University College London

Department of Economics

M.Sc. in Economics

# MC3: Econometric Theory and Methods

## Course notes: 1

## Econometric models, random variables, probability distributions and regression

Andrew Chesher

5/10/2006

## 1. Introduction

These notes contain:

1. a discussion of the nature of economic data and the concept of an econometric model,
2. a review of some important concepts in probability distribution theory that arise frequently in developing econometric theory and in the application of econometric methods,
3. an introduction to the concept of a regression function in the context of distribution theory, as a preparation for the study of estimation of and inference concerning regression functions.

Since 2005 the Probability and Statistics Refresher Course has been wider in scope than in earlier years and has taken place during the first week of term so that all M.Sc. students could attend. That covered most of the material in item 2 above (Sections 4 - 7 of these notes), so I will not lecture on that this term. You should read all the sections of these notes carefully and study any elements that are new to you. Raise questions in class if you need to.

## 2. Data

Econometric work employs data recording economic phenomena and usually the environment in which they were obtained.

Sometimes we are interested in measuring simple economic magnitudes, for example the proportion of households in poverty in a region of a country, the degree of concentration of economic power in an industry, the amount by which a company's costs exceed the efficient level of costs for companies in its industry.

Often we are interested in the way in which the environment (broadly defined) affects interesting economic magnitudes, for example the impact of indirect taxes on amounts of goods purchased, the impact of direct taxes on labour supply, the sensitivity of travel choices to alternative transport mode prices and characteristics.

Challenging and important problems arise when we want to understand how people, households and institutions will react in the face of a policy intervention.

The data used in econometric work exhibit variation, often considerable amounts. This variation arises for a variety of reasons.

Data recording responses of individual agents exhibit variation because of differences in agents' preferences, differences in their environments, because chance occurrences affect different agents in different ways and because of measurement error.

Data recording time series of aggregate flows and stocks exhibit variation because they are aggregations of responses of individual agents whose responses vary for the reasons just described, because macroeconomic aggregates we observe may be developed from survey data (e.g. of households, companies) whose results are subject to sampling variation, because of changes in the underlying environment, because of chance events and shocks, and, because of measurement error.

## 3. Econometric models

Economics tells us about some of the properties of data generating processes. The knowledge that economics gives us about data generating processes is embodied in *econometric models*. Econometric models are constructions which set out the admissible properties of data generating processes. As an example consider an econometric model which might be used in the study of the *returns to schooling*.

Suppose we are interested in the determination of a labour market outcome, say the log wage $W$, and a measure (say years) of schooling ($S$) given a value of another specified characteristic ($X$) of an individual. Here is an example of the equations of a model for the process generating wage and schooling data given a value of $X$.

$$\begin{aligned} W &= \alpha_0 + \alpha_1 S + \alpha_2 X + \varepsilon_1 + \lambda\varepsilon_2 \\ S &= \beta_0 + \beta_1 X + \varepsilon_2 \end{aligned}$$

The term $\varepsilon_2$ is unobserved and allows individuals with identical values of $X$ to have different values of $S$, something likely to be seen in practice. We could think of $\varepsilon_2$ as a measure of "ability". This term also appears in the log wage equation, expressing the idea that higher ability people tend to receive higher[1] wages, other things being equal. The term $\varepsilon_1$ is also unobserved and allows people with identical values of $S$, $X$ and $\varepsilon_2$ to receive different wages, again, something likely to occur in practice.

In econometric models unobservable terms like $\varepsilon_1$ and $\varepsilon_2$ are specified as *random variables*, varying across individuals (in this example) with *probability distributions*. Typically an econometric model will place restrictions on these probability distributions. In this example a model could require $\varepsilon_1$ and $\varepsilon_2$ to have expected value zero and to be uncorrelated with $X$. We will shortly review the theory of random variables. For now we just note that if $\varepsilon_1$ and $\varepsilon_2$ are random variables then so are $W$ and $S$.

The terms $\alpha_0$, $\alpha_1$, $\alpha_2$, $\lambda$, $\beta_0$ and $\beta_1$ are unknown *parameters* of this model. A *particular* data generating process that conforms to this model will have equations as set out above with *particular* numerical values of the parameters and *particular* distributions for the unobservable $\varepsilon_1$ and $\varepsilon_2$. We will call such a fully specified data generating process a *structure*.

Each structure implies a particular probability distribution for $W$, $S$ and $X$ and statistical analysis can inform us about this distribution. Part of this course will be concerned with the way in which this sort of statistical analysis can be done.

In general *distinct* structures can imply the *same* probability distribution for the observable random variables. If, across such observationally equivalent structures an interesting parameter takes *different* values then no amount of data can be informative about the value of that parameter. We talk then of the parameter's value being *not identified*. If an econometric model is sufficiently restrictive then parameter values are identified. We will focus on identification issues, which lie at the core of econometrics, later in the course.

To see how observationally equivalent structures can arise, return to the wage-schooling model, and consider what happens when we substitute for $\varepsilon_2$ in the log wage equation using $\varepsilon_2 = S - \beta_0 - \beta_1 X$ which is implied by the schooling equation. After collecting terms we obtain

[1] If $\lambda > 0$.

the following.

$$\begin{aligned} W &= (\alpha_0 - \lambda\beta_0) + (\alpha_1 + \lambda)\, S + (\alpha_2 - \lambda\beta_1)\, X + \varepsilon_1 \\ S &= \beta_0 + \beta_1 X + \varepsilon_2 \end{aligned}$$

Write this as follows.

$$\begin{aligned} W &= \gamma_0 + \gamma_1 S + \gamma_2 X + \varepsilon_1 \\ S &= \beta_0 + \beta_1 X + \varepsilon_2 \end{aligned}$$

$$\begin{aligned} \gamma_0 &= \alpha_0 - \lambda\beta_0 \\ \gamma_1 &= \alpha_1 + \lambda \\ \gamma_2 &= \alpha_2 - \lambda\beta_1 \end{aligned}$$

Suppose we had data generated by a structure of this form. Given the values of $W$, $S$, $X$ and also the values of $\varepsilon_1$ and $\varepsilon_2$ we could deduce the values of $\gamma_0$, $\gamma_1$, $\gamma_2$, $\beta_0$ and $\beta_1$. We would need only three observations to do this if we were indeed give the values of $\varepsilon_1$ and $\varepsilon_2$.[2] But, given any values of $\beta_0$ and $\beta_1$, many values of the four unknowns $\alpha_0$, $\alpha_1$, $\alpha_2$ and $\lambda$ would be consistent with any particular values of $\gamma_0$, $\gamma_1$, and $\gamma_2$. These various values are associated with observationally equivalent structures. We could not tell which particular set of values of $\alpha_0$, $\alpha_1$, $\alpha_2$ and $\lambda$ generated the data. This is true even if we are given the values of $\varepsilon_1$ and $\varepsilon_2$. This situation get no better once we do not have this information as will always be the case in practice.

Note that only one value of $\beta_0$ and $\beta_1$ could generate a particular set of values of $S$ given a particular set of values of $X$ and $\varepsilon_2$.[3] It appears that $\beta_0$ and $\beta_1$ are identified. However if, say, large values of $\varepsilon_2$ tend to be associated with large values of $X$, then data cannot distinguish the impact of $X$ and $\varepsilon_2$ on $S$, so models need to contain some restriction on the co-variation of unobservables and other variables if they are to have identifying power.

