises naturally in many economic models. A disadvantage of the mean is that it is not robust[4] especially in the presence of substantial skewness or thick tails, which are both features of the wage

[1] The distribution and density are estimated nonparametrically from the sample of 50,742 full-time non-military wage-earners reported in the March 2009 Current Population Survey. The wage rate is constructed as annual individual wage and salary earnings divided by hours worked.

[2] If F is not continuous the definition is $m = \inf\{u : F(u) \geq \frac{1}{2}\}$

[3] The median is not sensitive to pertubations in the tails of the distribution.

[4] The mean is sensitive to pertubations in the tails of the distribution.

distribution as can be seen easily in the right panel of Figure 2.1. Another way of viewing this is that 64% of workers earn less that the mean wage of $23.90, suggesting that it is incorrect to describe the mean as a "typical" wage rate.

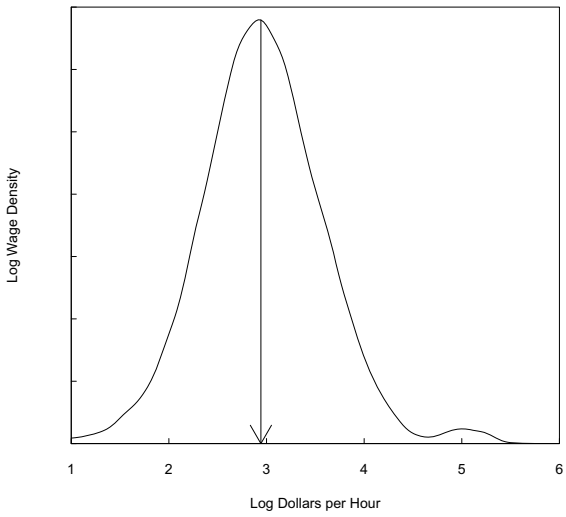


Figure 2.2: Log Wage Density

In this context it is useful to transform the data by taking the natural logarithm[5]. Figure 2.2 shows the density of log hourly wages $\log(wage)$ for the same population, with its mean 2.95 drawn in with the arrow. The density of log wages is much less skewed and fat-tailed than the density of the level of wages, so its mean

$$\mathbb{E}(\log(wage)) = 2.95$$

is a much better (more robust) measure[6] of central tendency of the distribution. For this reason, wage regressions typically use log wages as a dependent variable rather than the level of wages.

Another useful way to summarize the probability distribution $F(u)$ is in terms of its quantiles. For any $\alpha \in (0,1)$, the α^{th} quantile of the continuous[7] distribution F is the real number q_α which satisfies

$$F(q_\alpha) = \alpha.$$

The quantile function q_α, viewed as a function of α, is the inverse of the distribution function F. The most commonly used quantile is the median, that is, $q_{0.5} = m$. We sometimes refer to quantiles by the percentile representation of α, and in this case they are often called percentiles, e.g. the median is the 50^{th} percentile.

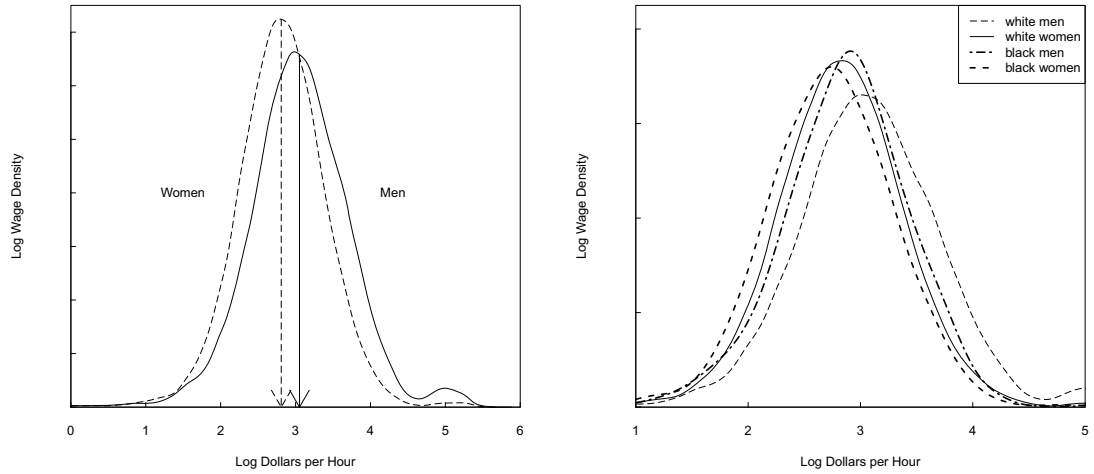
2.3 Conditional Expectation

We saw in Figure 2.2 the density of log wages. Is this distribution the same for all workers, or does the wage distribution vary across subpopulations? To answer this question, we can compare wage distributions for different groups – for example, men and women. The plot on the left in Figure 2.3 displays the densities of log wages for U.S. men and women with their means (3.05 and 2.81) indicated by the arrows. We can see that the two wage densities take similar shapes but the density for men is somewhat shifted to the right with a higher mean.

[5] Throughout the text, we will use $\log(y)$ or $\log y$ to denote the natural logarithm of y.

[6] More precisely, the geometric mean $\exp(\mathbb{E}(\log w)) = \$19.11$ is a robust measure of central tendency.

[7] If F is not continuous the definition is $q_\alpha = \inf\{u : F(u) \geq \alpha\}$



(a) Women and Men

(b) By Sex and Race

Figure 2.3: Log Wage Density by Sex and Race

The values 3.05 and 2.81 are the mean log wages in the subpopulations of men and women workers. They are called the **conditional means** (or **conditional expectations**) of log wages given sex. We can write their specific values as

$$\mathbb{E}\left(\log(wage) \mid sex = man\right) = 3.05 \tag{2.1}$$

$$\mathbb{E}\left(\log(wage) \mid sex = woman\right) = 2.81. \tag{2.2}$$

We call these means *conditional* as they are conditioning on a fixed value of the variable *sex*. While you might not think of a person's sex as a random variable, it is random from the viewpoint of econometric analysis. If you randomly select an individual, the sex of the individual is unknown and thus random. (In the population of U.S. workers, the probability that a worker is a woman happens to be 43%.) In observational data, it is most appropriate to view all measurements as random variables, and the means of subpopulations are then conditional means.

As the two densities in Figure 2.3 appear similar, a hasty inference might be that there is not a meaningful difference between the wage distributions of men and women. Before jumping to this conclusion let us examine the differences in the distributions of Figure 2.3 more carefully. As we mentioned above, the primary difference between the two densities appears to be their means. This difference equals

$$\begin{aligned}\mathbb{E}\left(\log(wage) \mid sex = man\right) - \mathbb{E}\left(\log(wage) \mid sex = woman\right) &= 3.05 - 2.81 \\ &= 0.24 \end{aligned} \tag{2.3}$$

A difference in expected log wages of 0.24 implies an average 24% difference between the wages of men and women, which is quite substantial. (For an explanation of logarithmic and percentage differences see Section 2.4.)

Consider further splitting the men and women subpopulations by race, dividing the population into whites, blacks, and other races. We display the log wage density functions of four of these groups on the right in Figure 2.3. Again we see that the primary difference between the four density functions is their central tendency.

	men	women
white	3.07	2.82
black	2.86	2.73
other	3.03	2.86

Table 2.1: Mean Log Wages by Sex and Race

Focusing on the means of these distributions, Table 2.1 reports the mean log wage for each of the six sub-populations.

The entries in Table 2.1 are the conditional means of $\log(wage)$ given *sex* and *race*. For example

$$\mathbb{E}\left(\log(wage) \mid sex = man,\ race = white\right) = 3.07$$

and

$$\mathbb{E}\left(\log(wage) \mid sex = woman,\ race = black\right) = 2.73$$

One benefit of focusing on conditional means is that they reduce complicated distributions to a single summary measure, and thereby facilitate comparisons across groups. Because of this simplifying property, conditional means are the primary interest of regression analysis and are a major focus in econometrics.

Table 2.1 allows us to easily calculate average wage differences between groups. For example, we can see that the wage gap between men and women continues after disaggregation by race, as the average gap between white men and white women is 25%, and that between black men and black women is 13%. We also can see that there is a race gap, as the average wages of blacks are substantially less than the other race categories. In particular, the average wage gap between white men and black men is 21%, and that between white women and black women is 9%.

2.4 Log Differences*

A useful approximation for the natural logarithm for small x is

$$\log(1 + x) \approx x. \tag{2.4}$$

This can be derived from the infinite series expansion of $\log(1 + x)$:

$$\begin{aligned} \log(1 + x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\ &= x + O(x^2). \end{aligned}$$

The symbol $O(x^2)$ means that the remainder is bounded by Ax^2 as $x \to 0$ for some $A < \infty$. A plot of $\log(1 + x)$ and the linear approximation x is shown in Figure 2.4. We can see that $\log(1 + x)$ and the linear approximation x are very close for $|x| \leq 0.1$, and reasonably close for $|x| \leq 0.2$, but the difference increases with $|x|$.

Now, if y^* is $c\%$ greater than y, then

$$y^* = (1 + c/100)y.$$

Taking natural logarithms,

$$\log y^* = \log y + \log(1 + c/100)$$

or

$$\log y^* - \log y = \log(1 + c/100) \approx \frac{c}{100}$$

where the approximation is (2.4). This shows that 100 multiplied by the difference in logarithms is approximately the percentage difference between y and y^*, and this approximation is quite good for $|c| \leq 10$.

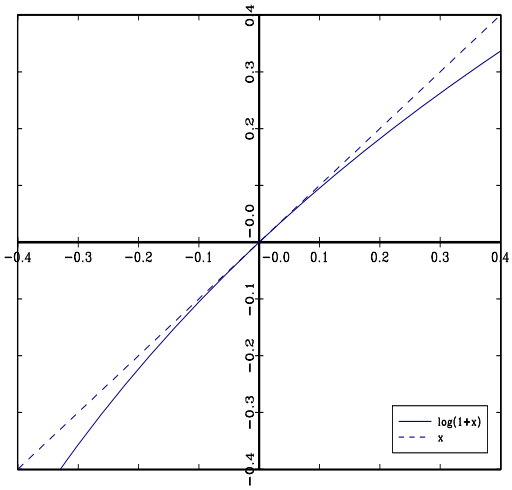


Figure 2.4: $\log(1+x)$

2.5 Conditional Expectation Function

An important determinant of wage levels is education. In many empirical studies economists measure educational attainment by the number of years of schooling, and we will write this variable as *education*[8].

The conditional mean of log wages given *sex*, *race*, and *education* is a single number for each category. For example

$$\mathbb{E}\left(\log(wage) \mid sex = man,\ race = white,\ education = 12\right) = 2.84$$

We display in Figure 2.5 the conditional means of $\log(wage)$ for white men and white women as a function of *education*. The plot is quite revealing. We see that the conditional mean is increasing in years of education, but at a different rate for schooling levels above and below nine years. Another striking feature of Figure 2.5 is that the gap between men and women is roughly constant for all education levels. As the variables are measured in logs this implies a constant average percentage gap between men and women regardless of educational attainment.

In many cases it is convenient to simplify the notation by writing variables using single characters, typically y, x and/or z. It is conventional in econometrics to denote the dependent variable (e.g. $\log(wage)$) by the letter y, a conditioning variable (such as *sex*) by the letter x, and multiple conditioning variables (such as *race*, *education* and *sex*) by the subscripted letters $x_1, x_2, ..., x_k$.

Conditional expectations can be written with the generic notation

$$\mathbb{E}\left(y \mid x_1, x_2, ..., x_k\right) = m(x_1, x_2, ..., x_k).$$

We call this the **conditional expectation function** (CEF). The CEF is a function of $(x_1, x_2, ..., x_k)$ as it varies with the variables. For example, the conditional expectation of $y = \log(wage)$ given $(x_1, x_2) = (sex,\ race)$ is given by the six entries of Table 2.1. The CEF is a function of $(sex,\ race)$ as it varies across the entries.

For greater compactness, we will typically write the conditioning variables as a vector in $\mathbb{R}^k$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}. \tag{2.5}$$

[8]Here, *education* is defined as years of schooling beyond kindergarten. A high school graduate has *education*=12, a college graduate has *education*=16, a Master's degree has *education*=18, and a professional degree (medical, law or PhD) has *education*=20.

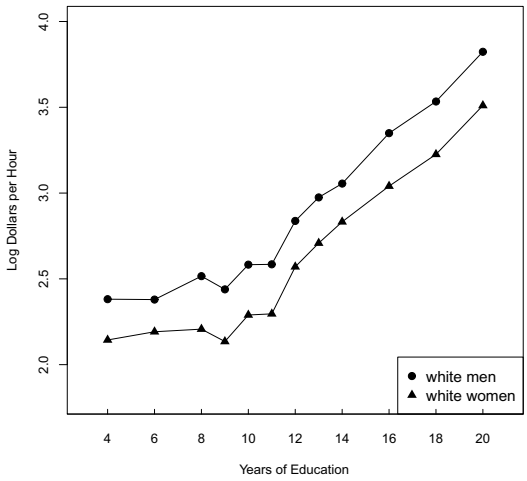


Figure 2.5: Mean Log Wage as a Function of Years of Education

Here we follow the convention of using lower case bold italics $\boldsymbol{x}$ to denote a vector. Given this notation, the CEF can be compactly written as

$$\mathbb{E}(y \mid \boldsymbol{x}) = m(\boldsymbol{x}).$$

The CEF $\mathbb{E}(y \mid \boldsymbol{x})$ is a random variable as it is a function of the random variable $\boldsymbol{x}$. It is also sometimes useful to view the CEF as a function of $\boldsymbol{x}$. In this case we can write $m(\boldsymbol{u}) = \mathbb{E}(y \mid \boldsymbol{x} = \boldsymbol{u})$, which is a function of the argument $\boldsymbol{u}$. The expression $\mathbb{E}(y \mid \boldsymbol{x} = \boldsymbol{u})$ is the conditional expectation of y, given that we know that the random variable $\boldsymbol{x}$ equals the specific value $\boldsymbol{u}$. However, sometimes in econometrics we take a notational shortcut and use $\mathbb{E}(y \mid \boldsymbol{x})$ to refer to this function. Hopefully, the use of $\mathbb{E}(y \mid \boldsymbol{x})$ should be apparent from the context.

2.6 Continuous Variables

In the previous sections, we implicitly assumed that the conditioning variables are discrete. However, many conditioning variables are continuous. In this section, we take up this case and assume that the variables $(y, \boldsymbol{x})$ are continuously distributed with a joint density function $f(y, \boldsymbol{x})$.

As an example, take $y = \log(wage)$ and $x = experience$, the number of years of potential labor market experience[9]. The contours of their joint density are plotted on the left side of Figure 2.6 for the population of white men with 12 years of education.

Given the joint density $f(y, \boldsymbol{x})$ the variable $\boldsymbol{x}$ has the marginal density

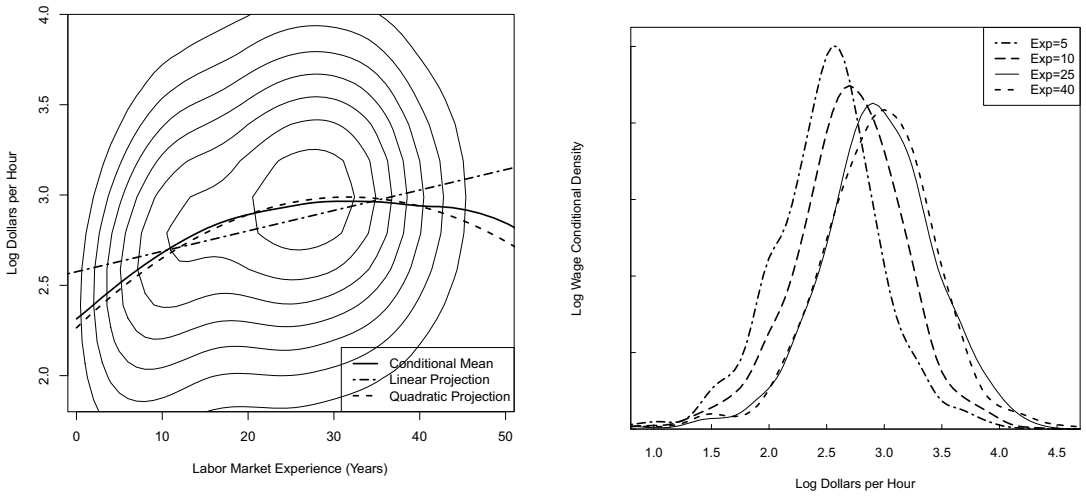
$$f_{\boldsymbol{x}}(\boldsymbol{x}) = \int_{\mathbb{R}} f(y, \boldsymbol{x}) dy.$$

For any $\boldsymbol{x}$ such that $f_{\boldsymbol{x}}(\boldsymbol{x}) > 0$ the conditional density of y given $\boldsymbol{x}$ is defined as

$$f_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}) = \frac{f(y, \boldsymbol{x})}{f_{\boldsymbol{x}}(\boldsymbol{x})}. \tag{2.6}$$

The conditional density is a (renormalized) slice of the joint density $f(y, \boldsymbol{x})$ holding $\boldsymbol{x}$ fixed. The slice is renormalized (divided by $f_{\boldsymbol{x}}(\boldsymbol{x})$ so that it integrates to one and is thus a density.) We can

[9] Here, *experience* is defined as potential labor market experience, equal to $age - education - 6$



(a) Joint density of log(wage) and experience and conditional mean

(b) Conditional density

Figure 2.6: White men with *education*=12

visualize this by slicing the joint density function at a specific value of $\boldsymbol{x}$ parallel with the y-axis. For example, take the density contours on the left side of Figure 2.6 and slice through the contour plot at a specific value of *experience*, and then renormalize the slice so that it is a proper density. This gives us the conditional density of log(*wage*) for white men with 12 years of education and this level of experience. We do this for four levels of *experience* (5, 10, 25, and 40 years), and plot these densities on the right side of Figure 2.6. We can see that the distribution of wages shifts to the right and becomes more diffuse as experience increases from 5 to 10 years, and from 10 to 25 years, but there is little change from 25 to 40 years experience.