Note also that if economic theory required that $\alpha_2 = 0$ and $\beta_1 \neq 0$ then there would be only one set of values of $\alpha_0$, $\alpha_1$ and $\lambda$ which could produce a given set of values of $W$ and $S$ given a particular set of values of $X$, $\varepsilon_1$ and $\varepsilon_2$.[4] Again data cannot be informative about those values unless data generating structures conform to a model in which there is some restriction on the co-variation of unobservables and other variables. These issues will arise again later in the course.

The model considered above is highly restrictive and embodies functional form restrictions which may not flow from economic theory. A less restrictive model has equations of the following form

$$\begin{aligned} W &= h_1(S, X, \varepsilon_1, \varepsilon_2) \\ S &= h_2(X, \varepsilon_2) \end{aligned}$$

where the functions $h_1$ and $h_2$ are left unspecified. This is an example of a *nonparametric* model. Note that structures which conform to this model have returns to schooling $(\partial h_1/\partial S)$

[2] Ruling out cases in which there was a linear dependence between the values of $S$ and $X$.

[3] Unless the $X$ data take special sets of values, for example each of the 100 values of $X$ is identical.

[4] Again unless special sets of values of $X$ arise.

which may depend upon $S$, $X$ and the values of the unobservables. In structures conforming to the linear model the returns to schooling is the constant $\alpha_1$.

The linear model we considered above is in fact, as we specified it, *semiparametric*, in the sense that, although the equations were written in terms of a finite number of unknown parameters, the distributions of $\varepsilon_1$ and $\varepsilon_2$ were not parametrically specified. If we further restricted the linear model, requiring $\varepsilon_1$ and $\varepsilon_2$ to have, say, normal distributions then we would have a fully *parametric* model.

In practice the "true" data generating process (structure) may not satisfy the restrictions of an econometric model. In this case we talk of the model as being *misspecified*. Part of our effort will be devoted to studying ways of detecting misspecification.

Since in econometric analysis we regard data as realisations of *random variables* it is essential to have a good understanding of the theory of random variables, and so some important elements of this are reviewed now. We first consider a single (scalar) random variable and then some extensions needed when many random variables are considered simultaneously, as is often the case.

## 4. Scalar random variables

A scalar random variable, $X$, takes values on the real line, $\Re^1$, such that for all $x \in \Re^1$, the probability $P[X \leq x]$ is defined. A proper random variable (we shall always deal with these) has

$$P[-\infty < X < \infty] = 1. \tag{4.1}$$

The function $F_X(x) = P[X \leq x]$ is called the *distribution function*[5]. This function is non-decreasing. The set of values at which the distribution function is increasing is called the support (of the distribution) of $X$. The probability that $X$ falls in an interval $(a, b]$ is[6]

$$P[X \in (a, b]] = F_X(b) - F_X(a).$$

If the support of $X$ contains a finite or countably infinite number of elements then $X$ has a *discrete distribution*. In econometric work we frequently encounter discrete *binary random variables* whose two points of support (zero and one) indicate the possession or not of an attribute, or occurrence or not of an event (e.g. having a job, owning an asset, buying a commodity). We also encounter discrete random variables with many points of support, e.g. in studying the returns to R&D investment when we might look at data on the number of patents a company registers in a year.

$X$ is continuously distributed over intervals for which the derivative

$$f_X(x) = \frac{\partial}{\partial x} F_X(x)$$

exists[7]. $X$ is a *continuous random variable* if it is continuously distributed over its support.

[5] Sometimes the cumulative distribution function.

[6] The square bracket, "]" indicates that the value "$b$" is contained in the interval whose probability of occurrence is being considered, "(" indicates that the value "$a$" is not.

[7] We require the left and right derivatives to be finite and equal, i.e.

$$\lim_{\Delta x \to 0_-} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = \lim_{\Delta x \to 0_+} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = f_X(x)$$

We often use continuous random variables as models for econometric data such as income, and the times between events (e.g. unemployment durations), even though in reality our data are recorded to finite accuracy. When data are coarsely grouped, as income responses in household surveys sometimes are, we employ discrete data models but often these are derived from an underlying model for a continuous, but unobserved, response. We do encounter random variables which are continuously distributed over only a part of their support. For example expenditures recorded over a period of time are often modelled as continuously distributed over positive values with a point mass of probability at zero.

For continuous random variables the function $f_X(x)$, defined over all the support of $X$, is called the *probability density function*. The probability that continuously distributed $X$ falls in intervals $[a, b]$, $(a, b]$, $[a, b)$ and $(a, b)$ is

$$F_X(b) - F_X(a) = \int_a^b dF_X(x) = \int_a^b \frac{d}{dx} F_X(x) dx = \int_a^b f_X(x) dx.$$

Because of (4.1)

$$\int_{-\infty}^{\infty} f_X(x) dx = 1,$$

that is, the probability density function integrates to one over the support of the random variable.

Purely discrete random variables have support on a set of points $\mathcal{X} = \{x_i\}_{i=1}^{M_X}$ where the number of points of support, $M_X$, may be infinite and $x_1 < x_2 < \cdots < x_m < \ldots$. Often these points are equally spaced on the real line in which case we say that $X$ has a *lattice distribution*. The probability mass on the $i$th point of support is $p_i = p(x_i) = F_X(x_i) - F_X(x_{i-1})$ where we define $x_0 = -\infty$, and $\sum_{i=1}^{M_X} p_i = 1$. If $\mathcal{A} \subset \mathcal{X}$ is a subset of the points of support then $P[X \in \mathcal{A}] = \sum_{x_i \in \mathcal{A}} p(x_i)$.

Example 1. *The exponential distribution.*

> Let $X$ be a continuously distributed random variable with support on $[0, \infty)$ with distribution function $F_X(x) = 1 - \exp(-\lambda x)$, $x \geq 0$, $F_X(x) = 0$, $x < 0$, where $\lambda > 0$. Note that $F_X(-\infty) = F_X(0) = 0$, $F_X(\infty) = 1$, and $F_X(\cdot)$ is strictly increasing over its support. The probability density function of $X$ is $f_X(x) = \lambda \exp(-\lambda x)$. Sketch this density function and the distribution function. This distribution is often used as a starting point for building econometric models of durations, e.g. of unemployment.

### 4.1. Functions of a random variable

Let $g(\cdot)$ be an increasing function and define the random variable $Z = g(X)$. Then, with $g^{-1}(x)$ denoting the inverse function satisfying

$$a = g(g^{-1}(a))$$

we have

$$F_Z(z) = P[Z \leq z] = P[g(X) \leq z] = P[X \leq g^{-1}(z)] = F_X(g^{-1}(z)). \tag{4.2}$$

with $f_X(x)$ finite.

The point here is that $\{Z \leq z\}$ is an event that occurs if and only if the event $\{g(X) \leq z\}$ occurs, and this event occurs if and only if the event $\left\{X \leq g^{-1}(z)\right\}$ occurs - and identical events must have the same probability of occurrence.

In summary

$$F_Z(z) = F_X(g^{-1}(z)).$$

Put another way[8],

$$F_X(x) = F_Z(g(x)).$$

For continuous random variables and differentiable functions $g(\cdot)$, we have, on differentiating with respect to $x$, and using the "chain rule"

$$f_X(x) = f_Z(g(x)) \times g'(x)$$

and using $z = g(x)$, $x = g^{-1}(z)$,

$$f_Z(z) = f_X(g^{-1}(z))/g'(g^{-1}(z)).$$

Here $'$ denotes the first derivative.

If $g(\cdot)$ is a *decreasing* function and $X$ is a continuous random variable then (4.2) is replaced by

$$F_Z(z) = P[Z \leq z] = P[g(X) \leq z] = P[X \geq g^{-1}(z)] = 1 - F_X(g^{-1}(z)).$$

Notice that, because $g(\cdot)$ is a decreasing function, the inequality is reversed when the inverse function, $g^{-1}(\cdot)$ is applied. Drawing a picture helps make this clear.

In summary,

$$F_X(x) = 1 - F_Z(g(x)).$$

For continuous random variables and differentiable $g(\cdot)$

$$\begin{aligned} f_X(x) &= -f_Z(g(x)) \times g'(x) \\ f_Z(z) &= -f_X(g^{-1}(z))/g'(g^{-1}(z)). \end{aligned}$$

The results for probability density functions for increasing and decreasing functions $g(\cdot)$ are combined in

$$\begin{aligned} f_X(x) &= f_Z(g(x)) \times |g'(x)| \\ f_Z(z) &= f_X(g^{-1}(z))/\left|g'(g^{-1}(z))\right|. \end{aligned}$$

If the function $g(\cdot)$ is not monotonic the increasing and decreasing segments must be treated separately and the results added together.

Example 2. *The normal (Gaussian) and log normal distributions.*

[8] Substitute $z = g(x)$ and use $g^{-1}(g(x)) = x$.

A normally distributed random variable $X$ has probability density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \qquad x \in (-\infty, \infty).$$

This density function is symmetric about $x = \mu$ with fast decreasing tails, and bell shaped. The smaller is $\sigma$ the faster the tails fall away and the more concentrated is the distribution around $\mu$. The normal distribution function cannot be expressed in terms of simple functions but most statistical software has a built in function which computes it.

A common model used in the study of income distributions supposes that log income has a normal distribution. In this case we say that income is log normally distributed. Suppose log income ($X$) has the normal density function above. What is the density function of income, that is of $Z = \exp(X)$?

First note that $Z$ has support on $(0, \infty)$. Applying the result above with

$$g(X) = \exp(X) = g'(X)$$

noting the $\exp(X)$ is an increasing function,

$$g^{-1}(Z) = \log(Z)$$

$$g'(g^{-1}(z)) = \exp(\log(z)) = z$$

gives

$$f_Z(z) = \frac{1}{z\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(z)-\mu}{\sigma}\right)^2\right).$$

This is a skewed distribution with a relatively long tail in the positive direction.

### 4.2. (*) Application: Simulation[9]

It is often useful to be able to generate realisations of random variables with specific distributions. We sometimes do this in order to study the properties of some statistical procedure, sometimes in order to get understanding of the implications of an econometric model. In most statistical software packages there is a facility for generating sequences of numbers which mimic realisations from a standard uniform distribution. Here we show how these can be transformed so that they mimic realisations from a distribution of our choice.

A standard *uniform random variable* takes all values on the unit interval and the probability that a value falls in any interval is proportional to the length of the interval. For a standard uniform random variable, $U$, the distribution and density functions are

$$F_U(u) = u, \qquad f_U(u) = 1, \qquad u \in [0, 1].$$

[9] Some may find the starred Sections more demanding. They can be omitted.

Suppose we want pseudo-random numbers mimicing realisations of a random variable $W$ which has distribution function $F_W(w)$ and let the inverse distribution function (we will sometimes call this the *quantile function*) be $Q_W(p)$ for $p \in [0,1]$, i.e.

$$Q_W(p) = F_W^{-1}(p), \qquad p \in [0,1]$$

equivalently

$$F_W(Q_W(p)) = p.$$

Let $U$ have a standard uniform distribution and let $V = Q_W(U)$. Then, using the results above (work through these steps) the distribution function of $V$ is

$$F_V(v) = P[V \leq v] = P[Q_W(U) \leq v] = P[U \leq Q_W^{-1}(v)] = F_U(Q_W^{-1}(v)) = Q_W^{-1}(v) = F_W(v)$$

So, the distribution function of $V$ is identical to the distribution function of $W$. To generate pseudo-random numbers mimicing a random variable with distribution function $F_W$ we generate standard uniform pseudo-random numbers, $u$, and use $Q_W(u)$ as our pseudo-random numbers mimicing values drawn from the distribution of $W$.

### 4.3. Quantiles

The values taken by the quantile function are known as the quantiles of the distribution of $X$. Some quantiles have special names. For example $Q_X(0.5)$ is called the *median* of $X$, $Q_X(p)$ for $p \in \{0.25, 0.5, 0.75\}$ are called the *quartiles* and $Q_X(p)$, $p \in \{0.1, , 0, 2, \ldots, 0.9\}$ are called the *deciles*. The median is often used as a measure of the location of a distribution and the *interquartile range*, $Q_X(0.75) - Q_X(0.25)$ is sometimes used as a measure of dispersion.

### 4.4. Expected values and moments

Let $Z = g(X)$ be a function of $X$. The expected value of $Z$ is defined for continuous and discrete random variables respectively as

$$E_Z[Z] = E_X[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{4.3}$$

$$E_Z[Z] = E_X[g(X)] = \sum_{i=1}^{M_X} g(x_i) p(x_i) \tag{4.4}$$

which certainly exists when $g(\cdot)$ is bounded, but may not exist for unbounded functions.

Expected values correspond to the familiar notion of an average. They are one measure of the location of the probability distribution of a random variable ($g(X) = Z$ above). They also turn up in decision theory as, under some circumstances[10], an optimal prediction of the value that a random variable will take.

[10] When the loss associated with predicting $y_p$ when $y$ is realised is quadratic: $L(y_p, y) = a + b(y - y_p)^2$, $b > 0$, and we choose a prediction that minimises expected loss.

The expected value of a constant is the value of the constant because, e.g. for continuous random variables (work through these steps for a discrete random variable) and the constant $a$:

$$E_X[a] = \int_{-\infty}^{\infty} a f_X(x)dx = a\int_{-\infty}^{\infty} f_X(x)dx = a \times 1 = a.$$

The expected value of a constant times a random variable is the value of the constant times the expected value of the random variable, because, again for continuous random variables and a constant $b$,

$$E_X[bX] = \int_{-\infty}^{\infty} bx f_X(x)dx = b\int_{-\infty}^{\infty} x f_X(x)dx = bE_X[X].$$

Therefore

$$E_X[a + bX] = a + bE_X[X].$$

For additively separable $g(X) = g_1(X) + g_2(X)$ we have

$$E_X[g(X)] = E_X[g_1(X)] + E_X[g_2(X)].$$

Show that this is true using the definitions (4.3) and (4.4).