The CEF of y given $\boldsymbol{x}$ is the mean of the conditional density (2.6)

$$m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x}) = \int_{\mathbb{R}} y f_{y|\boldsymbol{x}}(y \mid \boldsymbol{x})\, dy. \tag{2.7}$$

Intuitively, $m(\boldsymbol{x})$ is the mean of y for the idealized subpopulation where the conditioning variables are fixed at $\boldsymbol{x}$. This is idealized since $\boldsymbol{x}$ is continuously distributed so this subpopulation is infinitely small.

In Figure 2.6 the CEF of log(*wage*) given *experience* is plotted as the solid line. We can see that the CEF is a smooth but nonlinear function. The CEF is initially increasing in *experience*, flattens out around *experience* = 30, and then decreases for high levels of experience.

2.7 Law of Iterated Expectations

An extremely useful tool from probability theory is the **law of iterated expectations**. An important special case is the known as the Simple Law.

Theorem 2.7.1 ***Simple Law of Iterated Expectations***
If $\mathbb{E}|y| < \infty$ *then for any random vector* $\boldsymbol{x}$*,*

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x})) = \mathbb{E}(y)$$

The simple law states that the expectation of the conditional expectation is the unconditional expectation. In other words, the average of the conditional averages is the unconditional average. When $\boldsymbol{x}$ is discrete

$$\mathbb{E}\left(\mathbb{E}\left(y \mid \boldsymbol{x}\right)\right) = \sum_{j=1}^{\infty} \mathbb{E}\left(y \mid \boldsymbol{x}_j\right) \Pr\left(\boldsymbol{x} = \boldsymbol{x}_j\right)$$

and when $\boldsymbol{x}$ is continuous

$$\mathbb{E}\left(\mathbb{E}\left(y \mid \boldsymbol{x}\right)\right) = \int_{\mathbb{R}^k} \mathbb{E}\left(y \mid \boldsymbol{x}\right) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}.$$

Going back to our investigation of average log wages for men and women, the simple law states that

$$\begin{aligned}
&\mathbb{E}\left(\log(wage) \mid sex = man\right) \Pr\left(sex = man\right) \\
&+ \mathbb{E}\left(\log(wage) \mid sex = woman\right) \Pr\left(sex = woman\right) \\
&= \mathbb{E}\left(\log(wage)\right).
\end{aligned}$$

Or numerically,

$$3.05 \times 0.57 + 2.79 \times 0.43 = 2.92.$$

The general law of iterated expectations allows two sets of conditioning variables.

Theorem 2.7.2 ***Law of Iterated Expectations***
If $\mathbb{E}\left|y\right| < \infty$ *then for any random vectors* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*,*

$$\mathbb{E}\left(\mathbb{E}\left(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) \mid \boldsymbol{x}_1\right) = \mathbb{E}\left(y \mid \boldsymbol{x}_1\right)$$

Notice the way the law is applied. The inner expectation conditions on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, while the outer expectation conditions only on $\boldsymbol{x}_1$. The iterated expectation yields the simple answer $\mathbb{E}\left(y \mid \boldsymbol{x}_1\right)$, the expectation conditional on $\boldsymbol{x}_1$ alone. Sometimes we phrase this as: "The smaller information set wins."

As an example

$$\begin{aligned}
&\mathbb{E}\left(\log(wage) \mid sex = man,\ race = white\right) \Pr\left(race = white | sex = man\right) \\
&+ \mathbb{E}\left(\log(wage) \mid sex = man,\ race = black\right) \Pr\left(race = black | sex = man\right) \\
&+ \mathbb{E}\left(\log(wage) \mid sex = man,\ race = other\right) \Pr\left(race = other | sex = man\right) \\
&= \mathbb{E}\left(\log(wage) \mid sex = man\right)
\end{aligned}$$

or numerically

$$3.07 \times 0.84 + 2.86 \times 0.08 + 3.03 \times 0.08 = 3.05.$$

A property of conditional expectations is that when you condition on a random vector $\boldsymbol{x}$ you can effectively treat it as if it is constant. For example, $\mathbb{E}\left(\boldsymbol{x} \mid \boldsymbol{x}\right) = \boldsymbol{x}$ and $\mathbb{E}\left(g\left(\boldsymbol{x}\right) \mid \boldsymbol{x}\right) = g\left(\boldsymbol{x}\right)$ for any function $g(\cdot)$. The general property is known as the Conditioning Theorem.

Theorem 2.7.3 ***Conditioning Theorem***
If

$$\mathbb{E}\left|g\left(\boldsymbol{x}\right) y\right| < \infty \tag{2.8}$$

then

$$\mathbb{E}\left(g\left(\boldsymbol{x}\right) y \mid \boldsymbol{x}\right) = g\left(\boldsymbol{x}\right) \mathbb{E}\left(y \mid \boldsymbol{x}\right) \tag{2.9}$$

and

$$\mathbb{E}\left(g\left(\boldsymbol{x}\right) y\right) = \mathbb{E}\left(g\left(\boldsymbol{x}\right) \mathbb{E}\left(y \mid \boldsymbol{x}\right)\right). \tag{2.10}$$

The proofs of Theorems 2.7.1, 2.7.2 and 2.7.3 are given in Section 2.34.

2.8 CEF Error

The CEF error e is defined as the difference between y and the CEF evaluated at the random vector $\boldsymbol{x}$:

$$e = y - m(\boldsymbol{x}).$$

By construction, this yields the formula

$$y = m(\boldsymbol{x}) + e. \tag{2.11}$$

In (2.11) it is useful to understand that the error e is derived from the joint distribution of $(y, \boldsymbol{x})$, and so its properties are derived from this construction.

A key property of the CEF error is that it has a conditional mean of zero. To see this, by the linearity of expectations, the definition $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ and the Conditioning Theorem

$$\begin{aligned}
\mathbb{E}(e \mid \boldsymbol{x}) &= \mathbb{E}((y - m(\boldsymbol{x})) \mid \boldsymbol{x}) \\
&= \mathbb{E}(y \mid \boldsymbol{x}) - \mathbb{E}(m(\boldsymbol{x}) \mid \boldsymbol{x}) \\
&= m(\boldsymbol{x}) - m(\boldsymbol{x}) \\
&= 0.
\end{aligned}$$

This fact can be combined with the law of iterated expectations to show that the unconditional mean is also zero.

$$\mathbb{E}(e) = \mathbb{E}(\mathbb{E}(e \mid \boldsymbol{x})) = \mathbb{E}(0) = 0.$$

We state this and some other results formally.

Theorem 2.8.1 ***Properties of the CEF error***
If $\mathbb{E}|y| < \infty$ *then*

1. $\mathbb{E}(e \mid \boldsymbol{x}) = 0$.
2. $\mathbb{E}(e) = 0$.
3. *If* $\mathbb{E}|y|^r < \infty$ *for* $r \geq 1$ *then* $\mathbb{E}|e|^r < \infty$.
4. *For any function* $h(\boldsymbol{x})$ *such that* $\mathbb{E}|h(\boldsymbol{x})\, e| < \infty$ *then* $\mathbb{E}(h(\boldsymbol{x})\, e) = 0$.

The proof of the third result is deferred to Section 2.34.

The fourth result, whose proof is left to Exercise 2.3, implies that e is uncorrelated with any function of the regressors.

The equations

$$\begin{aligned}
y &= m(\boldsymbol{x}) + e \\
\mathbb{E}(e \mid \boldsymbol{x}) &= 0
\end{aligned}$$

together imply that $m(\boldsymbol{x})$ is the CEF of y given $\boldsymbol{x}$. It is important to understand that this is not a restriction. These equations hold true by definition.

The condition $\mathbb{E}(e \mid \boldsymbol{x}) = 0$ is implied by the definition of e as the difference between y and the CEF $m(\boldsymbol{x})$. The equation $\mathbb{E}(e \mid \boldsymbol{x}) = 0$ is sometimes called a conditional mean restriction, since the conditional mean of the error e is restricted to equal zero. The property is also sometimes called **mean independence**, for the conditional mean of e is 0 and thus independent of $\boldsymbol{x}$. However, it does not imply that the distribution of e is independent of $\boldsymbol{x}$. Sometimes the assumption "e is

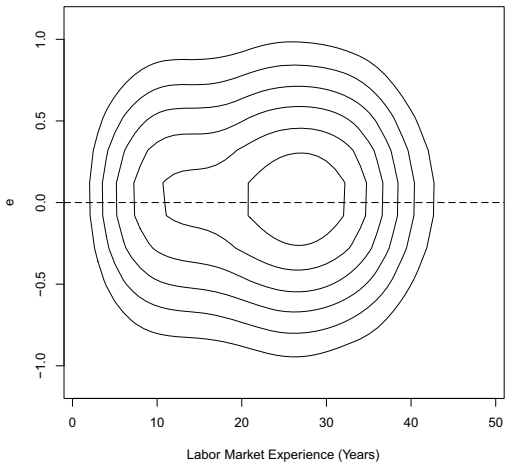


Figure 2.7: Joint density of CEF error e and *experience* for white men with *education*=12.

independent of $\boldsymbol{x}$" is added as a convenient simplification, but it is not generic feature of the conditional mean. Typically and generally, e and $\boldsymbol{x}$ are jointly dependent, even though the conditional mean of e is zero.

As an example, the contours of the joint density of e and *experience* are plotted in Figure 2.7 for the same population as Figure 2.6. The error e has a conditional mean of zero for all values of *experience*, but the shape of the conditional distribution varies with the level of *experience*.

As a simple example of a case where x and e are mean independent yet dependent, let $e = x\varepsilon$ where x and ε are independent $\mathrm{N}(0,1)$. Then conditional on x, the error e has the distribution $\mathrm{N}(0, x^2)$. Thus $\mathbb{E}(e \mid x) = 0$ and e is mean independent of x, yet e is not fully independent of x. Mean independence does not imply full independence.

2.9 Intercept-Only Model

A special case of the regression model is when there are no regressors $\boldsymbol{x}$. In this case $m(\boldsymbol{x}) = \mathbb{E}(y) = \mu$, the unconditional mean of y. We can still write an equation for y in the regression format:

$$\begin{aligned} y &= \mu + e \\ \mathbb{E}(e) &= 0 \end{aligned}$$

This is useful for it unifies the notation.

2.10 Regression Variance

An important measure of the dispersion about the CEF function is the unconditional variance of the CEF error e. We write this as

$$\sigma^2 = \operatorname{var}(e) = \mathbb{E}\left((e - \mathbb{E}e)^2\right) = \mathbb{E}\left(e^2\right).$$

Theorem 2.8.1.3 implies the following simple but useful result.

Theorem 2.10.1 *If* $\mathbb{E}y^2 < \infty$ *then* $\sigma^2 < \infty$.

We can call σ^2 the regression variance or the variance of the regression error. The magnitude of σ^2 measures the amount of variation in y which is not "explained" or accounted for in the conditional mean $\mathbb{E}(y \mid \boldsymbol{x})$.

The regression variance depends on the regressors $\boldsymbol{x}$. Consider two regressions

$$y = \mathbb{E}(y \mid \boldsymbol{x}_1) + e_1$$
$$y = \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) + e_2.$$

We write the two errors distinctly as e_1 and e_2 as they are different – changing the conditioning information changes the conditional mean and therefore the regression error as well.

In our discussion of iterated expectations, we have seen that by increasing the conditioning set, the conditional expectation reveals greater detail about the distribution of y. What is the implication for the regression error?

It turns out that there is a simple relationship. We can think of the conditional mean $\mathbb{E}(y \mid \boldsymbol{x})$ as the "explained portion" of y. The remainder $e = y - \mathbb{E}(y \mid \boldsymbol{x})$ is the "unexplained portion". The simple relationship we now derive shows that the variance of this unexplained portion decreases when we condition on more variables. This relationship is monotonic in the sense that increasing the amont of information always decreases the variance of the unexplained portion.

Theorem 2.10.2 *If* $\mathbb{E}y^2 < \infty$ *then*

$$\operatorname{var}(y) \geq \operatorname{var}(y - \mathbb{E}(y \mid \boldsymbol{x}_1)) \geq \operatorname{var}(y - \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)).$$

Theorem 2.10.2 says that the variance of the difference between y and its conditional mean (weakly) decreases whenever an additional variable is added to the conditioning information.

The proof of Theorem 2.10.2 is given in Section 2.34.

2.11 Best Predictor

Suppose that given a realized value of $\boldsymbol{x}$, we want to create a prediction or forecast of y. We can write any predictor as a function $g(\boldsymbol{x})$ of $\boldsymbol{x}$. The prediction error is the realized difference $y - g(\boldsymbol{x})$. A non-stochastic measure of the magnitude of the prediction error is the expectation of its square

$$\mathbb{E}(y - g(\boldsymbol{x}))^2. \tag{2.12}$$

We can define the best predictor as the function $g(\boldsymbol{x})$ which minimizes (2.12). What function is the best predictor? It turns out that the answer is the CEF $m(\boldsymbol{x})$. This holds regardless of the joint distribution of $(y, \boldsymbol{x})$.

To see this, note that the mean squared error of a predictor $g(\boldsymbol{x})$ is

$$\begin{aligned}
\mathbb{E}(y - g(\boldsymbol{x}))^2 &= \mathbb{E}(e + m(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&= \mathbb{E}e^2 + 2\mathbb{E}(e(m(\boldsymbol{x}) - g(\boldsymbol{x}))) + \mathbb{E}(m(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&= \mathbb{E}e^2 + \mathbb{E}(m(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&\geq \mathbb{E}e^2 \\
&= \mathbb{E}(y - m(\boldsymbol{x}))^2
\end{aligned}$$

where the first equality makes the substitution $y = m(\boldsymbol{x}) + e$ and the third equality uses Theorem 2.8.1.4. The right-hand-side after the third equality is minimized by setting $g(\boldsymbol{x}) = m(\boldsymbol{x})$, yielding

the inequality in the fourth line. The minimum is finite under the assumption $\mathbb{E}y^2 < \infty$ as shown by Theorem 2.10.1.

We state this formally in the following result.

Theorem 2.11.1 ***Conditional Mean as Best Predictor***
If $\mathbb{E}y^2 < \infty$, *then for any predictor* $g(\boldsymbol{x})$,

$$\mathbb{E}\left(y - g(\boldsymbol{x})\right)^2 \geq \mathbb{E}\left(y - m(\boldsymbol{x})\right)^2$$

where $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$.

It may be helpful to consider this result in the context of the intercept-only model

$$\begin{aligned} y &= \mu + e \\ \mathbb{E}(e) &= 0. \end{aligned}$$

Theorem 2.11.1 shows that the best predictor for y (in the class of constants) is the unconditional mean $\mu = \mathbb{E}(y)$, in the sense that the mean minimizes the mean squared prediction error.

2.12 Conditional Variance

While the conditional mean is a good measure of the location of a conditional distribution, it does not provide information about the spread of the distribution. A common measure of the dispersion is the **conditional variance**. We first give the general definition of the conditional variance of a random variable w.

Definition 2.12.1 *If* $\mathbb{E}w^2 < \infty$, *the* ***conditional variance*** *of* w *given* $\boldsymbol{x}$ *is*

$$\operatorname{var}(w \mid \boldsymbol{x}) = \mathbb{E}\left(\left(w - \mathbb{E}(w \mid \boldsymbol{x})\right)^2 \mid \boldsymbol{x}\right)$$

Notice that the conditional variance is the conditional second moment, centered around the conditional first moment. Given this definition, we define the conditional variance of the regression error.

Definition 2.12.2 *If* $\mathbb{E}e^2 < \infty$, *the* ***conditional variance*** *of the regression error* e *is*

$$\sigma^2(\boldsymbol{x}) = \operatorname{var}(e \mid \boldsymbol{x}) = \mathbb{E}\left(e^2 \mid \boldsymbol{x}\right).$$

Generally, $\sigma^2(\boldsymbol{x})$ is a non-trivial function of $\boldsymbol{x}$ and can take any form subject to the restriction that it is non-negative. One way to think about $\sigma^2(\boldsymbol{x})$ is that it is the conditional mean of e^2 given $\boldsymbol{x}$. Notice as well that $\sigma^2(\boldsymbol{x}) = \operatorname{var}(y \mid \boldsymbol{x})$ so it is equivalently the conditional variance of the dependent variable.