The expected values $E_X[X^j]$ for positive integer $j$ are the *moments of order $j$ about zero* of $X$ and $E_X[(X - E[X])^j]$ are the *central moments* and in particular $Var_X(X) = E[(X - E[X])^2]$ is the *variance*. Note that if $X$ does not have bounded support then these moments are expectations of unbounded functions and so in some cases may not exist.

It is sometimes helpful to think of the probability that $X$ lies in some region $A \subset \Re^1$ as being itself an expectation. To do this define the indicator function

$$\begin{aligned} 1_{[x\in A]} &= 1, \qquad x \in A \\ &= 0, \qquad x \notin A. \end{aligned}$$

Then

$$P[X \in A] = E_X[1_{[X\in A]}].$$

### 4.5. Moment generating functions

Expected values, and moments generally, attract a lot of interest in econometric work and it is useful to have a variety of methods for calculating them. Moment generating functions are sometimes helpful in this respect. The moment generating function of a random variable $X$ is defined, when it exists, as the expectation of the function $\exp(tX)$ where $t$ is a constant.

$$M_X(t) = E_X[\exp(tX)]. \tag{4.5}$$

This function clearly exists for all random variables with bounded support, discrete or continuous. It exists for some, but not all, random variables with support on the whole real line.

Considering the definition of the expectation of a function of a random variable we see that, if $M_X^{(i)}(0)$ denotes the $i$th derivative of $M_X(t)$ evaluated at $t = 0$ then $M_X^{(i)}(0) = E_X[X^i]$, that is, is equal to the $i$th moment about zero. For continuous random variables for example

$$M_X^{(1)}(t) = \frac{\partial}{\partial t}\int_{-\infty}^{\infty} \exp(tx) f_X(x)dx = \int_{-\infty}^{\infty} x\exp(tx) f_X(x)dx$$

and setting $t = 0$, and using $\exp(0) = 1$,

$$M_X^{(1)}(0) = \int_{-\infty}^{\infty} x f_X(x) dx = E_X[X].$$

Similarly

$$M_X^{(2)}(t) = \int_{-\infty}^{\infty} x^2 \exp(tx) f_X(x) dx$$

and setting $t = 0$

$$M_X^{(2)}(0) = \int_{-\infty}^{\infty} x^2 f_X(x) dx = E_X[X^2].$$

Work through these steps for a discrete random variable.

Example 2 (continued). *The normal moment generating function.*

The normal distribution's moment generating function is

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(tx) \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(x^2 - 2\left(\mu + t\sigma^2\right)x + \mu^2\right)\right) dx \\ &= \exp(\mu t + \frac{t^2\sigma^2}{2}) \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(x - \left(\mu + t\sigma^2\right)\right)^2\right) dx \\ &= \exp(\mu t + \frac{t^2\sigma^2}{2}) \end{aligned}$$

where the last line follows on noting that the normal density function integrates to one whatever its mean.

Differentiating with respect to $t$ and setting $t = 0$ after each differentiation gives the moments about zero of this normal random variable, $E_X[X] = \mu$, $E_X[X^2] = \mu^2 + \sigma^2$, whence $Var(X) = \sigma^2$. The "standard" normal distribution (with mean zero and variance one) has moment generating function equal to $\exp(t^2/2)$.

Example 3. *The Poisson distribution.*

As another example consider a Poisson random variable which is discrete with support on the non-negative integers and probability mass function

$$P[X = x] = \frac{\lambda^x \exp(-\lambda)}{x!}, \qquad x \in \{0, 1, 2, \dots\}$$

where $\lambda > 0$. Note that because

$$\exp(\lambda) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \tag{4.6}$$

this is a proper probability mass function. This distribution is often used as a starting point for modelling data which record counts of events.

The moment generating function of this Poisson random variable is

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} \exp(tx) \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x \exp(-\lambda)}{x!} \\ &= \exp(\lambda e^t - \lambda) \end{aligned} \tag{4.7}$$

where to get to the last line we have used (4.6) with $\lambda$ replaced by $\lambda e^t$.

The first two moments of this Poisson random variable are then easily got by differentiating the moment generating function with respect to $t$ and setting $t = 0$, $E_X[X] = \lambda$, $E_X[X^2] = \lambda^2 + \lambda$, from which we see that $Var[X] = E[X^2] - E[X]^2 = \lambda$. So a Poisson random variable has variance equal to its mean. One way to tell if a Poisson distribution is a suitable model for data which are counts of events is to see if the difference between the sample mean and sample variance is too large to be the result of chance sampling variation.

## 4.6. (*) Using moment generating functions to determine limiting behaviour of distributions

Consider a "standardised" Poisson random variable constructed to have mean zero and variance one, namely:

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}.$$

We will investigate the behaviour of the moment generating function of this random variable as $\lambda$ becomes large.

What is the moment generating function of $Z$? Applying the definition (4.5)

$$\begin{aligned} M_Z(t) &= E_Z[\exp(tZ)] \\ &= E_X[\exp(t\left(\frac{X - \lambda}{\sqrt{\lambda}}\right))] \\ &= E_X[\exp(\frac{t}{\sqrt{\lambda}} X)] \exp(-t\sqrt{\lambda}) \\ &= \exp(\lambda e^{t/\sqrt{\lambda}} - \lambda - t\sqrt{\lambda}) \end{aligned}$$

where the last line follows on substituting $t/\sqrt{\lambda}$ for $t$ in (4.7).

When $\lambda$ is large enough, $t/\sqrt{\lambda}$ is small for any positive $t$, and $e^{t/\sqrt{\lambda}} \simeq 1 + t/\sqrt{\lambda} + t^2/(2\lambda)$. Substituting in the last line above gives, for large $\lambda$, $M_Z(t) \simeq \exp(t^2/2)$ which is the moment

generating function of a standard (zero mean, unit variance) *normally distributed* random variable. This informal argument suggests that a Poisson random variable with a large mean is approximately distributed as a normal random variable.

In fact this is the case. However a formal demonstration would (a) have to be more careful about the limiting operation and (b) be conducted in terms of the *characteristic function* which is defined as $C_X(t) = E_X[\exp(itX)]$ where $i^2 = -1$. This generally complex valued function of $t$ always exists because $\exp(itX)$ is a bounded function[11] of $X$. Further, under very general conditions there is a one to one correspondence between characteristic functions and the distributions of random variables. That means that if we can show (as here we can) that the characteristic functions of a sequence of random variables converge to the characteristic function of a random variable, $Y$, say, then the distributions of the sequence of random variables converge to the distribution of $Y$.

## 5. Many random variables

In econometric work we usually deal with data recording many aspects of the economic phenomenon of interest. For example in a study of consumers' expenditures we will have records for each household of expenditures on many goods and services, perhaps for more than one period of time, and also data recording aspects of the households' environments (income, household composition etc.). And in macroeconometric work we will often observe many simultaneously evolving time series. We model each recorded item as a realisation of a random variable and so we have to be able to manipulate many random variables simultaneously. This requires us to extend some of the ideas above and to introduce some new ones.

For the moment consider two random variables, $X$ and $Y$. The extension of most of what we do now to more than two random variables is, for the most part obvious, and will be summarised later.