The variance is in a different unit of measurement than the original variable. To convert the variance back to the same unit of measure we define the **conditional standard deviation** as its square root $\sigma(\boldsymbol{x}) = \sqrt{\sigma^2(\boldsymbol{x})}$.

As an example of how the conditional variance depends on observables, compare the conditional log wage densities for men and women displayed in Figure 2.3. The difference between the densities is not purely a location shift, but is also a difference in spread. Specifically, we can see that the density for men's log wages is somewhat more spread out than that for women, while the density for women's wages is somewhat more peaked. Indeed, the conditional standard deviation for men's wages is 3.05 and that for women is 2.81. So while men have higher average wages, they are also somewhat more dispersed.

The unconditional error variance and the conditional variance are related by the law of iterated expectations

$$\sigma^2 = \mathbb{E}\left(e^2\right) = \mathbb{E}\left(\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right)\right) = \mathbb{E}\left(\sigma^2(\boldsymbol{x})\right).$$

That is, the unconditional error variance is the average conditional variance.

Given the conditional variance, we can define a rescaled error

$$\varepsilon = \frac{e}{\sigma(\boldsymbol{x})}. \tag{2.13}$$

We can calculate that since $\sigma(\boldsymbol{x})$ is a function of $\boldsymbol{x}$

$$\mathbb{E}\left(\varepsilon \mid \boldsymbol{x}\right) = \mathbb{E}\left(\frac{e}{\sigma(\boldsymbol{x})} \mid \boldsymbol{x}\right) = \frac{1}{\sigma(\boldsymbol{x})}\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$$

and

$$\operatorname{var}\left(\varepsilon \mid \boldsymbol{x}\right) = \mathbb{E}\left(\varepsilon^2 \mid \boldsymbol{x}\right) = \mathbb{E}\left(\frac{e^2}{\sigma^2(\boldsymbol{x})} \mid \boldsymbol{x}\right) = \frac{1}{\sigma^2(\boldsymbol{x})}\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \frac{\sigma^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} = 1.$$

Thus ε has a conditional mean of zero, and a conditional variance of 1.

Notice that (2.13) can be rewritten as

$$e = \sigma(\boldsymbol{x})\varepsilon.$$

and substituting this for e in the CEF equation (2.11), we find that

$$y = m(\boldsymbol{x}) + \sigma(\boldsymbol{x})\varepsilon. \tag{2.14}$$

This is an alternative (mean-variance) representation of the CEF equation.

Many econometric studies focus on the conditional mean $m(\boldsymbol{x})$ and either ignore the conditional variance $\sigma^2(\boldsymbol{x})$, treat it as a constant $\sigma^2(\boldsymbol{x}) = \sigma^2$, or treat it as a nuisance parameter (a parameter not of primary interest). This is appropriate when the primary variation in the conditional distribution is in the mean, but can be short-sighted in other cases. Dispersion is relevant to many economic topics, including income and wealth distribution, economic inequality, and price dispersion. Conditional dispersion (variance) can be a fruitful subject for investigation.

The perverse consequences of a narrow-minded focus on the mean has been parodied in a classic joke:

> An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, "On average I feel just fine."

Clearly, the economist in question ignored variance!

2.13 Homoskedasticity and Heteroskedasticity

An important special case obtains when the conditional variance $\sigma^2(\boldsymbol{x})$ is a constant and independent of $\boldsymbol{x}$. This is called **homoskedasticity**.

Definition 2.13.1 *The error is* ***homoskedastic*** *if* $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2$ *does not depend on* $\boldsymbol{x}$.

In the general case where $\sigma^2(\boldsymbol{x})$ depends on $\boldsymbol{x}$ we say that the error e is **heteroskedastic**.

Definition 2.13.2 *The error is* ***heteroskedastic*** *if* $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2(\boldsymbol{x})$ *depends on* $\boldsymbol{x}$.

It is helpful to understand that the concepts homoskedasticity and heteroskedasticity concern the conditional variance, not the unconditional variance. By definition, the unconditional variance σ^2 is a constant and independent of the regressors $\boldsymbol{x}$. So when we talk about the variance as a function of the regressors, we are talking about the conditional variance $\sigma^2(\boldsymbol{x})$.

Some older or introductory textbooks describe heteroskedasticity as the case where "the variance of e varies across observations". This is a poor and confusing definition. It is more constructive to understand that heteroskedasticity means that the conditional variance $\sigma^2(\boldsymbol{x})$ depends on observables.

Older textbooks also tend to describe homoskedasticity as a component of a correct regression specification, and describe heteroskedasticity as an exception or deviance. This description has influenced many generations of economists, but it is unfortunately backwards. The correct view is that heteroskedasticity is generic and "standard", while homoskedasticity is unusual and exceptional. The default in empirical work should be to assume that the errors are heteroskedastic, not the converse.

In apparent contradiction to the above statement, we will still frequently impose the homoskedasticity assumption when making theoretical investigations into the properties of estimation and inference methods. The reason is that in many cases homoskedasticity greatly simplifies the theoretical calculations, and it is therefore quite advantageous for teaching and learning. It should always be remembered, however, that homoskedasticity is never imposed because it is believed to be a correct feature of an empirical model, but rather because of its simplicity.

2.14 Regression Derivative

One way to interpret the CEF $m(\boldsymbol{x}) = \mathbb{E}\left(y \mid \boldsymbol{x}\right)$ is in terms of how marginal changes in the regressors $\boldsymbol{x}$ imply changes in the conditional mean of the response variable y. It is typical to consider marginal changes in a single regressor, say x_1, holding the remainder fixed. When a regressor x_1 is continuously distributed, we define the marginal effect of a change in x_1, holding the variables $x_2, ..., x_k$ fixed, as the partial derivative of the CEF

$$\frac{\partial}{\partial x_1} m(x_1, ..., x_k).$$

When x_1 is discrete we define the marginal effect as a discrete difference. For example, if x_1 is binary, then the marginal effect of x_1 on the CEF is

$$m(1, x_2, ..., x_k) - m(0, x_2, ..., x_k).$$

We can unify the continuous and discrete cases with the notation

$$\nabla_1 m(\boldsymbol{x}) = \begin{cases} \dfrac{\partial}{\partial x_1} m(x_1, ..., x_k), & \text{if } x_1 \text{ is continuous} \\ \\ m(1, x_2, ..., x_k) - m(0, x_2, ..., x_k), & \text{if } x_1 \text{ is binary.} \end{cases}$$

Collecting the k effects into one $k \times 1$ vector, we define the **regression derivative** with respect to $\boldsymbol{x}$:

$$\boldsymbol{\nabla} m(\boldsymbol{x}) = \begin{bmatrix} \nabla_1 m(\boldsymbol{x}) \\ \nabla_2 m(\boldsymbol{x}) \\ \vdots \\ \nabla_k m(\boldsymbol{x}) \end{bmatrix}$$

When all elements of $\boldsymbol{x}$ are continuous, then we have the simplification $\boldsymbol{\nabla} m(\boldsymbol{x}) = \dfrac{\partial}{\partial \boldsymbol{x}} m(\boldsymbol{x})$, the vector of partial derivatives.

There are two important points to remember concerning our definition of the regression derivative.

First, the effect of each variable is calculated holding the other variables constant. This is the **ceteris paribus** concept commonly used in economics. But in the case of a regression derivative, the conditional mean does not literally hold *all else* constant. It only holds constant the variables included in the conditional mean. This means that the regression derivative depends on which regressors are included. For example, in a regression of wages on education, experience, race and sex, the regression derivative with respect to education shows the marginal effect of education on mean wages, holding constant experience, race and sex. But it does not hold constant an individual's unobservable characteristics (such as ability), nor variables not included in the regression (such as the quality of education).

Second, the regression derivative is the change in the conditional expectation of y, not the change in the actual value of y for an individual. It is tempting to think of the regression derivative as the change in the actual value of y, but this is not a correct interpretation. The regression derivative $\boldsymbol{\nabla} m(\boldsymbol{x})$ is the change in the actual value of y only if the error e is unaffected by the change in the regressor $\boldsymbol{x}$. We return to a discussion of causal effects in Section 2.30.

2.15 Linear CEF

An important special case is when the CEF $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is linear in $\boldsymbol{x}$. In this case we can write the mean equation as

$$m(\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \beta_{k+1}.$$

Notationally it is convenient to write this as a simple function of the vector $\boldsymbol{x}$. An easy way to do so is to augment the regressor vector $\boldsymbol{x}$ by listing the number "1" as an element. We call this the "constant" and the corresponding coefficient is called the "intercept". Equivalently, specify that the final element[10] of the vector $\boldsymbol{x}$ is $x_k = 1$. Thus (2.5) has been redefined as the $k \times 1$ vector

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \\ 1 \end{pmatrix}. \tag{2.15}$$

[10] The order doesn't matter. It could be any element.

With this redefinition, the CEF is

$$\begin{aligned} m(\boldsymbol{x}) &= x_1\beta_1 + x_2\beta_2 + \cdots + \beta_k \\ &= \boldsymbol{x}'\boldsymbol{\beta} \end{aligned} \tag{2.16}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{2.17}$$

is a $k \times 1$ coefficient vector. This is the **linear CEF model**. It is also often called the **linear regression model**, or the regression of y on $\boldsymbol{x}$.

In the linear CEF model, the regression derivative is simply the coefficient vector. That is

$$\boldsymbol{\nabla} m(\boldsymbol{x}) = \boldsymbol{\beta}.$$

This is one of the appealing features of the linear CEF model. The coefficients have simple and natural interpretations as the marginal effects of changing one variable, holding the others constant.

Linear CEF Model

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}\right) &= 0 \end{aligned}$$

If in addition the error is homoskedastic, we call this the homoskedastic linear CEF model.

Homoskedastic Linear CEF Model

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}\right) &= 0 \\ \mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) &= \sigma^2 \end{aligned}$$

2.16 Linear CEF with Nonlinear Effects

The linear CEF model of the previous section is less restrictive than it might appear, as we can include as regressors nonlinear transformations of the original variables. In this sense, the linear CEF framework is flexible and can capture many nonlinear effects.

For example, suppose we have two scalar variables x_1 and x_2. The CEF could take the quadratic form

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6. \tag{2.18}$$

This equation is quadratic in the regressors (x_1, x_2) yet linear in the coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_6)'$. We will descriptively call (2.18) a **quadratic CEF**, and yet (2.18) is also a **linear CEF** in the sense of being linear in the coefficients. The key is to understand that (2.18) is quadratic in the variables (x_1, x_2) yet linear in the coefficients $\boldsymbol{\beta}$.

To simplify the expression, we define the transformations $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1x_2$, and $x_6 = 1$, and redefine the regressor vector as $\boldsymbol{x} = (x_1, ..., x_6)'$. With this redefinition,

$$m(x_1, x_2) = \boldsymbol{x}'\boldsymbol{\beta}$$

which is linear in $\boldsymbol{\beta}$. For most econometric purposes (estimation and inference on $\boldsymbol{\beta}$) the linearity in $\boldsymbol{\beta}$ is all that is important.

An exception is in the analysis of regression derivatives. In nonlinear equations such as (2.18), the regression derivative should be defined with respect to the original variables, not with respect to the transformed variables. Thus

$$\frac{\partial}{\partial x_1} m(x_1, x_2) = \beta_1 + 2x_1\beta_3 + x_2\beta_5$$
$$\frac{\partial}{\partial x_2} m(x_1, x_2) = \beta_2 + 2x_2\beta_4 + x_1\beta_5$$

We see that in the model (2.18), the regression derivatives are not a simple coefficient, but are functions of several coefficients plus the levels of (x_1,x_2). Consequently it is difficult to interpret the coefficients individually. It is more useful to interpret them as a group.

We typically call β_5 the **interaction effect**. Notice that it appears in both regression derivative equations, and has a symmetric interpretation in each. If $\beta_5 > 0$ then the regression derivative with respect to x_1 is increasing in the level of x_2 (and the regression derivative with respect to x_2 is increasing in the level of x_1), while if $\beta_5 < 0$ the reverse is true. It is worth noting that this symmetry is an artificial implication of the quadratic equation (2.18), and is not a general feature of nonlinear conditional means $m(x_1, x_2)$.

2.17 Linear CEF with Dummy Variables

When all regressors take a finite set of values, it turns out the CEF can be written as a linear function of regressors.

This simplest example is a **binary** variable, which takes only two distinct values. For example, the variable *sex* typically takes only the values *man* and *woman*. Binary variables are extremely common in econometric applications, and are alternatively called **dummy variables** or **indicator variables**.

Consider the simple case of a single binary regressor. In this case, the conditional mean can only take two distinct values. For example,

$$\mathbb{E}(y \mid sex) = \begin{cases} \mu_0 & \text{if} \quad sex\text{=}man \\ \\ \mu_1 & \text{if} \quad sex\text{=}woman \end{cases}$$

To facilitate a mathematical treatment, we typically record dummy variables with the values $\{0, 1\}$. For example

$$x_1 = \begin{cases} 0 & \text{if} \quad sex\text{=}man \\ 1 & \text{if} \quad sex\text{=}woman \end{cases} \tag{2.19}$$

Given this notation we can write the conditional mean as a linear function of the dummy variable x_1, that is

$$\mathbb{E}(y \mid x_1) = \beta_1 x_1 + \beta_2$$

where $\beta_1 = \mu_1 - \mu_0$ and $\beta_2 = \mu_0$. In this simple regression equation the intercept β_2 is equal to the conditional mean of y for the $x_1 = 0$ subpopulation (men) and the slope β_1 is equal to the difference in the conditional means between the two subpopulations.

Equivalently, we could have defined x_1 as

$$x_1 = \begin{cases} 1 & \text{if} \quad \textit{sex=man} \\ 0 & \text{if} \quad \textit{sex=woman} \end{cases} \tag{2.20}$$

In this case, the regression intercept is the mean for women (rather than for men) and the regression slope has switched signs. The two regressions are equivalent but the interpretation of the coefficients has changed. Therefore it is always important to understand the precise definitions of the variables, and illuminating labels are helpful. For example, labelling x_1 as "sex" does not help distinguish between definitions (2.19) and (2.20). Instead, it is better to label x_1 as "women" or "female" if definition (2.19) is used, or as "men" or "male" if (2.20) is used.

Now suppose we have two dummy variables x_1 and x_2. For example, $x_2 = 1$ if the person is married, else $x_2 = 0$. The conditional mean given x_1 and x_2 takes at most four possible values:

$$\mathbb{E}(y \mid x_1, x_2) = \begin{cases} \mu_{00} & \text{if} \quad x_1 = 0 \text{ and } x_2 = 0 & \text{(unmarried men)} \\ \mu_{01} & \text{if} \quad x_1 = 0 \text{ and } x_2 = 1 & \text{(married men)} \\ \mu_{10} & \text{if} \quad x_1 = 1 \text{ and } x_2 = 0 & \text{(unmarried women)} \\ \mu_{11} & \text{if} \quad x_1 = 1 \text{ and } x_2 = 1 & \text{(married women)} \end{cases}$$

In this case we can write the conditional mean as a linear function of x_1, x_2 and their product x_1x_2 :

$$\mathbb{E}(y \mid x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4$$

where $\beta_1 = \mu_{10} - \mu_{00}$, $\beta_2 = \mu_{01} - \mu_{00}$, $\beta_3 = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$, and $\beta_4 = \mu_{00}$.

We can view the coefficient β_1 as the effect of sex on expected log wages for unmarried wages earners, the coefficient β_2 as the effect of marriage on expected log wages for men wage earners, and the coefficient β_3 as the difference between the effects of marriage on expected log wages among women and among men. Alternatively, it can also be interpreted as the difference between the effects of sex on expected log wages among married and non-married wage earners. Both interpretations are equally valid. We often describe β_3 as measuring the **interaction** between the two dummy variables, or the **interaction effect**, and describe $\beta_3 = 0$ as the case when the interaction effect is zero.

In this setting we can see that the CEF is linear in the three variables (x_1, x_2, x_1x_2). Thus to put the model in the framework of Section 2.15, we would define the regressor $x_3 = x_1x_2$ and the regressor vector as

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}.$$

So even though we started with only 2 dummy variables, the number of regressors (including the intercept) is 4.

If there are 3 dummy variables x_1, x_2, x_3, then $\mathbb{E}(y \mid x_1, x_2, x_3)$ takes at most $2^3 = 8$ distinct values and can be written as the linear function

$$\mathbb{E}(y \mid x_1, x_2, x_3) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \beta_8$$

which has eight regressors including the intercept.

In general, if there are p dummy variables $x_1, ..., x_p$ then the CEF $\mathbb{E}(y \mid x_1, x_2, ..., x_p)$ takes at most 2^p distinct values, and can be written as a linear function of the 2^p regressors including $x_1, x_2, ..., x_p$ and all cross-products. This might be excessive in practice if p is modestly large. In the next section we will discuss projection approximations which yield more parsimonious parameterizations.