The *joint distribution function* of $X$ and $Y$ is[12]

$$P[X \leq x \cap Y \leq y] = F_{XY}(x, y), \qquad (x, y) \in \Re^2$$

which is a non decreasing function of both its arguments, with $F_{XY}(-\infty, -\infty) = 0$, $F_{XY}(\infty, \infty) = 1$. The support of a pair of random variables is a set of points in the real two dimensional plane, $\Re^2$. The distribution function of, say, $X$ alone is extracted from this by noting that

$$P[X \leq x] = P[X \leq x \cap Y \leq \infty] = F_{XY}(x, \infty) = F_X(x).$$

We call this a *marginal distribution function.*

---

[11]
$$\exp(itX) = \cos(tX) + i\sin(tX)$$
and
$$|\cos(tX) + i\sin(tX)| = \cos^2(tX) + \sin^2(tX) = 1.$$

[12] $A \cap B$ is the event which occurs if and only if the events $A$ and $B$ both occur.

The probability that $X$ and $Y$ lie respectively in intervals $(x_L, x_U)$, $(y_L, y_U)$ is

$$\begin{aligned} P[(x_L < X \leq x_U) \cap (y_L < Y \leq y_U)] &= F_{XY}(x_U, y_U) \\ &\quad -F_{XY}(x_L, y_U) - F_{XY}(x_U, y_L) \\ &\quad +F_{XY}(x_L, y_L). \end{aligned} \tag{5.1}$$

To understand this it helps to draw a picture showing the support of $X$ and $Y$.

Now suppose $X$ and $Y$ are jointly *continuously* distributed and above let

$$x_U = x_L + \Delta x, \qquad y_U = y_L + \Delta y$$

where $\Delta x$ and $\Delta y$ are vanishingly small.

Then

$$\begin{aligned} P[(x_L < X \leq x_L + \Delta x) \cap (y_L < Y \leq y_L + \Delta y)] &= F_{XY}(x_L + \Delta x, y_L + \Delta y) \\ &\quad -F_{XY}(x_L, y_L + \Delta y) \\ &\quad -F_{XY}(x_L + \Delta x, y_L) \\ &\quad +F_{XY}(x_L, y_L) \end{aligned}$$

and[13]

$$\begin{aligned} \lim_{\Delta x \to 0, \Delta y \to 0} \frac{1}{\Delta x \Delta y} P[(x_L < X \leq x_L + \Delta x) \cap (y_L < Y \leq y_L + \Delta y)] &= \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) \\ &= f_{XY}(x, y) \end{aligned}$$

We call $f_{XY}(x, y)$ the *joint probability density function* of $X$ and $Y$.

It follows that the probability that $X$ and $Y$ lie respectively in intervals $(x_L, x_U)$, $(y_L, y_U)$ can be written as

$$P[(x_L < X \leq x_U) \cap (y_L < Y \leq y_U)] = \int_{y_L}^{y_U} \int_{x_L}^{x_U} f_{XY}(x, y) dx dy$$

and generally, for a subset of the real plane, $A \subset \Re^2$,

$$P[(X, Y) \in A] = \int\int_{(x,y) \in A} f_{XY}(x, y) dx dy.$$

For pairs of discrete random variables we define the joint probability (mass) function

$$P[X = x_i \cap Y = y_j] = p_{XY}(x_i, y_j)$$

which can be obtained from the joint distribution function using (5.1) and the associated marginal probability (mass) functions:

$$\begin{aligned} P[X &= x_i] = \sum_{j=1}^{M_Y} p_{XY}(x_i, y_j) = p_X(x_i) \\ P[Y &= y_j] = \sum_{i=1}^{M_X} p_{XY}(x_i, y_j) = p_Y(y_j). \end{aligned}$$

[13] Check that indeed what we have here is the definition of the second cross partial derivative.

Thinking of the joint probabilities being arrayed in a table these last two operations involve adding up entries across rows or columns of the table to produce totals to appear in the margins of the table, hence the expression, "marginal distribution".

### 5.1. Expected values, variance and covariance

Let $g(\cdot,\cdot)$ be a scalar function of two arguments. The *expected value* of $Z = g(X,Y)$ is defined for continuous and discrete random variables respectively as

$$\begin{aligned} E_Z[Z] &= E_{XY}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)dxdy \\ E_Z[Z] &= E_{XY}[g(X,Y)] = \sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} g(x_i,y_j) p_{XY}(x_i,y_j) \end{aligned}$$

where for discrete random variables with $X \in \{x_i\}_{i=1}^{M_X}$, $Y \in \{y_i\}_{i=1}^{M_Y}$

$$P[X = x_i \cap Y = y_j] = p_{XY}(x_i,y_j).$$

For additively separable functions,

$$E_{XY}[g_1(X,Y) + g_2(X,Y)] = E_{XY}[g_1(X,Y)] + E_{XY}[g_2(X,Y)].$$

Check this using the definitions above. Also note[14] that for functions of one random variable alone, say $Y$,

$$E_{XY}[g(Y)] = E_Y[g(Y)]$$

which is determined entirely by the marginal distribution of $Y$.

Once we deal with multiple random variables there are some functions of interest which require consideration of the joint distribution. Of particular interest are the cross central moments, $E[(X - E[X])^i(Y - E[Y])^j]$ which may of course not exist for all $i$ and $j$.

The variances of $X$ and $Y$, when they exist, are obtained when we set $i = 2$, $j = 0$ and $i = 0$, $j = 2$, respectively. Setting $i = 1$, $j = 1$ gives the *covariance* of $X$ and $Y$

$$Cov(X,Y) = E_{XY}[(X - E[X])(Y - E[Y])].$$

The *correlation* between $X$ and $Y$ is defined as

$$Cor(X,Y) = \frac{Cov(X,Y)}{(Var(X)Var(Y))^{1/2}}.$$

This quantity, when it exists, always lies in $[-1,1]$.

[14] For continuous random variables,

$$\begin{aligned} E_{XY}[g(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y) f_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty} g(y)\left(\int_{-\infty}^{\infty} f_{XY}(x,y)dx\right)dy \\ &= \int_{-\infty}^{\infty} g(y) f_Y(y)dy = E_Y[g(Y)] \end{aligned}$$

### 5.2. Conditional probabilities

Conditional distributions are of crucial importance in econometric work. They tell us how the probabilities of events concerning one set of random variables depend (or not) on values taken by other random variables.

For events $A$ and $B$ the *conditional probability* that event $A$ occurs given that event $B$ occurs is

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \tag{5.2}$$

We require that $B$ occurs with non-zero probability. Then we can write

$$\begin{aligned} P[A \cap B] &= P[A|B] \times P[B] \\ &= P[B|A] \times P[A] \end{aligned}$$

the second line following on interchanging the roles of $A$ and $B$, from which

$$\begin{aligned} P[A|B] &= \frac{P[B|A]P[A]}{P[B]} = \frac{P[B|A]P[A]}{P[B \cap A] + P[B \cap \bar{A}]} \\ &= \frac{P[B|A]P[A]}{P[B|A]P[A] + P[B|\bar{A}]P[\bar{A}]} \end{aligned}$$

where $\bar{A}$ is the event occurring when the event $A$ does not occur. This is known as Bayes Theorem.

For three events,

$$P[A \cap B \cap C] = P[A|B \cap C] \times P[B|C] \times P[C]$$

and so on. This sort of iteration is particularly important when we deal with time series, or the results of sequential decisions, in which $A$, $B$, $C$, and so on, are a sequence of events ordered in time with $C$ preceding $B$ preceding $A$ and so forth.

### 5.3. Conditional distributions

Let $X$ and $Y$ be discrete random variables. Then the conditional probability mass function of $Y$ given $X$ is, applying (5.2)

$$p_{Y|X}(y_j|x_i) = P[Y = y_j|X = x_i] = p_{XY}(x_i, y_j)/p_X(x_i).$$

For continuous random variables a direct application of (5.2) is problematic because $P[X \in (x, x + \Delta x)]$ approaches zero as $\Delta x$ approaches zero. However it is certainly possible to define the conditional distribution function of $Y$ given $X \in (x, x + \Delta x)$ for any non-zero value of $\Delta x$ directly from (5.2) as

$$\begin{aligned} P[Y \leq y|X \in (x, x + \Delta x)] &= \frac{F_{XY}(x + \Delta x, y) - F_{XY}(x, y)}{F_X(x + \Delta x) - F_X(x)} \\ &= \frac{\left(F_{XY}(x + \Delta x, y) - F_{XY}(x, y)\right)/\Delta x}{\left(F_X(x + \Delta x) - F_X(x)\right)/\Delta x} \end{aligned}$$

and letting $\Delta x$ pass to zero, gives what we will use as the definition of the *conditional distribution function* of $Y$ given $X = x$:

$$P[Y \leq y|X = x] = \frac{\partial}{\partial x}F_{XY}(x, y)/\frac{\partial}{\partial x}F_X(x) = \frac{1}{f_X(x)}\frac{\partial}{\partial x}F_{XY}(x, y)$$

from which, on differentiating with respect to $y$, we obtain the *conditional probability density function* of $Y$ given $X$ as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}. \tag{5.3}$$

Note that this is a proper probability density function wherever $f_X(x) \neq 0$ in the sense that $f_{Y|X}(y|x) \geq 0$ and

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} f_{XY}(x, y)dy/f_X(x) = f_X(x)/f_X(x) = 1.$$

Turning (5.3) around,

$$\begin{aligned} f_{XY}(x, y) &= f_{Y|X}(y|x)f_X(x) \\ &= f_{X|Y}(x|y)f_Y(y) \end{aligned}$$

where the second line follows on interchanging the roles of $x$ and $y$. It follows that

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int f_{X|Y}(x|y)f_Y(y)dy} = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

which is the equivalent for density functions of Bayes Theorem given above. This simple expression lies at the heart of a complete school of inference - Bayesian inference.

### 5.4. Independence

Two random variables are said to be *independently distributed* if for all sets $A$ and $B$,

$$P[X \in A \cap Y \in B] = P[X \in A]P[Y \in B].$$

Consider two random variables $X$ and $Y$ such that the support of the conditional distribution of $X$ given $Y$ is independent of $Y$ and vice versa. Then the random variables are independent if the joint distribution function of $X$ and $Y$ is the product of the marginal distribution functions of $X$ and $Y$ for all values of their arguments. For jointly continuously distributed random variables this implies that the joint density is the product of the two marginal densities and that the conditional distributions are equal to their marginal distributions.

We use the idea of independence extensively in econometric work. For example when analysing data from household surveys it is common to proceed on the basis that data from different households at a common point in time are realisations of independent random variables, at least conditional on a set of household characteristics. That would be a reasonable basis for analysis under some survey sampling schemes.

### 5.5. Regression

Consider a function of $Y$, $g(Y)$. The conditional expectation of $g(Y)$ given $X = x$ is defined for continuous random variables as

$$E_{Y|X}[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) dy$$

and for discrete random variables as

$$E_{Y|X}[g(Y)|X = x_i] = \sum_{j=1}^{M_Y} g(y_j) p_{Y|X}(y_j|x_i).$$

These functions are given the generic name *regression functions*.

When $g(Y) = Y$ we have the *mean regression function* which describes how the conditional expectation of $Y$ given $X = x$ varies with $x$. This is often referred to as *the* regression function. $Var[Y|X = x]$ is less commonly referred to as the *scedastic function*.

In econometric work we are often interested in the forms of these functions. We will shortly consider how regression functions can be estimated using realisations of random variables and study the properties of alternative estimators. Much of the interest in current econometric work is in the mean regression function but scedastic functions are also of interest.

For example in studying the returns to schooling we might think of the wage rate a person obtains after completing education as a random variable with a conditional distribution given $X$, years of schooling. The mean regression tells us how the average wage rate varies with years of schooling - we might be interested in the linearity or otherwise of this regression function and in the magnitude of the derivative of the regression function with respect to years of schooling.

The scedastic function tells us how the dispersion of wage rates varies with years of schooling. If we are interested in wage inequality then this is an interesting function in its own right. As we will see the form of the scedastic function is also important when we come to consider the properties of estimators of the (mean) regression function.

We can think of the conditional distribution function as a regression function. Define $Z(Y, c) = 1_{[Y<c]}$. Then

$$E_{Y|X}[Z(Y, c)|X = x] = \int_{-\infty}^{\infty} 1_{[Y<c]} f_{Y|X}(y|x) dy = \int_{-\infty}^{c} f_{Y|X}(y|x) dy = F_{Y|X}(c|x).$$

Sometimes we are interested in the way in which the conditional distribution of some random variable (e.g. wages) varies with conditioning variables and then we might consider the conditional quantile functions. Let $Q_{Y|X}(p, x)$ be such that $F_{Y|X}(Q_{Y|X}(p, x)|x) = p$. For $p = 0.5$ this is called the median regression function and generally a *quantile regression function*.

The $p$-quantile regression function satisfies

$$p = \int_{-\infty}^{Q_{Y|X}(p,x)} f_{Y|X}(y|x) dy.$$

It is possible to estimate quantile regression functions. If we do this and find that for different values of $p$ they are not parallel then we have evidence that the dispersion and/or shape of the conditional distribution of $Y$ given $X$ depends upon the value of $X$.

## 6. Iterated expectations

We will frequently make use of the following important result, known as the law of iterated expectations.

$$E_Y[Y] = E_X[E_{Y|X}[Y|X]]$$

Loosely speaking this says that to obtain the expectation of $Y$ we can average the expected value of $Y$ obtained at each possible value of $X$, weighting these conditional expectations (of $Y$ given $X = x$) by the probability that $X = x$ occurs. Formally, for continuous random variables we have

$$\begin{aligned}
E_X[E_{Y|X}[Y|X]] &= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x,y) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X|Y}(x|y) f_Y(y) dy dx \\
&\quad \int_{-\infty}^{\infty} y \left( \int_{-\infty}^{\infty} f_{X|Y}(x|y) dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= E_Y[Y].
\end{aligned}$$

Work carefully through the steps in this argument. Repeat the steps for a pair of discrete random variables[15].

## 7. Many random variables

Here the extension of the previous results to many random variables is sketched for the case in which the random variables are jointly continuously distributed.