We started this section by saying that the conditional mean is linear whenever all regressors take only a finite number of possible values. How can we see this? Take a **categorical** variable,

such as *race*. For example, we earlier divided race into three categories. We can record categorical variables using numbers to indicate each category, for example

$$x_3 = \begin{cases} 1 & \text{if} \quad \textit{white} \\ 2 & \text{if} \quad \textit{black} \\ 3 & \text{if} \quad \textit{other} \end{cases}$$

When doing so, the values of x_3 have no meaning in terms of magnitude, they simply indicate the relevant category.

When the regressor is categorical the conditional mean of y given x_3 takes a distinct value for each possibility:

$$\mathbb{E}(y \mid x_3) = \begin{cases} \mu_1 & \text{if} \quad x_3 = 1 \\ \mu_2 & \text{if} \quad x_3 = 2 \\ \mu_3 & \text{if} \quad x_3 = 3 \end{cases}$$

This is not a linear function of x_3 itself, but it can be made a linear function by constructing dummy variables for two of the three categories. For example

$$x_4 = \begin{cases} 1 & \text{if} \quad \textit{black} \\ 0 & \text{if} \quad \textit{not black} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if} \quad \textit{other} \\ 0 & \text{if} \quad \textit{not other} \end{cases}$$

In this case, the categorical variable x_3 is equivalent to the pair of dummy variables (x_4, x_5). The explicit relationship is

$$x_3 = \begin{cases} 1 & \text{if} \quad x_4 = 0 \text{ and } x_5 = 0 \\ 2 & \text{if} \quad x_4 = 1 \text{ and } x_5 = 0 \\ 3 & \text{if} \quad x_4 = 0 \text{ and } x_5 = 1 \end{cases}$$

Given these transformations, we can write the conditional mean of y as a linear function of x_4 and x_5

$$\mathbb{E}(y \mid x_3) = \mathbb{E}(y \mid x_4, x_5) = \beta_1 x_4 + \beta_2 x_5 + \beta_3$$

We can write the CEF as either $\mathbb{E}(y \mid x_3)$ or $\mathbb{E}(y \mid x_4, x_5)$ (they are equivalent), but it is only linear as a function of x_4 and x_5.

This setting is similar to the case of two dummy variables, with the difference that we have not included the interaction term $x_4 x_5$. This is because the event $\{x_4 = 1 \text{ and } x_5 = 1\}$ is empty by construction, so $x_4 x_5 = 0$ by definition.

2.18 Best Linear Predictor

While the conditional mean $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is the best predictor of y among all functions of $\boldsymbol{x}$, its functional form is typically unknown. In particular, the linear CEF model is empirically unlikely to be accurate unless $\boldsymbol{x}$ is discrete and low-dimensional so all interactions are included. Consequently in most cases it is more realistic to view the linear specification (2.16) as an approximation. In this section we derive a specific approximation with a simple interpretation.

Theorem 2.11.1 showed that the conditional mean $m(\boldsymbol{x})$ is the best predictor in the sense that it has the lowest mean squared error among all predictors. By extension, we can define an approximation to the CEF by the linear function with the lowest mean squared error among all linear predictors.

For this derivation we require the following regularity condition.

Assumption 2.18.1

1. $\mathbb{E}y^2 < \infty$.
2. $\mathbb{E}\left\|\boldsymbol{x}\right\|^2 < \infty$.
3. $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ *is positive definite.*

In Assumption 2.18.1.2 we use the notation $\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2}$ to denote the Euclidean length of the vector $\boldsymbol{x}$.

The first two parts of Assumption 2.18.1 imply that the variables y and $\boldsymbol{x}$ have finite means, variances, and covariances. The third part of the assumption is more technical, and its role will become apparent shortly. It is equivalent to imposing that the columns of the matrix $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ are linearly independent, or equivalently that the matrix is invertible.

A linear predictor for y is a function of the form $\boldsymbol{x}'\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. The mean squared prediction error is

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2.$$

The **best linear predictor** of y given $\boldsymbol{x}$, written $\mathcal{P}(y \mid \boldsymbol{x})$, is found by selecting the vector $\boldsymbol{\beta}$ to minimize $S(\boldsymbol{\beta})$.

Definition 2.18.1 *The* ***Best Linear Predictor*** *of y given $\boldsymbol{x}$ is*

$$\mathcal{P}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ minimizes the mean squared prediction error

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2.$$

The minimizer

$$\boldsymbol{\beta} = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^k} S(\boldsymbol{b}) \tag{2.21}$$

is called the ***Linear Projection Coefficient****.*

We now calculate an explicit expression for its value. The mean squared prediction error can be written out as a quadratic function of $\boldsymbol{\beta}$:

$$S(\boldsymbol{\beta}) = \mathbb{E}y^2 - 2\boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}y\right) + \boldsymbol{\beta}'\mathbb{E}\big(\boldsymbol{x}\boldsymbol{x}'\big)\boldsymbol{\beta}.$$

The quadratic structure of $S(\boldsymbol{\beta})$ means that we can solve explicitly for the minimizer. The first-order condition for minimization (from Appendix A.10) is

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2\mathbb{E}\left(\boldsymbol{x}y\right) + 2\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}. \tag{2.22}$$

Rewriting (2.22) as

$$2\mathbb{E}\left(\boldsymbol{x}y\right) = 2\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}$$

and dividing by 2, this equation takes the form

$$\boldsymbol{Q}_{\boldsymbol{xy}} = \boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{\beta} \tag{2.23}$$

where $\boldsymbol{Q}_{\boldsymbol{xy}} = \mathbb{E}(\boldsymbol{x}y)$ is $k \times 1$ and $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ is $k \times k$. The solution is found by inverting the matrix $\boldsymbol{Q}_{\boldsymbol{xx}}$, and is written

$$\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}}$$

or

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y). \tag{2.24}$$

It is worth taking the time to understand the notation involved in the expression (2.24). $\boldsymbol{Q}_{\boldsymbol{xx}}$ is a $k \times k$ matrix and $\boldsymbol{Q}_{\boldsymbol{xy}}$ is a $k \times 1$ column vector. Therefore, alternative expressions such as $\frac{\mathbb{E}(\boldsymbol{x}y)}{\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')}$ or $\mathbb{E}(\boldsymbol{x}y)\left(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')\right)^{-1}$ are incoherent and incorrect. We also can now see the role of Assumption 2.18.1.3. It is equivalent to assuming that $\boldsymbol{Q}_{\boldsymbol{xx}}$ has an inverse $\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ which is necessary for the normal equations (2.23) to have a solution or equivalently for (2.24) to be uniquely defined. In the absence of Assumption 2.18.1.3 there could be multiple solutions to the equation (2.23).

We now have an explicit expression for the best linear predictor:

$$\mathcal{P}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y).$$

This expression is also referred to as the **linear projection** of y on $\boldsymbol{x}$.

The **projection error** is

$$e = y - \boldsymbol{x}'\boldsymbol{\beta}. \tag{2.25}$$

This equals the error from the regression equation when (and only when) the conditional mean is linear in $\boldsymbol{x}$, otherwise they are distinct.

Rewriting, we obtain a decomposition of y into linear predictor and error

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e. \tag{2.26}$$

In general we call equation (2.26) or $\boldsymbol{x}'\boldsymbol{\beta}$ the best linear predictor of y given $\boldsymbol{x}$**,** or the linear projection of y on $\boldsymbol{x}$. Equation (2.26) is also often called the **regression** of y on $\boldsymbol{x}$ but this can sometimes be confusing as economists use the term *regression* in many contexts. (Recall that we said in Section 2.15 that the linear CEF model is also called the linear regression model.)

An important property of the projection error e is

$$\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}. \tag{2.27}$$

To see this, using the definitions (2.25) and (2.24) and the matrix properties $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ and $\boldsymbol{I}\boldsymbol{a} = \boldsymbol{a}$,

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}e) &= \mathbb{E}\left(\boldsymbol{x}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\right) \\ &= \mathbb{E}(\boldsymbol{x}y) - \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y) \\ &= \mathbf{0}\end{aligned} \tag{2.28}$$

as claimed.

Equation (2.27) is a set of k equations, one for each regressor. In other words, (2.27) is equivalent to

$$\mathbb{E}(x_j e) = 0 \tag{2.29}$$

for $j = 1, ..., k$. As in (2.15), the regressor vector $\boldsymbol{x}$ typically contains a constant, e.g. $x_k = 1$. In this case (2.29) for $j = k$ is the same as

$$\mathbb{E}(e) = 0. \tag{2.30}$$

Thus the projection error has a mean of zero when the regressor vector contains a constant. (When $\boldsymbol{x}$ does not have a constant, (2.30) is not guaranteed. As it is desirable for e to have a zero mean, this is a good reason to always include a constant in any regression model.)

It is also useful to observe that since $\operatorname{cov}(x_j, e) = \mathbb{E}(x_j e) - \mathbb{E}(x_j)\mathbb{E}(e)$, then (2.29)-(2.30) together imply that the variables x_j and e are uncorrelated.

This completes the derivation of the model. We summarize some of the most important properties.

Theorem 2.18.1 ***Properties of Linear Projection Model***
Under Assumption 2.18.1,

1. *The moments $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ and $\mathbb{E}(\boldsymbol{x}y)$ exist with finite elements.*
2. *The Linear Projection Coefficient (2.21) exists, is unique, and equals*
$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y).$$
3. *The best linear predictor of y given $\boldsymbol{x}$ is*
$$\mathcal{P}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y).$$
4. *The projection error $e = y - \boldsymbol{x}'\boldsymbol{\beta}$ exists and satisfies*
$$\mathbb{E}\left(e^2\right) < \infty$$
and
$$\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}.$$
5. *If $\boldsymbol{x}$ contains an constant, then*
$$\mathbb{E}(e) = 0.$$
6. *If $\mathbb{E}|y|^r < \infty$ and $\mathbb{E}\|\boldsymbol{x}\|^r < \infty$ for $r \geq 2$ then $\mathbb{E}|e|^r < \infty$.*

A complete proof of Theorem 2.18.1 is given in Section 2.34.

It is useful to reflect on the generality of Theorem 2.18.1. The only restriction is Assumption 2.18.1. Thus for any random variables $(y, \boldsymbol{x})$ with finite variances we can define a linear equation (2.26) with the properties listed in Theorem 2.18.1. Stronger assumptions (such as the linear CEF model) are not necessary. In this sense the linear model (2.26) exists quite generally. However, it is important not to misinterpret the generality of this statement. The linear equation (2.26) is defined as the best linear predictor. It is not necessarily a conditional mean, nor a parameter of a structural or causal economic model.

Linear Projection Model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e.$$
$$\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$$
$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y)$$

We illustrate projection using three log wage equations introduced in earlier sections.

For our first example, we consider a model with the two dummy variables for sex and race similar to Table 2.1. As we learned in Section 2.17, the entries in this table can be equivalently expressed by a linear CEF. For simplicity, let's consider the CEF of $\log(wage)$ as a function of *Black* and *Female*.

$$\mathbb{E}(\log(wage) \mid Black, Female) = -0.20Black - 0.24Female + 0.10Black \times Female + 3.06. \quad (2.31)$$

This is a CEF as the variables are binary and all interactions are included.

Now consider a simpler model omitting the interaction effect. This is the linear projection on the variables *Black* and *Female*

$$\mathcal{P}(\log(wage) \mid Black, Female) = -0.15Black - 0.23Female + 3.06. \quad (2.32)$$

What is the difference? The full CEF (2.31) shows that the race gap is differentiated by sex: it is 20% for black men (relative to non-black men) and 10% for black women (relative to non-black women). The projection model (2.32) simplifies this analysis, calculating an average 15% wage gap for blacks, ignoring the role of sex. Notice that this is despite the fact that the sex variable is included in (2.32).

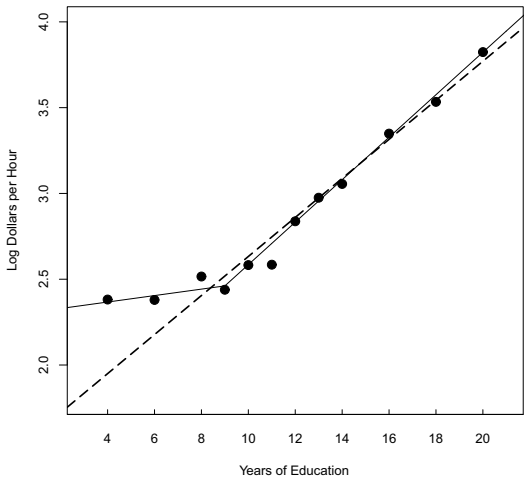


Figure 2.8: Projections of $\log(wage)$ onto Education

For our second example we consider the CEF of log wages as a function of years of education for white men which was illustrated in Figure 2.5 and is repeated in Figure 2.8. Superimposed on the figure are two projections. The first (given by the dashed line) is the linear projection of log wages on years of education

$$\mathcal{P}(\log(wage) \mid Education) = 0.11Education + 1.5$$

This simple equation indicates an average 11% increase in wages for every year of education. An inspection of the Figure shows that this approximation works well for *education*$\geq$ 9, but underpredicts for individuals with lower levels of education. To correct this imbalance we use a linear spline equation which allows different rates of return above and below 9 years of education:

$$\begin{aligned}&\mathcal{P}\left(\log(wage) \mid Education, (Education - 9) \times 1\left(Education > 9\right)\right)\\ &= 0.02Education + 0.10 \times (Education - 9) \times 1\left(Education > 9\right) + 2.3\end{aligned}$$

This equation is displayed in Figure 2.8 using the solid line, and appears to fit much better. It indicates a 2% increase in mean wages for every year of education below 9, and a 12% increase in

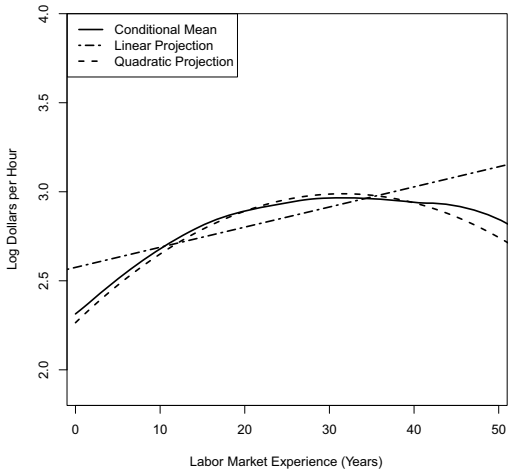


Figure 2.9: Linear and Quadratic Projections of log(*wage*) onto Experience

mean wages for every year of education above 9. It is still an approximation to the conditional mean but it appears to be fairly reasonable.

For our third example we take the CEF of log wages as a function of years of experience for white men with 12 years of education, which was illustrated in Figure 2.6 and is repeated as the solid line in Figure 2.9. Superimposed on the figure are two projections. The first (given by the dot-dashed line) is the linear projection on experience

$$\mathcal{P}(\log(wage) \mid Experience) = 0.011 Experience + 2.5$$

and the second (given by the dashed line) is the linear projection on experience and its square

$$\mathcal{P}(\log(wage) \mid Experience) = 0.046 Experience - 0.0007 Experience^2 + 2.3.$$

It is fairly clear from an examination of Figure 2.9 that the first linear projection is a poor approximation. It over-predicts wages for young and old workers, and under-predicts for the rest. Most importantly, it misses the strong downturn in expected wages for older wage-earners. The second projection fits much better. We can call this equation a **quadratic projection** since the function is quadratic in *experience*.

Invertibility and Identification

The linear projection coefficient $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ is invertible. The matrix $\boldsymbol{Q}_{\boldsymbol{xx}}$ is sometimes called the **design matrix**, as in experimental settings the researcher is able to control $\boldsymbol{Q}_{\boldsymbol{xx}}$ by manipulating the distribution of the regressors $\boldsymbol{x}$.

Observe that for any non-zero $\boldsymbol{\alpha} \in \mathbb{R}^k$,

$$\boldsymbol{\alpha}'\boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{\alpha} = \mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\alpha}\right) = \mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\right)^2 \geq 0$$

so $\boldsymbol{Q}_{\boldsymbol{xx}}$ by construction is positive semi-definite. The assumption that it is positive definite means that this is a strict inequality, $\mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\right)^2 > 0$. Equivalently, there cannot exist a non-zero vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}'\boldsymbol{x} = 0$ identically. This occurs when redundant variables are included in $\boldsymbol{x}$. Positive semi-definite matrices are invertible if and only if they are positive definite. When $\boldsymbol{Q}_{\boldsymbol{xx}}$ is invertible then $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ exists and is uniquely defined. In other words, in order for $\boldsymbol{\beta}$ to be uniquely defined, we must exclude the degenerate situation of redundant varibles.

Theorem 2.18.1 shows that the linear projection coefficient $\boldsymbol{\beta}$ is **identified** (uniquely determined) under Assumption 2.18.1. The key is invertibility of $\boldsymbol{Q}_{\boldsymbol{xx}}$. Otherwise, there is no unique solution to the equation

$$\boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{xy}}. \tag{2.33}$$

When $\boldsymbol{Q}_{\boldsymbol{xx}}$ is not invertible there are multiple solutions to (2.33), all of which yield an equivalent best linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$. In this case the coefficient $\boldsymbol{\beta}$ is **not identified** as it does not have a unique value. Even so, the best linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$ still identified. One solution is to set

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-}\mathbb{E}\left(\boldsymbol{x}y\right)$$

where $\boldsymbol{A}^{-}$ denotes the generalized inverse of $\boldsymbol{A}$ (see Appendix A.5).

2.19 Linear Predictor Error Variance

As in the CEF model, we define the error variance as

$$\sigma^2 = \mathbb{E}\left(e^2\right).$$

Setting $Q_{yy} = \mathbb{E}\left(y^2\right)$ and $\boldsymbol{Q}_{\boldsymbol{yx}} = \mathbb{E}\left(y\boldsymbol{x}'\right)$ we can write σ^2 as

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 \\
&= \mathbb{E}y^2 - 2\mathbb{E}\left(y\boldsymbol{x}'\right)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta} \\
&= Q_{yy} - 2\boldsymbol{Q}_{\boldsymbol{yx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}} + \boldsymbol{Q}_{\boldsymbol{yx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}} \\
&= Q_{yy} - \boldsymbol{Q}_{\boldsymbol{yx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}} \\
&\overset{def}{=} Q_{yy\cdot\boldsymbol{x}}.
\end{aligned} \tag{2.34}$$

One useful feature of this formula is that it shows that $Q_{yy\cdot\boldsymbol{x}} = Q_{yy} - \boldsymbol{Q}_{\boldsymbol{yx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}}$ equals the variance of the error from the linear projection of y on $\boldsymbol{x}$.

2.20 Regression Coefficients

Sometimes it is useful to separate the constant from the other regressors, and write the linear projection equation in the format

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e \tag{2.35}$$

where α is the intercept and $\boldsymbol{x}$ does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}y = \mathbb{E}\boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}\alpha + \mathbb{E}e$$

or

$$\mu_y = \mu_x'\boldsymbol{\beta} + \alpha$$

where $\mu_y = \mathbb{E}y$ and $\mu_x = \mathbb{E}\boldsymbol{x}$, since $\mathbb{E}(e) = 0$ from (2.30). (While $\boldsymbol{x}$ does not contain a constant, the equation does so (2.30) still applies.) Rearranging, we find

$$\alpha = \mu_y - \mu_x'\boldsymbol{\beta}.$$

Subtracting this equation from (2.35) we find

$$y - \mu_y = (\boldsymbol{x} - \mu_x)'\boldsymbol{\beta} + e, \tag{2.36}$$

a linear equation between the centered variables $y - \mu_y$ and $\boldsymbol{x} - \mu_x$. (They are centered at their means, so are mean-zero random variables.) Because $\boldsymbol{x} - \mu_x$ is uncorrelated with e, (2.36) is also a linear projection, thus by the formula for the linear projection model,

$$\begin{aligned}\boldsymbol{\beta} &= \left(\mathbb{E}\left((\boldsymbol{x} - \mu_x)(\boldsymbol{x} - \mu_x)'\right)\right)^{-1}\mathbb{E}\left((\boldsymbol{x} - \mu_x)(y - \mu_y)\right) \\ &= \operatorname{var}(\boldsymbol{x})^{-1}\operatorname{cov}(\boldsymbol{x}, y)\end{aligned}$$

a function only of the covariances[11] of $\boldsymbol{x}$ and y.

Theorem 2.20.1 *In the linear projection model*

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e,$$

then

$$\alpha = \mu_y - \mu_x'\boldsymbol{\beta} \tag{2.37}$$

and

$$\boldsymbol{\beta} = \operatorname{var}(\boldsymbol{x})^{-1}\operatorname{cov}(\boldsymbol{x}, y). \tag{2.38}$$

2.21 Regression Sub-Vectors

Let the regressors be partitioned as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}. \tag{2.39}$$

[11] The **covariance matrix** between vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ is $\operatorname{cov}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}\left((\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{z} - \mathbb{E}\boldsymbol{z})'\right)$. The (co)variance matrix of the vector $\boldsymbol{x}$ is $\operatorname{var}(\boldsymbol{x}) = \operatorname{cov}(\boldsymbol{x}, \boldsymbol{x}) = \mathbb{E}\left((\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})'\right)$.

We can write the projection of y on $\boldsymbol{x}$ as

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ &= \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + e \\ \mathbb{E}(\boldsymbol{x}e) &= \mathbf{0}. \end{aligned} \tag{2.40}$$

In this section we derive formula for the sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

Partition $\boldsymbol{Q}_{xx}$ comformably with $\boldsymbol{x}$

$$\boldsymbol{Q}_{xx} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1') & \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_2') \\ \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_1') & \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_2') \end{bmatrix}$$

and similarly $\boldsymbol{Q}_{xy}$

$$\boldsymbol{Q}_{xy} = \begin{bmatrix} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\boldsymbol{x}_1 y) \\ \mathbb{E}(\boldsymbol{x}_2 y) \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.4)

$$\boldsymbol{Q}_{xx}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}^{-1} \overset{def}{=} \begin{bmatrix} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{bmatrix}. \tag{2.41}$$

where $\boldsymbol{Q}_{11\cdot2} \overset{def}{=} \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot1} \overset{def}{=} \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$. Thus

$$\begin{aligned} \boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\ &= \begin{bmatrix} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q}_{11\cdot2}^{-1}\left(\boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y}\right) \\ \boldsymbol{Q}_{22\cdot1}^{-1}\left(\boldsymbol{Q}_{2y} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{1y}\right) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{1y\cdot2} \\ \boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{2y\cdot1} \end{pmatrix} \end{aligned}$$

We have shown that

$$\begin{aligned} \boldsymbol{\beta}_1 &= \boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{1y\cdot2} \\ \boldsymbol{\beta}_2 &= \boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{2y\cdot1} \end{aligned}$$

2.22 Coefficient Decomposition

In the previous section we derived formulae for the coefficient sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We now use these formulae to give a useful interpretation of the coefficients in terms of an iterated projection.

Take equation (2.40) for the case $\dim(x_1) = 1$ so that $\beta_1 \in \mathbb{R}$.

$$y = x_1\beta_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + e. \tag{2.42}$$

Now consider the projection of x_1 on $\boldsymbol{x}_2$:

$$\begin{aligned} x_1 &= \boldsymbol{x}_2'\boldsymbol{\gamma}_2 + u_1 \\ \mathbb{E}(\boldsymbol{x}_2 u_1) &= \mathbf{0}. \end{aligned}$$

From (2.24) and (2.34), $\boldsymbol{\gamma}_2 = \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\mathbb{E}u_1^2 = \boldsymbol{Q}_{11\cdot2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$. We can also calculate that

$$\mathbb{E}(u_1 y) = \mathbb{E}\left(\left(x_1 - \boldsymbol{\gamma}_2'\boldsymbol{x}_2\right)y\right) = \mathbb{E}(x_1 y) - \boldsymbol{\gamma}_2'\mathbb{E}(\boldsymbol{x}_2 y) = \boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y} = \boldsymbol{Q}_{1y\cdot2}.$$

We have found that

$$\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1y\cdot 2} = \frac{\mathbb{E}(u_1 y)}{\mathbb{E}u_1^2}$$

the coefficient from the simple regression of y on u_1.

What this means is that in the multivariate projection equation (2.42), the coefficient β_1 equals the projection coefficient from a regression of y on u_1, the error from a projection of x_1 on the other regressors $\boldsymbol{x}_2$. The error u_1 can be thought of as the component of x_1 which is not linearly explained by the other regressors. Thus the coefficient β_1 equals the linear effect of x_1 on y, after stripping out the effects of the other variables.

There was nothing special in the choice of the variable x_1. This derivation applies symmetrically to all coefficients in a linear projection. Each coefficient equals the simple regression of y on the error from a projection of that regressor on all the other regressors. Each coefficient equals the linear effect of that variable on y, after linearly controlling for all the other regressors.

2.23 Omitted Variable Bias

Again, let the regressors be partitioned as in (2.39). Consider the projection of y on $\boldsymbol{x}_1$ only. Perhaps this is done because the variables $\boldsymbol{x}_2$ are not observed. This is the equation

$$\begin{aligned} y &= \boldsymbol{x}_1'\boldsymbol{\gamma}_1 + u \\ \mathbb{E}(\boldsymbol{x}_1 u) &= \mathbf{0}. \end{aligned} \tag{2.43}$$

Notice that we have written the coefficient on $\boldsymbol{x}_1$ as $\boldsymbol{\gamma}_1$ rather than $\boldsymbol{\beta}_1$ and the error as u rather than e. This is because (2.43) is different than (2.40). Goldberger (1991) introduced the catchy labels **long regression** for (2.40) and **short regression** for (2.43) to emphasize the distinction.

Typically, $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \boldsymbol{\gamma}_1 &= \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1 y\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1\left(\boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + e\right)\right) \\ &= \boldsymbol{\beta}_1 + \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_2'\right)\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2 \end{aligned}$$

where

$$\boldsymbol{\Gamma} = \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_2'\right)$$

is the coefficient matrix from a projection of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$.

Observe that $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1$ unless $\boldsymbol{\Gamma} = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$. Thus the short and long regressions have different coefficients on $\boldsymbol{x}_1$. They are the same only under one of two conditions. First, if the projection of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$ yields a set of zero coefficients (they are uncorrelated), or second, if the coefficient on $\boldsymbol{x}_2$ in (2.40) is zero. In general, the coefficient in (2.43) is $\boldsymbol{\gamma}_1$ rather than $\boldsymbol{\beta}_1$. The difference $\boldsymbol{\Gamma}\boldsymbol{\beta}_2$ between $\boldsymbol{\gamma}_1$ and $\boldsymbol{\beta}_1$ is known as **omitted variable bias**. It is the consequence of omission of a relevant correlated variable.

To avoid omitted variables bias the standard advice is to include all potentially relevant variables in estimated models. By construction, the general model will be free of such bias. Unfortunately in many cases it is not feasible to completely follow this advice as many desired variables are not observed. In this case, the possibility of omitted variables bias should be acknowledged and discussed in the course of an empirical investigation.

For example, suppose y is log wages, x_1 is education, and x_2 is intellectual ability. It seems reasonable to suppose that education and intellectual ability are positively correlated (highly able individuals attain higher levels of education) which means $\Gamma > 0$. It also seems reasonable to suppose that conditional on education, individuals with higher intelligence will earn higher wages

on average, so that $\beta_2 > 0$. This implies that $\Gamma\beta_2 > 0$ and $\gamma_1 = \beta_1 + \Gamma\beta_2 > \beta_1$. Therefore, it seems reasonable to expect that in a regression of wages on education with ability omitted, the coefficient on education is higher than in a regression where ability is included. In other words, in this context the omitted variable biases the regression coefficient upwards.

2.24 Best Linear Approximation

There are alternative ways we could construct a linear approximation $\boldsymbol{x}'\boldsymbol{\beta}$ to the conditional mean $m(\boldsymbol{x})$. In this section we show that one alternative approach turns out to yield the same answer as the best linear predictor.

We start by defining the mean-square approximation error of $\boldsymbol{x}'\boldsymbol{\beta}$ to $m(\boldsymbol{x})$ as the expected squared difference between $\boldsymbol{x}'\boldsymbol{\beta}$ and the conditional mean $m(\boldsymbol{x})$

$$d(\boldsymbol{\beta}) = \mathbb{E}\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2. \tag{2.44}$$

The function $d(\boldsymbol{\beta})$ is a measure of the deviation of $\boldsymbol{x}'\boldsymbol{\beta}$ from $m(\boldsymbol{x})$. If the two functions are identical then $d(\boldsymbol{\beta}) = 0$, otherwise $d(\boldsymbol{\beta}) > 0$. We can also view the mean-square difference $d(\boldsymbol{\beta})$ as a density-weighted average of the function $(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta})^2$, since

$$d(\boldsymbol{\beta}) = \int_{\mathbb{R}^k} \left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}$$

where $f_{\boldsymbol{x}}(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$.

We can then define the best linear approximation to the conditional $m(\boldsymbol{x})$ as the function $\boldsymbol{x}'\boldsymbol{\beta}$ obtained by selecting $\boldsymbol{\beta}$ to minimize $d(\boldsymbol{\beta})$:

$$\boldsymbol{\beta} = \underset{\boldsymbol{b} \in \mathbb{R}^k}{\operatorname{argmin}}\, d(\boldsymbol{b}). \tag{2.45}$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.21) selects $\boldsymbol{\beta}$ to minimize the expected squared prediction error, while the best linear approximation (2.45) selects $\boldsymbol{\beta}$ to minimize the expected squared approximation error.

Despite the different definitions, it turns out that the best linear predictor and the best linear approximation are identical. By the same steps as in (2.18) plus an application of conditional expectations we can find that

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} m(\boldsymbol{x})\right) \tag{2.46}$$

$$= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} y\right) \tag{2.47}$$

(see Exercise 2.19). Thus (2.45) equals (2.21). We conclude that the definition (2.45) can be viewed as an alternative motivation for the linear projection coefficient.

2.25 Normal Regression

Suppose the variables $(y, \boldsymbol{x})$ are jointly normally distributed. Consider the best linear predictor of y given $\boldsymbol{x}$

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e$$
$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} y\right).$$

Since the error e is a linear transformation of the normal vector $(y, \boldsymbol{x})$, it follows that $(e, \boldsymbol{x})$ is jointly normal, and since they are jointly normal and uncorrelated (since $\mathbb{E}\left(\boldsymbol{x} e\right) = 0$) they are also independent (see Appendix B.9). Independence implies that

$$\mathbb{E}\left(e \mid \boldsymbol{x}\right) = \mathbb{E}\left(e\right) = 0$$

and

$$\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \mathbb{E}\left(e^2\right) = \sigma^2$$

which are properties of a homoskedastic linear CEF.

We have shown that when $(y, \boldsymbol{x})$ are jointly normally distributed, they satisfy a normal linear CEF

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e$$

where

$$e \sim \mathrm{N}(0, \sigma^2)$$

is independent of $\boldsymbol{x}$.

This is an alternative (and traditional) motivation for the linear CEF model. This motivation has limited merit in econometric applications since economic data is typically non-normal.

2.26 Regression to the Mean

The term **regression** originated in an influential paper by Francis Galton published in 1886, where he examined the joint distribution of the stature (height) of parents and children. Effectively, he was estimating the conditional mean of children's height given their parent's height. Galton discovered that this conditional mean was approximately linear with a slope of 2/3. This implies that *on average* a child's height is more mediocre (average) than his or her parent's height. Galton called this phenomenon **regression to the mean**, and the label **regression** has stuck to this day to describe most conditional relationships.

One of Galton's fundamental insights was to recognize that if the marginal distributions of y and x are the same (e.g. the heights of children and parents in a stable environment) then the regression slope in a linear projection is always less than one.

To be more precise, take the simple linear projection

$$y = x\beta + \alpha + e \tag{2.48}$$

where y equals the height of the child and x equals the height of the parent. Assume that y and x have the same mean, so that $\mu_y = \mu_x = \mu$. Then from (2.37)

$$\alpha = (1 - \beta)\,\mu$$

so we can write the linear projection (2.48) as

$$\mathcal{P}\left(y \mid x\right) = (1 - \beta)\,\mu + x\beta.$$

This shows that the projected height of the child is a weighted average of the population average height μ and the parent's height x, with the weight equal to the regression slope β. When the height distribution is stable across generations, so that $\mathrm{var}(y) = \mathrm{var}(x)$, then this slope is the simple correlation of y and x. Using (2.38)

$$\beta = \frac{\mathrm{cov}\left(x, y\right)}{\mathrm{var}(x)} = \mathrm{corr}(x, y).$$

By the properties of correlation (e.g. equation (B.7) in the Appendix), $-1 \leq \mathrm{corr}(x, y) \leq 1$, with $\mathrm{corr}(x, y) = 1$ only in the degenerate case $y = x$. Thus if we exclude degeneracy, β is strictly less than 1.

This means that on average a child's height is more mediocre (closer to the population average) than the parent's.

Sir Francis Galton

Sir Francis Galton (1822-1911) of England was one of the leading figures in late 19th century statistics. In addition to inventing the concept of regression, he is credited with introducing the concepts of correlation, the standard deviation, and the bivariate normal distribution. His work on heredity made a significant intellectual advance by examing the joint distributions of observables, allowing the application of the tools of mathematical statistics to the social sciences.

A common error – known as the **regression fallacy** – is to infer from $\beta < 1$ that the population is **converging**, meaning that its variance is declining towards zero. This is a fallacy because we derived the implication $\beta < 1$ under the assumption of constant means and variances. So certainly $\beta < 1$ does not imply that the variance y is less than than the variance of x.