Let the $N-$element vector

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$$

denote $N$ random variables with joint distribution function

$$P[X \leq x] = P[\bigcap_{i=1}^{N} (X_i \leq x_i)] = F_X(x)$$

[15] More difficult - how would you prove the result if the support of $Y$ depended upon $X$, so that given $X = x$, $Y \in (-\infty, h(x))$ where $h(x)$ is an increasing function of $x$, and $h(\infty) = \infty$?

where $x = (x_1, \ldots, x_N)'$. Here and later in this section $'$ denotes *transposition* not differentiation.

The joint density function of $X$ is

$$f_X(x) = \frac{\partial^N}{\partial x_1 \ldots \partial x_N} F_X(x).$$

The expected value of $X$ is

$$E_X[X] = \begin{bmatrix} E_{X_1}[X_1] \\ \vdots \\ E_{X_N}[X_N] \end{bmatrix}$$

and we define the $N \times N$ variance covariance matrix of $X$ as $E_X[(X - E_X[X])(X - E_X[X])'] = E_X[XX'] - E_X[X]E_X[X]'$ whose $(i,j)$ element is $Cov(X_i, X_j)$ which is $Var(X_i)$ when $i = j$.

Here for example

$$\begin{aligned} E_{X_1}[X_1] &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_1 f_X(x) dx_1 \ldots dx_N \\ &= \end{aligned}$$

Now consider a vector random variable $X$ partitioned thus: $X' = (X_1', X_2')$ with joint distribution and density functions respectively $F_X(x_1, x_2)$ and $f_X(x_1, x_2)$ where $X_i$ has $M_i$ elements.

The *marginal distribution function* of, say, $X_2$ is

$$F_{X_2}(x_2) = F_X(\infty, x_2),$$

the *marginal density function* of $X_2$ is

$$f_{X_2}(x_2) = \frac{\partial}{\partial x_2} F_X(\infty, x_2).$$

Alternatively

$$f_{X_2}(x_2) = \int\limits_{x_1 \in \Re^{M_1}} f_X(x_1, x_2) dx_1$$

and the *conditional density function* of $X_1$ given $X_2$ is

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)}.$$

The *regression* (function) of scalar $g(X_1)$ on $X_2$ is

$$E_{X_1|X_2}[g(X_1)|X_2 = x_2] = \int\limits_{x_1 \in \Re^{M_1}} g(x_1) f_{X_1|X_2}(x_1|x_2) dx_1.$$

### 7.1. The multivariate normal distribution

If $M-$element $X$ has a *multivariate normal distribution* then its probability density function takes the form

$$f_X(x) = (2\pi)^{-M/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(x-\mu)' \Sigma^{-1}(x-\mu))$$

where $\Sigma$ is symmetric positive definite, $M \times M$. We write $X \sim N_M(\mu, \Sigma)$.

To develop the moments of $X$ and some other properties of this distribution it is particularly helpful to employ the *multivariate moment generating function*, $M_X(t) = E_X[\exp(t'X)]$ where $t$ is a $M-$element vector. This is just an extension of the idea of the simple moment generating function introduced earlier. We can get moments of $X$ by differentiating $M_X(t)$. For example,

$$\frac{\partial}{\partial t_i} M_X(t) = E_X[X_i \exp(t'X)]$$

and so

$$\frac{\partial}{\partial t_i} M_X(t)|_{t=0} = E_X[X_i].$$

Check that the derivative of $M_X(t)$ with respect to $t_i$ and $t_j$ evaluated at zero gives $E_X[X_i X_j]$.

The multivariate normal moment generating function is obtained as follows.

$$\begin{aligned}
M_X(t) &= \int \cdots \int (2\pi)^{-M/2} |\Sigma|^{-1/2} \exp(t'x) \exp(-\frac{1}{2}(x-\mu)' \Sigma^{-1}(x-\mu)) dx \\
&= \int \cdots \int (2\pi)^{-M/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}\left(x'\Sigma^{-1}x - 2x'\Sigma^{-1}(\mu + \Sigma t) + \mu'\Sigma^{-1}\mu\right)) dx \\
&= \exp(t'\mu + \frac{1}{2}t'\Sigma t) \\
&\quad \times \int \cdots \int (2\pi)^{-M/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(x-(\mu+\Sigma t))' \Sigma^{-1}(x-(\mu+\Sigma t))) dx \\
&= \exp(t'\mu + \frac{1}{2}t'\Sigma t).
\end{aligned}$$

Note that this reproduces the result for the univariate normal distribution if we set $M = 1$. Differentiating with respect to $t$ once and then twice, on each occasion setting $t = 0$ gives $E_X[X] = \mu$, $Var[X] = \Sigma$.

Here is another use of the moment generating function. Consider a linear function of $Z = BX$ where $B$ is $R \times M$. The moment generating function of $R-$element $Z$ is

$$\begin{aligned}
M_Z(t) &= E_Z[\exp(t'Z)] \\
&= E_X[\exp(t'BX)] \\
&= \exp(t'B\mu + \frac{1}{2}t'B\Sigma B't) \qquad (7.1)
\end{aligned}$$

from which we can conclude that $Z \sim N_R[B\mu, B\Sigma B']$. So, *all linear functions of normal random variables are normally distributed.* In particular every element of $X$, $X_i$ is univariate normal with mean $\mu_i$ and variance equal to $\Sigma_{ii}$ which is the $(i,i)$ element of $\Sigma$.

Partition $X$ so that $X' = (X_1' \vdots X_2')$ where $X_i$ has $M_i$ elements and partition $\mu$ and $\Sigma$ conformably,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Note that

$$X_1 = Q_1 X$$

where $Q_1 = \left[ I_{M_1} \vdots 0 \right]$. Employing this matrix $Q_1$ in (7.1) leads to $X_1 \sim N_{M_1}[\mu_1, \Sigma_{11}]$ and similarly for $X_2$.

With the marginal density functions in hand we can now develop the *conditional distributions* for multivariate normal random variables. Dividing the joint density of $X_1$ and $X_2$ by the marginal density of $X_2$ gives, after some algebra, the conditional distribution of $X_1$ given $X_2$,

$$X_1 | X_2 = x_2 \sim N_{M_1}[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}].$$

So, we have

$$E[X_1 | X_2 = x_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2).$$

and

$$Var[X_1 | X_2 = x_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

In the multivariate normal case then, mean regression functions are all linear, and conditional variances are *not* functions of the conditioning variable. We say that the variation about the mean regression function is *homoscedastic*. Of course conditional variances change as we condition on different variables.

Notice that if the covariance of $X_1$ and $X_2$ is small, the regression of $X_1$ on $X_2$ is insensitive to the value of $X_2$ and the conditional variance of $X_1$ given $X_2$ is only a little smaller than the marginal variance of $X_1$.

Suppose we consider only a subset, $X_2^I$ say, of the variables in $X_2$ ("I" for included). The conditional distribution of $X_1$ given $X_2^I$ is derived as above, but from the joint distribution of $X_1$ and $X_2^I$ *alone*. We have

$$\begin{bmatrix} X_1 \\ X_2^I \end{bmatrix} \sim N\left( \begin{bmatrix} \mu_1 \\ \mu_2^I \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12}^I \\ \Sigma_{21}^I & \Sigma_{22}^I \end{bmatrix} \right)$$

where $\mu_2^I$, $\Sigma_{21}^I$ contain only the rows in $\mu_2$ and $\Sigma_{21}$ respectively relevant to $X_2^I$. Similarly $\Sigma_{22}^I$ contains only rows and columns relevant to $X_2^I$.