Another way of seeing this is to examine the conditions for convergence in the context of equation (2.48). Since x and e are uncorrelated, it follows that

$$\operatorname{var}(y) = \beta^2 \operatorname{var}(x) + \operatorname{var}(e).$$

Then $\operatorname{var}(y) < \operatorname{var}(x)$ if and only if

$$\beta^2 < 1 - \frac{\operatorname{var}(e)}{\operatorname{var}(x)}$$

which is not implied by the simple condition $|\beta| < 1$.

The regression fallacy arises in related empirical situations. Suppose you sort families into groups by the heights of the parents, and then plot the average heights of each subsequent generation over time. If the population is stable, the regression property implies that the plots lines will converge – children's height will be more average than their parents. The regression fallacy is to incorrectly conclude that the population is converging. A message to be learned from this example is that such plots are misleading for inferences about convergence.

The regression fallacy is subtle. It is easy for intelligent economists to succumb to its temptation. A famous example is *The Triumph of Mediocrity in Business* by Horace Secrist, published in 1933. In this book, Secrist carefully and with great detail documented that in a sample of department stores over 1920-1930, when he divided the stores into groups based on 1920-1921 profits, and plotted the average profits of these groups for the subsequent 10 years, he found clear and persuasive evidence for convergence "toward mediocrity". Of course, there was no discovery – regression to the mean is a necessary feature of stable distributions.

2.27 Reverse Regression

Galton noticed another interesting feature of the bivariate distribution. There is nothing special about a regression of y on x. We can also regress x on y. (In his heredity example this is the best linear predictor of the height of parents given the height of their children.) This regression takes the form

$$x = y\beta^* + \alpha^* + e^*. \tag{2.49}$$

This is sometimes called the **reverse regression**. In this equation, the coefficients α^*, β^* and error e^* are defined by linear projection. In a stable population we find that

$$\beta^* = \operatorname{corr}(x, y) = \beta$$

$$\alpha^* = (1 - \beta)\,\mu = \alpha$$

which are exactly the same as in the projection of y on x! The intercept and slope have exactly the same values in the forward and reverse projections!

While this algebraic discovery is quite simple, it is counter-intuitive. Instead, a common yet mistaken guess for the form of the reverse regression is to take the equation (2.48), divide through by β and rewrite to find the equation

$$x = y\frac{1}{\beta} - \frac{\alpha}{\beta} - \frac{1}{\beta}e \tag{2.50}$$

suggesting that the projection of x on y should have a slope coefficient of $1/\beta$ instead of β, and intercept of $-\alpha/\beta$ rather than α. What went wrong? Equation (2.50) is perfectly valid, because it is a simple manipulation of the valid equation (2.48). The trouble is that (2.50) is neither a CEF nor a linear projection. Inverting a projection (or CEF) does not yield a projection (or CEF). Instead, (2.49) is a valid projection, not (2.50).

In any event, Galton's finding was that when the variables are standardized, the slope in both projections (y on x, and x and y) equals the correlation, and both equations exhibit regression to the mean. It is not a causal relation, but a natural feature of all joint distributions.

2.28 Limitations of the Best Linear Predictor

Let's compare the linear projection and linear CEF models.

From Theorem 2.8.1.4 we know that the CEF error has the property $\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$. Thus a linear CEF is a linear projection. However, the converse is not true as the projection error does not necessarily satisfy $\mathbb{E}(e \mid \boldsymbol{x}) = 0$. Furthermore, the linear projection may be a poor approximation to the CEF.

To see these points in a simple example, suppose that the true process is $y = x + x^2$ with $x \sim \mathrm{N}(0, 1)$. In this case the true CEF is $m(x) = x + x^2$ and there is no error. Now consider the linear projection of y on x and a constant, namely the model $y = \beta x + \alpha + u$. Since $x \sim \mathrm{N}(0, 1)$ then x and x^2 are uncorrelated the linear projection takes the form $\mathcal{P}(y \mid x) = x + 1$. This is quite different from the true CEF $m(x) = x + x^2$. The projection error equals $e = x^2 - 1$, which is a deterministic function of x, yet is uncorrelated with x. We see in this example that a projection error need not be a CEF error, and a linear projection can be a poor approximation to the CEF.

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is non-linear. We illustrate the issue in Figure 2.10 for a constructed[12] joint distribution of y and x. The solid line is the non-linear CEF of y given x. The data are divided in two – Group 1 and Group 2 – which have different marginal distributions for the regressor x, and Group 1 has a lower mean value of x than Group 2. The separate linear projections of y on x for these two groups are displayed in the Figure by the dashed lines. These two projections are distinct approximations to the CEF. A defect with linear projection is that it leads to the incorrect conclusion that the effect of x on y is different for individuals in the two groups. This conclusion is incorrect because in fact there is no difference in the conditional mean function. The apparant difference is a by-product of a linear approximation to a non-linear mean, combined with different marginal distributions for the conditioning variables.

2.29 Random Coefficient Model

A model which is notationally similar to but conceptually distinct from the linear CEF model is the linear random coefficient model. It takes the form

$$y = \boldsymbol{x}'\boldsymbol{\eta}$$

[12] The x in Group 1 are $\mathrm{N}(2, 1)$ and those in Group 2 are $\mathrm{N}(4, 1)$, and the conditional distribution of y given x is $\mathrm{N}(m(x), 1)$ where $m(x) = 2x - x^2/6$.

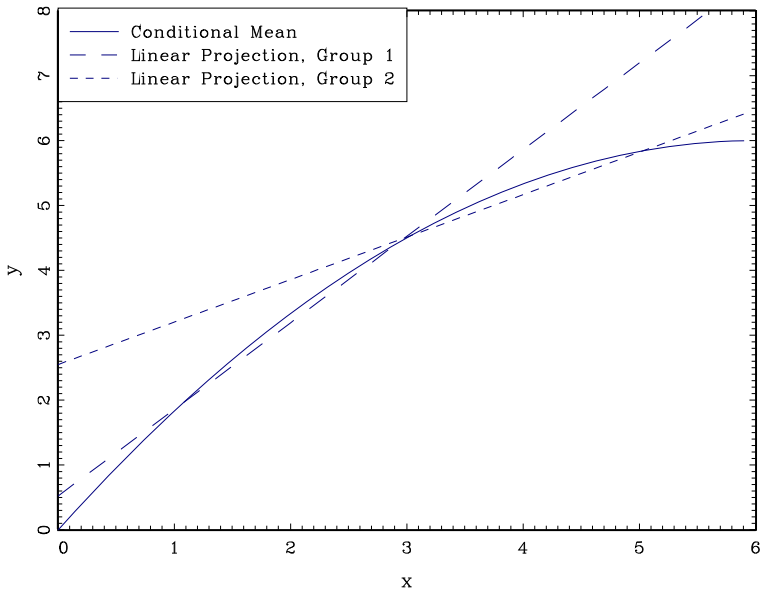


Figure 2.10: Conditional Mean and Two Linear Projections

where the individual-specific coefficient $\boldsymbol{\eta}$ is random and independent of $\boldsymbol{x}$. For example, if $\boldsymbol{x}$ is years of schooling and y is log wages, then $\boldsymbol{\eta}$ is the individual-specific returns to schooling. If a person obtains an extra year of schooling, $\boldsymbol{\eta}$ is the actual change in their wage. The random coefficient model allows the returns to schooling to vary in the population. Some individuals might have a high return to education (a high $\boldsymbol{\eta}$) and others a low return, possibly 0, or even negative.

In the linear CEF model the regressor coefficient equals the regression derivative – the change in the conditional mean due to a change in the regressors, $\boldsymbol{\beta} = \nabla m(\boldsymbol{x})$. This is not the effect on a given individual, it is the effect on the population average. In contrast, in the random coefficient model, the random vector $\boldsymbol{\eta} = \nabla (\boldsymbol{x}'\boldsymbol{\eta})$ is the true causal effect – the change in the response variable y itself due to a change in the regressors.

It is interesting, however, to discover that the linear random coefficient model implies a linear CEF. To see this, let $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ denote the mean and covariance matrix of $\boldsymbol{\eta}$ **:**

$$\begin{aligned}\boldsymbol{\beta} &= \mathbb{E}(\boldsymbol{\eta}) \\ \boldsymbol{\Sigma} &= \operatorname{var}(\boldsymbol{\eta})\end{aligned}$$

and then decompose the random coefficient as

$$\boldsymbol{\eta} = \boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{u}$ is distributed independently of $\boldsymbol{x}$ with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Then we can write

$$\mathbb{E}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\mathbb{E}(\boldsymbol{\eta} \mid \boldsymbol{x}) = \boldsymbol{x}'\mathbb{E}(\boldsymbol{\eta}) = \boldsymbol{x}'\boldsymbol{\beta}$$

so the CEF is linear in $\boldsymbol{x}$**,** and the coefficients $\boldsymbol{\beta}$ equal the mean of the random coefficient $\boldsymbol{\eta}$.

We can thus write the equation as a linear CEF

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e \tag{2.51}$$

where $e = \boldsymbol{x}'\boldsymbol{u}$ and $\boldsymbol{u} = \boldsymbol{\eta} - \boldsymbol{\beta}$. The error is conditionally mean zero:

$$\mathbb{E}(e \mid \boldsymbol{x}) = 0.$$

Furthermore

$$\begin{aligned}\operatorname{var}(e \mid \boldsymbol{x}) &= \boldsymbol{x}' \operatorname{var}(\boldsymbol{\eta}) \boldsymbol{x} \\ &= \boldsymbol{x}' \boldsymbol{\Sigma} \boldsymbol{x}\end{aligned}$$

so the error is conditionally heteroskedastic with its variance a quadratic function of $\boldsymbol{x}$.

Theorem 2.29.1 *In the linear random coefficient model $y = \boldsymbol{x}'\boldsymbol{\eta}$ with $\boldsymbol{\eta}$ independent of $\boldsymbol{x}$, $\mathbb{E}\|\boldsymbol{x}\|^2 < \infty$, and $\mathbb{E}\|\boldsymbol{\eta}\|^2 < \infty$, then*

$$\begin{aligned}\mathbb{E}(y \mid \boldsymbol{x}) &= \boldsymbol{x}'\boldsymbol{\beta} \\ \operatorname{var}(y \mid \boldsymbol{x}) &= \boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}\end{aligned}$$

where $\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{\eta})$ and $\boldsymbol{\Sigma} = \operatorname{var}(\boldsymbol{\eta})$.

2.30 Causal Effects

So far we have avoided the concept of causality, yet often the underlying goal of an econometric analysis is to uncover a causal relationship between variables. It is often of great interest to understand the causes and effects of decisions, actions, and policies. For example, we may be interested in the effect of class sizes on test scores, police expenditures on crime rates, climate change on economic activity, years of schooling on wages, institutional structure on growth, the effectiveness of rewards on behavior, the consequences of medical procedures for health outcomes, or any variety of possible causal relationships. In each case, the goal is to understand what is the actual effect on the outcome y due to a change in the input x. We are not just interested in the conditional mean or linear projection, we would like to know the actual change.

Two inherent barriers are that the causal effect is typically specific to an individual and that it is unobserved.

Consider the effect of schooling on wages. The causal effect is the actual difference a person would receive in wages if we could change their level of education *holding all else constant*. This is specific to each individual as their employment outcomes in these two distinct situations is individual. The causal effect is unobserved because the most we can observe is their actual level of education and their actual wage, but not the counterfactual wage if their education had been different.

To be even more specific, suppose that there are two individuals, Jennifer and George, and both have the possibility of being high-school graduates or college graduates, but both would have received different wages given their choices. For example, suppose that Jennifer would have earned \$10 an hour as a high-school graduate and \$20 an hour as a college graduate while George would have earned \$8 as a high-school graduate and \$12 as a college graduate. In this example the causal effect of schooling is \$10 a hour for Jennifer and \$4 an hour for George. The causal effects are specific to the individual and neither causal effect is observed.

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. We can write a full model for the response variable y as

$$y = h(x_1, \boldsymbol{x}_2, \boldsymbol{u}) \tag{2.52}$$

where x_1 and $\boldsymbol{x}_2$ are the observed variables, $\boldsymbol{u}$ is an $\ell \times 1$ unobserved random factor, and h is a functional relationship. This framework includes as a special case the random coefficient model

(2.29) studied earlier. 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 $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.

Definition 2.30.1 *In the model (2.52) the* ***causal effect*** *of x_1 on y is*

$$C(x_1, \boldsymbol{x}_2, \boldsymbol{u}) = \boldsymbol{\nabla}_1 h(x_1, \boldsymbol{x}_2, \boldsymbol{u}), \tag{2.53}$$

the change in y due to a change in x_1, holding $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.

To understand this concept, imagine taking a single individual. As far as our structural model is concerned, this person is described by their observables x_1 and $\boldsymbol{x}_2$ and their unobservables $\boldsymbol{u}$. In a wage regression the unobservables would include characteristics such as the person's abilities, skills, work ethic, interpersonal connections, and preferences. The causal effect of x_1 (say, education) is the change in the wage as x_1 changes, holding constant all other observables **and** unobservables.

It may be helpful to understand that (2.53) is a definition, and does not necessarily describe causality in a fundamental or experimental sense. Perhaps it would be more appropriate to label (2.53) as a **structural effect** (the effect within the structural model).

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

$$y(x_1) = h(x_1, \boldsymbol{x}_2, \boldsymbol{u})$$

where the notation implies that $y(x_1)$ is holding $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.

A popular example arises in the analysis of treatment effects with a binary regressor x_1. Let $x_1 = 1$ indicate treatment (e.g. a medical procedure) and $x_1 = 0$ indicate non-treatment. In this case $y(x_1)$ can be written

$$\begin{aligned} y(0) &= h(0, \boldsymbol{x}_2, \boldsymbol{u}) \\ y(1) &= h(1, \boldsymbol{x}_2, \boldsymbol{u}). \end{aligned}$$

In the literature on treatment effects, it is common to refer to $y(0)$ and $y(1)$ as the latent outcomes associated with non-treatment and treatment, respectively. That is, for a given individual, $y(0)$ is the health outcome if there is no treatment, and $y(1)$ is the health outcome if there is treatment. The causal effect of treatment for the individual is the change in their health outcome due to treatment – the change in y as we hold both $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant:

$$C(\boldsymbol{x}_2, \boldsymbol{u}) = y(1) - y(0).$$

This is random (a function of $\boldsymbol{x}_2$ and $\boldsymbol{u}$) as both potential outcomes $y(0)$ and $y(1)$ are different across individuals.

In a sample, we cannot observe both outcomes from the same individual, we only observe the realized value

$$y = \begin{cases} y(0) & \text{if} \quad x_1 = 0 \\ \\ y(1) & \text{if} \quad x_1 = 1. \end{cases}$$

As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the average causal effect.

Definition 2.30.2 *In the model (2.52) the* ***average causal effect*** *of* x_1 *on* y *conditional on* $\boldsymbol{x}_2$ *is*

$$\begin{aligned} ACE(x_1, \boldsymbol{x}_2) &= \mathbb{E}\left(C(x_1, \boldsymbol{x}_2, \boldsymbol{u}) \mid x_1, \boldsymbol{x}_2\right) \\ &= \int_{\mathbb{R}^\ell} \boldsymbol{\nabla}_1 h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u} \end{aligned} \tag{2.54}$$

where $f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2)$ *is the conditional density of* $\boldsymbol{u}$ *given* $x_1, \boldsymbol{x}_2$.

We can think of the average causal effect $ACE(x_1, \boldsymbol{x}_2)$ as the average effect in the general population. In our Jennifer & George schooling example given earlier, supposing that half of the population are Jennifer's and the other half George's, then the average causal effect of college is $(10+4)/2 = \$7$ an hour. This is not the individual causal effect, it is the average of the causal effect across all individuals in the population. Given data on only educational attainment and wages, the ACE of \$7 is the best we can hope to learn.

When we conduct a regression analysis (that is, consider the regression of observed wages on educational attainment) we might hope that the regression reveals the average causal effect. Technically, that the regression derivative (the coefficient on education) equals the ACE. Is this the case? In other words, what is the relationship between the average causal effect $ACE(x_1, \boldsymbol{x}_2)$ and the regression derivative $\boldsymbol{\nabla}_1 m\left(x_1, \boldsymbol{x}_2\right)$? Equation (2.52) implies that the CEF is

$$\begin{aligned} m(x_1, \boldsymbol{x}_2) &= \mathbb{E}\left(h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) \mid x_1, \boldsymbol{x}_2\right) \\ &= \int_{\mathbb{R}^\ell} h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u}, \end{aligned}$$

the average causal equation, averaged over the conditional distribution of the unobserved component $\boldsymbol{u}$.

Applying the marginal effect operator, the regression derivative is

$$\begin{aligned} \boldsymbol{\nabla}_1 m(x_1, \boldsymbol{x}_2) &= \int_{\mathbb{R}^\ell} \boldsymbol{\nabla}_1 h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u} \\ &\quad + \int_{\mathbb{R}^\ell} h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) \boldsymbol{\nabla}_1 f(\boldsymbol{u} | x_1, \boldsymbol{x}_2) d\boldsymbol{u} \\ &= ACE(x_1, \boldsymbol{x}_2) + \int_{\mathbb{R}^\ell} h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) \boldsymbol{\nabla}_1 f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u}. \end{aligned} \tag{2.55}$$

Equation (2.55) shows that in general, the regression derivative does not equal the average causal effect. The difference is the second term on the right-hand-side of (2.55). The regression derivative and ACE equal in the special case when this term equals zero, which occurs when $\boldsymbol{\nabla}_1 f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = 0$, that is, when the conditional density of $\boldsymbol{u}$ given $(x_1, \boldsymbol{x}_2)$ does not depend on x_1. When this condition holds then the regression derivative equals the ACE, which means that regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.

The condition is sufficiently important that it has a special name in the treatment effects literature.

Definition 2.30.3 ***Conditional Independence Assumption (CIA).*** *Conditional on* $\boldsymbol{x}_2$, *the random variables* x_1 *and* $\boldsymbol{u}$ *are statistically independent.*

The CIA implies $f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = f(\boldsymbol{u} \mid \boldsymbol{x}_2)$ does not depend on x_1, and thus $\boldsymbol{\nabla}_1 f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = 0$. Thus the CIA implies that $\boldsymbol{\nabla}_1 m(x_1, \boldsymbol{x}_2) = ACE(x_1, \boldsymbol{x}_2)$, the regression derivative equals the average causal effect.

Theorem 2.30.1 *In the structural model (2.52), the Conditional Independence Assumption implies*

$$\boldsymbol{\nabla}_1 m(x_1, \boldsymbol{x}_2) = ACE(x_1, \boldsymbol{x}_2)$$

the regression derivative equals the average causal effect for x_1 on y conditional on $\boldsymbol{x}_2$.

This is a fascinating result. It shows that whenever the unobservable is independent of the treatment variable (after conditioning on appropriate regressors) the regression derivative equals the average causal effect. In this case, the CEF has causal economic meaning, giving strong justification to estimation of the CEF. Our derivation also shows the critical role of the CIA. If CIA fails, then the equality of the regression derivative and ACE fails.

This theorem is quite general. It applies equally to the treatment-effects model where x_1 is binary or to more general settings where x_1 is continuous.

It is also helpful to understand that the CIA is weaker than full independence of $\boldsymbol{u}$ from the regressors $(x_1, \boldsymbol{x}_2)$. The CIA was introduced precisely as a minimal sufficient condition to obtain the desired result. Full independence implies the CIA and implies that each regression derivative equals that variable's average causal effect, but full independence is not necessary in order to causally interpret a subset of the regressors.

To illustrate, let's return to our education example involving a population with equal numbers of Jennifer's and George's. Recall that Jennifer earns \$10 as a high-school graduate and \$20 as a college graduate (and so has a causal effect of \$10) while George earns \$8 as a high-school graduate and \$12 as a college graduate (so has a causal effect of \$4). Given this information, the average causal effect of college is \$7, which is what we hope to learn from a regression analysis.

Now suppose that while in high school all students take an aptitude test, and if a student gets a high (H) score he or she goes to college with probability 3/4, and if a student gets a low (L) score he or she goes to college with probability 1/4. Suppose further that Jennifer's get an aptitude score of H with probability 3/4, while George's get a score of H with probability 1/4. Given this situation, 62.5% of Jennifer's will go to college[13], while 37.5% of George's will go to college[14].

An econometrician who randomly samples 32 individuals and collects data on educational attainment and wages will find the following wage distribution:

	\$8	\$10	\$12	\$20	Mean
High-School Graduate	10	6	0	0	\$8.75
College Graduate	0	0	6	10	\$17.00

Let *college* denote a dummy variable taking the value of 1 for a college graduate, otherwise 0. Thus the regression of wages on college attendence takes the form

$$\mathbb{E}(wage \mid college) = 8.25 college + 8.75.$$

The coefficient on the college dummy, \$8.25, is the regression derivative, and the implied wage effect of college attendence. But \$8.25 overstates the average causal effect of \$7. The reason is because

[13] $\Pr(College|Jennifer) = \Pr(College|H)\Pr(H|Jennifer) + \Pr(College|L)\Pr(L|Jennifer) = (3/4)^2 + (1/4)^2$

[14] $\Pr(College|George) = \Pr(College|H)\Pr(H|George) + \Pr(College|L)\Pr(L|George) = (3/4)(1/4) + (1/4)(3/4)$

the CIA fails. In this model the unobservable $\boldsymbol{u}$ is the individual's type (Jennifer or George) which is not independent of the regressor x_1 (education), since Jennifer is more likely to go to college than George. Since Jennifer's causal effect is higher than George's, the regression derivative overstates the ACE. The coefficient \$8.25 is not the average benefit of college attendence, rather it is the observed difference in realized wages in a population whose decision to attend college is correlated with their individual causal effect. At the risk of repeating myself, in this example, \$8.25 is the true regression derivative, it is the difference in average wages between those with a college education and those without. It is not, however, the average causal effect of college education in the population.

This does not mean that it is impossible to estimate the ACE. The key is conditioning on the appropriate variables. The CIA says that we need to find a variable x_2 such that conditional on x_2, $\boldsymbol{u}$ and x_1 (type and education) are independent. In this example a variable which will achieve this is the aptitude test score. The decision to attend college was based on the test score, not on an individual's type. Thus educational attainment and type are independent once we condition on the test score.

This also alters the ACE. Notice that Definition 2.30.2 is a function of x_2 (the test score). Among the students who receive a high test score, 3/4 are Jennifer's and 1/4 are George's. Thus the ACE for students with a score of H is $(3/4) \times 10 + (1/4) \times 4 = \$8.50$. Among the students who receive a low test score, 1/4 are Jennifer's and 3/4 are George's. Thus the ACE for students with a score of L is $(1/4) \times 10 + (3/4) \times 4 = \$5.50$. The ACE varies between these two observable groups (those with high test scores and those with low test scores). Again, we would hope to be able to learn the ACE from a regression analysis, this time from a regression of wages on education and test scores.

To see this in the wage distribution, suppose that the econometrician collects data on the aptitude test score as well as education and wages. Given a random sample of 32 individuals we would expect to find the following wage distribution:

	\$8	\$10	\$12	\$20	Mean
High-School Graduate + High Test Score	1	3	0	0	\$9.50
College Graduate + High Test Score	0	0	3	9	\$18.00
High-School Graduate + Low Test Score	9	3	0	0	\$8.50
College Graduate + Low Test Score	0	0	3	1	\$14.00

Define the dummy variable *highscore* which takes the value 1 for students who received a high test score, else zero. The regression of wages on college attendence and test scores (with interactions) takes the form

$$\mathbb{E}\left(wage \mid college, highscore\right) = 1.00 highscore + 5.50 college + 3.00 highscore \times college + 8.50.$$

The cofficient on *college*, \$5.50, is the regression derivative of college attendence for those with low test scores, and the sum of this coefficient with the interaction coefficient, \$8.50, is the regression derivative for college attendence for those with high test scores. These equal the average causal effect.

In this example, by conditioning on the aptitude test score, the average causal effect of education on wages can be learned from a regression analyis. What this shows is that by conditioning on the proper variables, it may be possible to achieve the CIA, in which case regression analysis measures average causal effects.

2.31 Expectation: Mathematical Details*

We define the **mean** or **expectation** $\mathbb{E}y$ of a random variable y as follows. If y is discrete on the set $\{\tau_1, \tau_2, ...\}$ then

$$\mathbb{E}y = \sum_{j=1}^{\infty} \tau_j \Pr\left(y = \tau_j\right),$$

and if y is continuous with density f then

$$\mathbb{E}y = \int_{-\infty}^{\infty} yf(y)dy.$$

We can unify these definitions by writing the expectation as the Lebesgue integral with respect to the distribution function F

$$\mathbb{E}y = \int_{-\infty}^{\infty} ydF(y). \tag{2.56}$$

In the event that the integral (2.56) is not finite, separately evaluate the two integrals

$$I_1 = \int_{0}^{\infty} ydF(y) \tag{2.57}$$

$$I_2 = -\int_{-\infty}^{0} ydF(y). \tag{2.58}$$

If $I_1 = \infty$ and $I_2 < \infty$ then it is typical to define $\mathbb{E}y = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $\mathbb{E}y = -\infty$. However, if both $I_1 = \infty$ and $I_2 = \infty$ then $\mathbb{E}y$ is undefined. If

$$\mathbb{E}\left|y\right| = \int_{-\infty}^{\infty} \left|y\right| dF(y) = I_1 + I_2 < \infty$$

then $\mathbb{E}y$ exists and is finite. In this case it is common to say that the mean $\mathbb{E}y$ is "well-defined".

More generally, y has a finite r'th moment if

$$\mathbb{E}\left|y\right|^r < \infty. \tag{2.59}$$

By Liapunov's Inequality (B.20), (2.59) implies $\mathbb{E}\left|y\right|^s < \infty$ for all $1 \leq s \leq r$. Thus, for example, if the fourth moment is finite then the first, second and third moments are also finite.

It is common in econometric theory to assume that the variables, or certain transformations of the variables, have finite moments of a certain order. How should we interpret this assumption? How restrictive is it?

One way to visualize the importance is to consider the class of Pareto densities given by

$$f(y) = ay^{-a-1}, \qquad y > 1.$$

The parameter a of the Pareto distribution indexes the rate of decay of the tail of the density. Larger a means that the tail declines to zero more quickly. See Figure 2.11 below where we show the Pareto density for $a = 1$ and $a = 2$. The parameter a also determines which moments are finite. We can calculate that

$$\mathbb{E}\left|y\right|^r = \begin{cases} a\int_1^\infty y^{r-a-1}dy = \dfrac{a}{a-r} & \text{if} \quad r < a \\ \\ \infty & \text{if} \quad r \geq a. \end{cases}$$

This shows that if y is Pareto distributed with parameter a, then the r'th moment of y is finite if and only if $r < a$. Higher a means higher finite moments. Equivalently, the faster the tail of the density declines to zero, the more moments are finite.

This connection between tail decay and finite moments is not limited to the Pareto distribution. We can make a similar analysis using a tail bound. Suppose that y has density $f(y)$ which satisfies the bound $f(y) \leq A\left|y\right|^{-a-1}$ for some $A < \infty$ and $a > 0$. Since $f(y)$ is bounded below a scale of a Pareto density, its tail behavior is similarly bounded. This means that for $r < a$

$$\mathbb{E}\left|y\right|^r = \int_{-\infty}^{\infty} \left|y\right|^r f(y)dy \leq \int_{-1}^{1} f(y)dy + 2A\int_{1}^{\infty} y^{r-a-1}dy \leq 1 + \frac{2A}{a-r} < \infty.$$

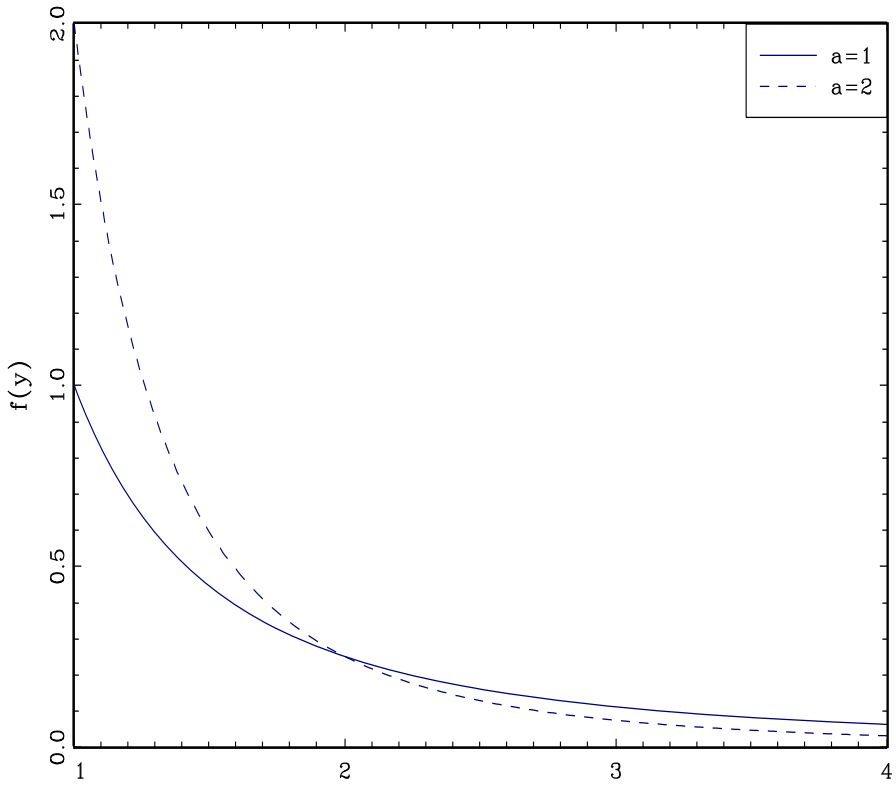


Figure 2.11: Pareto Densities, $a = 1$ and $a = 2$

Thus if the tail of the density declines at the rate $|y|^{-a-1}$ or faster, then y has finite moments up to (but not including) a. Broadly speaking, the restriction that y has a finite r^{th} moment means that the tail of y's density declines to zero faster than y^{-r-1}. The faster decline of the tail means that the probability of observing an extreme value of y is a more rare event.

We complete this section by adding an alternative representation of expectation in terms of the distribution function.

Theorem 2.31.1 *For any non-negative random variable y*

$$\mathbb{E}y = \int_0^\infty \Pr(y > u)\, du$$

Proof of Theorem 2.31.1: Let $F^*(x) = \Pr(y > x) = 1 - F(x)$, where $F(x)$ is the distribution function. By integration by parts

$$\mathbb{E}y = \int_0^\infty y dF(y) = -\int_0^\infty y dF^*(y) = -\left[yF^*(y)\right]_0^\infty + \int_0^\infty F^*(y)dy = \int_0^\infty \Pr(y > u)\, du$$

as stated. ■

2.32 Existence and Uniqueness of the Conditional Expectation*

In Sections 2.3 and 2.6 we defined the conditional mean when the conditioning variables $\boldsymbol{x}$ are discrete and when the variables $(y, \boldsymbol{x})$ have a joint density. We have explored these cases because

these are the situations where the conditional mean is easiest to describe and understand. However, the conditional mean exists quite generally without appealing to the properties of either discrete or continuous random variables.

To justify this claim we now present a deep result from probability theory. What it says is that the conditional mean exists for all joint distributions $(y, \boldsymbol{x})$ for which y has a finite mean.

Theorem 2.32.1 ***Existence of the Conditional Mean***
If $\mathbb{E}|y| < \infty$ then there exists a function $m(\boldsymbol{x})$ such that for all measurable sets $\mathcal{X}$

$$\mathbb{E}\left(1\left(\boldsymbol{x} \in \mathcal{X}\right) y\right) = \mathbb{E}\left(1\left(\boldsymbol{x} \in \mathcal{X}\right) m(\boldsymbol{x})\right). \tag{2.60}$$

The function $m(\boldsymbol{x})$ is almost everywhere unique, in the sense that if $h(\boldsymbol{x})$ satisfies (2.60), then there is a set S such that $\Pr(S) = 1$ and $m(\boldsymbol{x}) = h(\boldsymbol{x})$ for $\boldsymbol{x} \in S$. The function $m(\boldsymbol{x})$ is called the ***conditional mean*** *and is written $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$.*

See, for example, Ash (1972), Theorem 6.3.3.

The conditional mean $m(\boldsymbol{x})$ defined by (2.60) specializes to (2.7) when $(y, \boldsymbol{x})$ have a joint density. The usefulness of definition (2.60) is that Theorem 2.32.1 shows that the conditional mean $m(\boldsymbol{x})$ exists for all finite-mean distributions. This definition allows y to be discrete or continuous, for $\boldsymbol{x}$ to be scalar or vector-valued, and for the components of $\boldsymbol{x}$ to be discrete or continuously distributed.

2.33 Identification*

A critical and important issue in structural econometric modeling is identification, meaning that a parameter is uniquely determined by the distribution of the observed variables. It is relatively straightforward in the context of the unconditional and conditional mean, but it is worthwhile to introduce and explore the concept at this point for clarity.

Let F denote the distribution of the observed data, for example the distribution of the pair (y, x). Let $\mathcal{F}$ be a collection of distributions F. Let θ be a parameter of interest (for example, the mean $\mathbb{E}y$).

Definition 2.33.1 *A parameter $\theta \in \mathbb{R}$ is identified on $\mathcal{F}$ if for all $F \in \mathcal{F}$, there is a uniquely determined value of θ.*

Equivalently, θ is identified if we can write it as a mapping $\theta = g(F)$ on the set $\mathcal{F}$. The restriction to the set $\mathcal{F}$ is important. Most parameters are identified only on a strict subset of the space of all distributions.