It follows directly that the conditional distribution of $X_1$ given $X_2^I$ is

$$X_1 | X_2^I = x_2^I \sim N_{M_1}[\mu_1 + \Sigma_{12}^I \left(\Sigma_{22}^I\right)^{-1} \left(x_2^I - \mu_2^I\right), \Sigma_{11} - \Sigma_{12}^I \left(\Sigma_{22}^I\right)^{-1} \Sigma_{21}^I].$$

Notice that the coefficients in the regression function and the conditional variance both alter as we condition on different variables but that in this normal case the regression function remains linear with homoscedastic variation around it.

### 7.2. Iterated expectations

Now consider the extension of the law of iterated expectations. Consider three random variables, $X_1$, $X_2$ and $X_3$. First, using the law for the two variable case, and conditioning throughout on $X_1$ we have

$$E_{X_3|X_1}[X_3|X_1] = E_{X_2|X_1}[E_{X_3|X_2X_1}[X_3|X_2, X_1]|X_1].$$

The result is some function of $X_1$. Now apply the law for the two variable case again. We get the following.

$$E_{X_3}[X_3] = E_{X_1}[E_{X_2|X_1}[E_{X_3|X_2X_1}[X_3|X_2, X_1]|X_1]].$$

Now develop the law for the case of four random variables. You should see the structure of the general law for $N$ random variables.

### 7.3. Omitted variables?

In some econometrics textbooks you will read a lot of discussion of "omitted variables" and the "bias" in estimators that results when we "omit regressors" from models. We too will look at this "bias". The development in the previous section suggests that we can think of this in the following way.

When we estimate regression functions using different regressors we are estimating *different parameters*, that is different regression coefficients. In the multivariate normal setting, when $X_2$ is used as the set of conditioning variables, we estimate $\Sigma_{12}\Sigma_{22}^{-1}$ as the coefficients on $x_2$, and when the set of conditioning variables $X_2^I$ is used, we estimate $\Sigma_{12}^I \left(\Sigma_{22}^I\right)^{-1}$ as the coefficients on $x_2^I$.

Of course $X_2^I$ consists of variables that appear in $X_2$. These common variables may have different coefficients in the two regression functions, but they may not have. In particular if the covariance between $X_2^I$ and the remaining elements in $X_2$ is zero then the coefficients on $X_2^I$ will be the same in the two regression equations.

In this multivariate normal setting the "bias" that is talked of arises when we take estimates of one set of regression coefficients and regard them (usually incorrectly) as estimates of a different set of regression coefficients. Outside the multivariate normal setting there are additional considerations.

These arise because the multivariate normal model is very special in that its regression functions are all linear. In most other joint distributions this uniform linearity of regression functions, regardless of the conditioning variables, is *not generally present.*

We can write the regression of $X_1$ on $X_2^I$ as equal to the conditional expectation of the regression of $X_1$ on the complete $X_2$ with respect to the conditional distribution of $X_2$ given $X_2^I$. Let $X_2^E$ denote the excluded elements of $X_2$ and write the regression of $X_1$ on $X_2$ as

$$E[X_1|X_2 = x_2] = \beta_I' x_2^I + \beta_E' x_2^E.$$

Then the regression of $X_1$ on $X_2^I$ is

$$E[X_1|X_2^I = x_2^I] = \beta_I' x_2^I + \beta_E' E[X_2^E|X_2^I] = x_2^I].$$

The additional consideration alluded to above is that except in very special circumstances, outside a multivariate normal setting, $E[X_2^E|X_2^I = x_2^I]$ is *not a linear function of* $x_2^I$.

One implication of this is that when we see non-linearity in a scatter plot for data on two variables, it may be the case that there is a linear effect for one variable on the other but in the context of a wider model in which we condition on a larger set of variables.

### 7.4. Regression functions and linearity

As noted earlier, much econometric work focuses on the estimation of regression functions and it is common to find restrictions imposed on the functional form of a regression function, sometimes flowing from economic theory, but often not. In microeconometric work the conditioning variables in an econometric regression model usually capture features of the agents' environments.

From now on we will use the symbol $Y$ to denote the random variable whose conditional distribution is of interest and we will use the symbol $X$ to denote regressors, $k$ in number unless noted.

The elementary textbooks all start, as we shall shortly do, by considering *linear regression functions* and a single response, that is the case in which $Y$ is a scalar random variable, and there exists a column vector of constants $\beta$ such that, for all $x$,

$$E[Y|X = x] = \beta' x.$$

In analysing multiple responses (vector $Y$) too, it is common to find a linear regression function assumed, that is that there exists a matrix of constants, $B$, such that for all $x$,

$$E[Y|X = x] = Bx.$$

Surprisingly, given the ubiquity of linearity assumptions like these, it is hard to find any element of economic theory which predicts linearity of regression functions. Linearity is usually an empirical issue - if we employ a linear model then we should try to see if the linearity restriction is appropriate.

Suppose that in fact the regression of $Y$ on $X$ is a nonlinear function of $x$, say

$$E[Y|X = x] = g(x, \theta). \tag{7.2}$$

Taking a Taylor series expansion of $g(x, \theta)$ around some central point $x_0$, in the distribution of $X$ will lead to a linear approximation

$$E[Y|X = x] \doteqdot g(x_0, \theta) + \beta'(x - x_0)$$

where the $i$th element of this vector $\beta$ is

$$\beta_i = \frac{\partial}{\partial x_i} g(x, \theta)|_{x=x_0}.$$

So, if the second derivatives of the function $g(x, \theta)$ are not very large over the main part of the range of $X$ then a linear model may be a good approximation. Taking the Taylor series one more step produces a quadratic approximation. We might find (but not necessarily) that a quadratic approximation

$$E[Y|X = x] = \beta_0 + \beta' x + x' A x$$

where $\beta_0 = g(x_0, \theta)$, is close to the true nonlinear regression. In some applied work you will see linear models extended by the addition of polynomial functions of regressors.

A simpler, and it turns out easier to estimate version of the general nonlinear regression function (7.2) is the following

$$E[Y|X = x] = g(x'\theta).$$

in which the conditioning variables combine linearly but their *combined effect* on the expectation of $Y$ is *nonlinear*. This sort of restriction is known as a "single index" restriction.

An important, implication of a single index restriction of this sort is that

$$\frac{\partial}{\partial x_i} E[Y|X = x] = g'(x'\theta)\theta_i$$

where

$$g'(z) = \frac{\partial}{\partial z} g(z).$$

This implies

$$\frac{\frac{\partial}{\partial x_i} E[Y|X = x]}{\frac{\partial}{\partial x_j} E[Y|X = x]} = \frac{\theta_i}{\theta_j}.$$

The ratio of two partial derivatives of the regression function at every value of $x$ is independent of $g(\cdot)$ and of $x$. This provides us with a route to investigating whether a single index assumption is appropriate and to a way of estimating ratios of the $\theta_i$'s that does not require specification of $g(\cdot)$. Estimators of this sort are known as *semi-parametric estimators*.

We have started talking about estimation of regression functions. It is time to consider how this can be done.