Take, for example, the mean $\mu = \mathbb{E}y$. It is uniquely determined if $\mathbb{E}|y| < \infty$, so it is clear that μ is identified for the set $\mathcal{F} = \left\{F : \int_{-\infty}^{\infty} |y| \, dF(y) < \infty\right\}$. However, μ is also well defined when it is either positive or negative infinity. Hence, defining I_1 and I_2 as in (2.57) and (2.58), we can deduce that μ is identified on the set $\mathcal{F} = \{F : \{I_1 < \infty\} \cup \{I_2 < \infty\}\}$.

Next, consider the conditional mean. Theorem 2.32.1 demonstrates that $\mathbb{E}|y| < \infty$ is a sufficient condition for identification.

Theorem 2.33.1 ***Identification of the Conditional Mean***
If $\mathbb{E}|y| < \infty$*, the conditional mean* $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ *is identified almost everywhere.*

It might seem as if identification is a general property for parameters, so long as we exclude degenerate cases. This is true for moments of observed data, but not necessarily for more complicated models. As a case in point, consider the context of censoring. Let y be a random variable with distribution F. Instead of observing y, we observe y^* defined by the censoring rule

$$y^* = \begin{cases} y & \text{if } y \leq \tau \\ \tau & \text{if } y > \tau \end{cases}.$$

That is, y^* is capped at the value τ. A common example is income surveys, where income responses are "top-coded", meaning that incomes above the top code τ are recorded as equalling the top code. The observed variable y^* has distribution

$$F^*(u) = \begin{cases} F(u) & \text{for } u \leq \tau \\ 1 & \text{for } u \geq \tau. \end{cases}$$

We are interested in features of the distribution F not the censored distribution F^*. For example, we are interested in the mean wage $\mu = \mathbb{E}(y)$. The difficulty is that we cannot calculate μ from F^* except in the trivial case where there is no censoring $\Pr(y \geq \tau) = 0$. Thus the mean μ is not generically identified from the censored distribution.

A typical solution to the identification problem is to assume a parametric distribution. For example, let $\mathcal{F}$ be the set of normal distributions $y \sim \mathrm{N}(\mu, \sigma^2)$. It is possible to show that the parameters (μ, σ^2) are identified for all $F \in \mathcal{F}$. That is, if we know that the uncensored distribution is normal, we can uniquely determine the parameters from the censored distribution. This is often called **parametric identification** as identification is restricted to a parametric class of distributions. In modern econometrics this is generally viewed as a second-best solution, as identification has been achieved only through the use of an arbitrary and unverifiable parametric assumption.

A pessimistic conclusion might be that it is impossible to identify parameters of interest from censored data without parametric assumptions. Interestingly, this pessimism is unwarranted. It turns out that we can identify the quantiles q_α of F for $\alpha \leq \Pr(y \leq \tau)$. For example, if 20% of the distribution is censored, we can identify all quantiles for $\alpha \in (0, 0.8)$. This is often called **nonparametric identification** as the parameters are identified without restriction to a parametric class.

What we have learned from this little exercise is that in the context of censored data, moments can only be parametrically identified, while (non-censored) quantiles are nonparametrically identified. Part of the message is that a study of identification can help focus attention on what can be learned from the data distributions available.

2.34 Technical Proofs*

Proof of Theorem 2.7.1: For convenience, assume that the variables have a joint density $f(y, \boldsymbol{x})$. Since $\mathbb{E}(y \mid \boldsymbol{x})$ is a function of the random vector $\boldsymbol{x}$ only, to calculate its expectation we integrate with respect to the density $f_{\boldsymbol{x}}(\boldsymbol{x})$ of $\boldsymbol{x}$, that is

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y \mid \boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) \, d\boldsymbol{x}.$$

Substituting in (2.7) and noting that $f_{y|\boldsymbol{x}}(y|\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) = f(y, \boldsymbol{x})$, we find that the above expression equals

$$\int_{\mathbb{R}^k} \left(\int_{\mathbb{R}} y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\, dy \right) f_{\boldsymbol{x}}(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathbb{R}^k} \int_{\mathbb{R}} y f(y, \boldsymbol{x})\, dy d\boldsymbol{x} = \mathbb{E}(y)$$

the unconditional mean of y. ■

Proof of Theorem 2.7.2: Again assume that the variables have a joint density. It is useful to observe that

$$f(y|\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{f(y, \boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_1, \boldsymbol{x}_2)} \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_1)} = f(y, \boldsymbol{x}_2|\boldsymbol{x}_1), \tag{2.61}$$

the density of $(y, \boldsymbol{x}_2)$ given $\boldsymbol{x}_1$. Here, we have abused notation and used a single symbol f to denote the various unconditional and conditional densities to reduce notational clutter.

Note that

$$\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2)\, dy. \tag{2.62}$$

Integrating (2.62) with respect to the conditional density of $\boldsymbol{x}_2$ given $\boldsymbol{x}_1$, and applying (2.61) we find that

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \mid \boldsymbol{x}_1) &= \int_{\mathbb{R}^{k_2}} \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1)\, d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \left(\int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2)\, dy \right) f(\boldsymbol{x}_2|\boldsymbol{x}_1)\, d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1)\, dy d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y, \boldsymbol{x}_2|\boldsymbol{x}_1)\, dy d\boldsymbol{x}_2 \\
&= \mathbb{E}(y \mid \boldsymbol{x}_1)
\end{aligned}$$

as stated. ■

Proof of Theorem 2.7.3:

$$\mathbb{E}(g(\boldsymbol{x}) y \mid \boldsymbol{x}) = \int_{\mathbb{R}} g(\boldsymbol{x}) y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\, dy = g(\boldsymbol{x}) \int_{\mathbb{R}} y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\, dy = g(\boldsymbol{x}) \mathbb{E}(y \mid \boldsymbol{x})$$

This is (2.9). The assumption that $\mathbb{E}|g(\boldsymbol{x}) y| < \infty$ is required for the first equality to be well-defined. Equation (2.10) follows by applying the Simple Law of Iterated Expectations to (2.9). ■

Proof of Theorem 2.10.2: The assumption that $\mathbb{E}y^2 < \infty$ implies that all the conditional expectations below exist.

Set $z = \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$. By the conditional Jensen's inequality (B.13),

$$(\mathbb{E}(z \mid \boldsymbol{x}_1))^2 \leq \mathbb{E}\left(z^2 \mid \boldsymbol{x}_1\right).$$

Taking unconditional expectations, this implies

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x}_1))^2 \leq \mathbb{E}\left((\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2))^2\right).$$

Similarly,

$$(\mathbb{E}y)^2 \leq \mathbb{E}\left((\mathbb{E}(y \mid \boldsymbol{x}_1))^2\right) \leq \mathbb{E}\left((\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2))^2\right). \tag{2.63}$$

The variables y, $\mathbb{E}(y \mid \boldsymbol{x}_1)$ and $\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$ all have the same mean $\mathbb{E}y$, so the inequality (2.63) implies that the variances are ranked monotonically:

$$0 \leq \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x}_1)\right) \leq \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)\right). \tag{2.64}$$

Next, for $\mu = \mathbb{E}y$ observe that

$$\mathbb{E}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right)\left(\mathbb{E}(y \mid \boldsymbol{x}) - \mu\right) = \mathbb{E}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right)\left(\mathbb{E}(y \mid \boldsymbol{x}) - \mu\right) = 0$$

so the decomposition

$$y - \mu = y - \mathbb{E}(y \mid \boldsymbol{x}) + \mathbb{E}(y \mid \boldsymbol{x}) - \mu$$

satisfies

$$\operatorname{var}(y) = \operatorname{var}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right) + \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x})\right). \tag{2.65}$$

The monotonicity of the variances of the conditional mean (2.64) applied to the variance decomposition (2.65) implies the reverse monotonicity of the variances of the differences, completing the proof. ■

Proof of Theorem 2.8.1. Applying Minkowski's Inequality (B.19) to $e = y - m(\boldsymbol{x})$,

$$\left(\mathbb{E}\left|e\right|^r\right)^{1/r} = \left(\mathbb{E}\left|y - m(\boldsymbol{x})\right|^r\right)^{1/r} \leq \left(\mathbb{E}\left|y\right|^r\right)^{1/r} + \left(\mathbb{E}\left|m(\boldsymbol{x})\right|^r\right)^{1/r} < \infty,$$

where the two parts on the right-hand are finite since $\mathbb{E}\left|y\right|^r < \infty$ by assumption and $\mathbb{E}\left|m(\boldsymbol{x})\right|^r < \infty$ by the Conditional Expectation Inequality (B.14). The fact that $\left(\mathbb{E}\left|e\right|^r\right)^{1/r} < \infty$ implies $\mathbb{E}\left|e\right|^r < \infty$. ■

Proof of Theorem 2.18.1. For part 1, by the Expectation Inequality (B.15), (A.19) and Assumption 2.18.1,

$$\left\|\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right\| \leq \mathbb{E}\left\|\boldsymbol{x}\boldsymbol{x}'\right\| = \mathbb{E}\left\|\boldsymbol{x}\right\|^2 < \infty.$$

Similarly, using the Expectation Inequality (B.15), the Cauchy-Schwarz Inequality (B.17) and Assumption 2.18.1,

$$\left\|\mathbb{E}\left(\boldsymbol{x}y\right)\right\| \leq \mathbb{E}\left\|\boldsymbol{x}y\right\| \leq \left(\mathbb{E}\left\|\boldsymbol{x}\right\|^2\right)^{1/2}\left(\mathbb{E}y^2\right)^{1/2} < \infty.$$

Thus the moments $\mathbb{E}\left(\boldsymbol{x}y\right)$ and $\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ are finite and well defined.

For part 2, the coefficient $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ is well defined since $\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}$ exists under Assumption 2.18.1.

Part 3 follows from Definition 2.18.1 and part 2.

For part 4, first note that

$$\begin{aligned}
\mathbb{E}e^2 &= \mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 \\
&= \mathbb{E}y^2 - 2\mathbb{E}\left(y\boldsymbol{x}'\right)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta} \\
&= \mathbb{E}y^2 - 2\mathbb{E}\left(y\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right) \\
&\leq \mathbb{E}y^2 \\
&< \infty.
\end{aligned}$$

The first inequality holds because $\mathbb{E}\left(y\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ is a quadratic form and therefore necessarily non-negative. Second, by the Expectation Inequality (B.15), the Cauchy-Schwarz Inequality (B.17) and Assumption 2.18.1,

$$\left\|\mathbb{E}\left(\boldsymbol{x}e\right)\right\| \leq \mathbb{E}\left\|\boldsymbol{x}e\right\| = \left(\mathbb{E}\left\|\boldsymbol{x}\right\|^2\right)^{1/2}\left(\mathbb{E}e^2\right)^{1/2} < \infty.$$

It follows that the expectation $\mathbb{E}\left(\boldsymbol{x}e\right)$ is finite, and is zero by the calculation (2.28).

For part 6, Applying Minkowski's Inequality (B.19) to $e = y - \boldsymbol{x}'\boldsymbol{\beta}$,

$$\begin{aligned}
\left(\mathbb{E}\left|e\right|^r\right)^{1/r} &= \left(\mathbb{E}\left|y - \boldsymbol{x}'\boldsymbol{\beta}\right|^r\right)^{1/r} \\
&\leq \left(\mathbb{E}\left|y\right|^r\right)^{1/r} + \left(\mathbb{E}\left|\boldsymbol{x}'\boldsymbol{\beta}\right|^r\right)^{1/r} \\
&\leq \left(\mathbb{E}\left|y\right|^r\right)^{1/r} + \left(\mathbb{E}\left\|\boldsymbol{x}\right\|^r\right)^{1/r} \left\|\boldsymbol{\beta}\right\| \\
&< \infty,
\end{aligned}$$

the final inequality by assumption. ■

Exercises

Exercise 2.1 Find $\mathbb{E}\left(\mathbb{E}\left(\mathbb{E}\left(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\right) \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) \mid \boldsymbol{x}_1\right)$.

Exercise 2.2 If $\mathbb{E}\left(y \mid x\right) = a + bx$, find $\mathbb{E}\left(yx\right)$ as a function of moments of x.

Exercise 2.3 Prove Theorem 2.8.1.4 using the law of iterated expectations.

Exercise 2.4 Suppose that the random variables y and x only take the values 0 and 1, and have the following joint probability distribution

	$x=0$	$x=1$
$y=0$	.1	.2
$y=1$	.4	.3

Find $\mathbb{E}\left(y \mid x\right)$, $\mathbb{E}\left(y^2 \mid x\right)$ and $\operatorname{var}\left(y \mid x\right)$ for $x=0$ and $x=1$.

Exercise 2.5 Show that $\sigma^2(\boldsymbol{x})$ is the best predictor of e^2 given $\boldsymbol{x}$:

(a) Write down the mean-squared error of a predictor $h(\boldsymbol{x})$ for e^2.

(b) What does it mean to be predicting e^2?

(c) Show that $\sigma^2(\boldsymbol{x})$ minimizes the mean-squared error and is thus the best predictor.

Exercise 2.6 Use $y = m(\boldsymbol{x}) + e$ to show that

$$\operatorname{var}\left(y\right) = \operatorname{var}\left(m(\boldsymbol{x})\right) + \sigma^2$$

Exercise 2.7 Show that the conditional variance can be written as

$$\sigma^2(\boldsymbol{x}) = \mathbb{E}\left(y^2 \mid \boldsymbol{x}\right) - \left(\mathbb{E}\left(y \mid \boldsymbol{x}\right)\right)^2.$$

Exercise 2.8 Suppose that y is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of y given $\boldsymbol{x}$ is Poisson:

$$\Pr\left(y = j \mid \boldsymbol{x}\right) = \frac{\exp\left(-\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\beta}\right)^j}{j!}, \qquad j = 0, 1, 2, ...$$

Compute $\mathbb{E}\left(y \mid \boldsymbol{x}\right)$ and $\operatorname{var}\left(y \mid \boldsymbol{x}\right)$. Does this justify a linear regression model of the form $y = \boldsymbol{x}'\boldsymbol{\beta} + e$?

Hint: If $\Pr\left(y = j\right) = \frac{\exp(-\lambda)\lambda^j}{j!}$, then $\mathbb{E}y = \lambda$ and $\operatorname{var}(y) = \lambda$.

Exercise 2.9 Suppose you have two regressors: x_1 is binary (takes values 0 and 1) and x_2 is categorical with 3 categories (A, B, C). Write $\mathbb{E}\left(y \mid x_1, x_2\right)$ as a linear regression.

Exercise 2.10 True or False. If $y = x\beta + e$, $x \in \mathbb{R}$, and $\mathbb{E}\left(e \mid x\right) = 0$, then $\mathbb{E}\left(x^2 e\right) = 0$.

Exercise 2.11 True or False. If $y = x\beta + e$, $x \in \mathbb{R}$, and $\mathbb{E}\left(xe\right) = 0$, then $\mathbb{E}\left(x^2 e\right) = 0$.

Exercise 2.12 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ and $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$, then e is independent of $\boldsymbol{x}$.

Exercise 2.13 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ and $\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$, then $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$.

Exercise 2.14 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$, $\mathbb{E}(e \mid \boldsymbol{x}) = 0$, and $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2$, a constant, then e is independent of $\boldsymbol{x}$.

Exercise 2.15 Consider the intercept-only model $y = \alpha + e$ defined as the best linear predictor. Show that $\alpha = \mathbb{E}(y)$.

Exercise 2.16 Let x and y have the joint density $f(x, y) = \frac{3}{2}\left(x^2 + y^2\right)$ on $0 \leq x \leq 1$, $0 \leq y \leq 1$. Compute the coefficients of the best linear predictor $y = \alpha + \beta x + e$. Compute the conditional mean $m(x) = \mathbb{E}(y \mid x)$. Are the best linear predictor and conditional mean different?

Exercise 2.17 Let x be a random variable with $\mu = \mathbb{E}x$ and $\sigma^2 = \mathrm{var}(x)$. Define

$$g\left(x \mid \mu, \sigma^2\right) = \begin{pmatrix} x - \mu \\ (x-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Show that $\mathbb{E}g(x \mid m, s) = 0$ if and only if $m = \mu$ and $s = \sigma^2$.

Exercise 2.18 Suppose that

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x_2 \\ x_3 \end{pmatrix}$$

and $x_3 = \alpha_1 + \alpha_2 x_2$ is a linear function of x_2.

(a) Show that $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ is not invertible.

(b) Use a linear transformation of $\boldsymbol{x}$ to find an expression for the best linear predictor of y given $\boldsymbol{x}$. (Be explicit, do not just use the generalized inverse formula.)

Exercise 2.19 Show (2.46)-(2.47), namely that for

$$d(\boldsymbol{\beta}) = \mathbb{E}\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2$$

then

$$\begin{aligned} \boldsymbol{\beta} &= \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^k} d(\boldsymbol{b}) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} m(\boldsymbol{x})\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} y\right). \end{aligned}$$

Hint: To show $\mathbb{E}(\boldsymbol{x}m(\boldsymbol{x})) = \mathbb{E}(\boldsymbol{x}y)$ use the law of iterated expectations.

Exercise 2.20 Verify that (2.60) holds with $m(\boldsymbol{x})$ defined in (2.7) when $(y, \boldsymbol{x})$ have a joint density $f(y, \boldsymbol{x})$.