ECON 468

CHAPTER 1 : REGRESSION MODELS

In this chapter, we'll look at:

- simple linear regression model
- distributions, densities and moments
- specification of linear regression models
- estimating the regression model (method of moments, OLS)

1.1 SIMPLE LINEAR REGRESSION MODEL

$$\underset{(n\times 1)}{y} = \underset{(n\times k)}{X} \cdot \underset{(k\times 1)}{\beta} + \underset{(n\times 1)}{u}$$

- y: dependent variable
  The $(n\times 1)$ vector comprises observations of the variable, sample size $= n$. There are many kinds of variables, eg time series variables, flow variables such as GDP per annum, or cross-sectional variables as in census data.
- X: independent / explanatory variable(s)
  It is an $n\times k$ matrix, with k representing the number of independent variables – each of them take a column in X. As a partitioned matrix, we can write $X = [\underset{\sim}{x}_1 \; \underset{\sim}{x}_2 \; \cdots \; \underset{\sim}{x}_k]$, each $\underset{\sim}{x}_i$ corresponds to a single explanatory variable.
- β : regressor
  β is not observed; it is an unknown parameter that the model seeks to identify. ("how much does X explain y"?)
- u: error term
  u is also not observed, it is assumed that it is a random variable. We use randomness to model our ignorance of other real world factors that determine y. Generally, it is assumed that $E(u) = 0$. We can express u as a function of β: $u = y - X\beta$.

We can think of models as a set of data-generating processes (DGP). DGP refers to whatever mechanism is at work in the real world of economic activity giving rise to the numbers in our samples.

1.2 DISTRIBUTIONS, DENSITIES AND MOMENTS

Random variables: ~~[illegible]~~ → The variables that appear in an econometric model are treated as r.v.s, they are representations of real world variables we wish to consider random / unexplained.

A r.v. is a collection of possibilities; what we observe are the realizations of the r.v., which is one value out of the set of possible values.

- discrete r.v.: takes on a finite, or a countably infinite number of values which can be denoted x_i, $i = 1, 2, \ldots, n$
  - $0 \leq p_i \leq 1$, where p_i = prob of x_i
  - $\sum_{i=1}^{n} p_i = 1$

- continuous r.v.: for a scalar r.v, this means it could take any value on the real number line.
  eg standard uniform distribution: $x \sim U(0,1)$
  - $P[a, b] = b - a$.
  - $P[0, 1] = 1$
  - $P[a, a] = 0$ (ie probability of any specific value is 0.)

3 rules for probability distributions:

1. All probabilities lie between 0 and 1
2. The null set is assigned probability 0, and the full set of possibilities is assigned probability 1
3. The probability assigned to an event that is the union of 2 disjoint events is the sum of the probabilities assigned to those disjoint events.

Cumulative distribution function (CDF):

The CDF is denoted $F(x) = Pr(X \leq x)$, the probability that X is equal to or less than some value x.

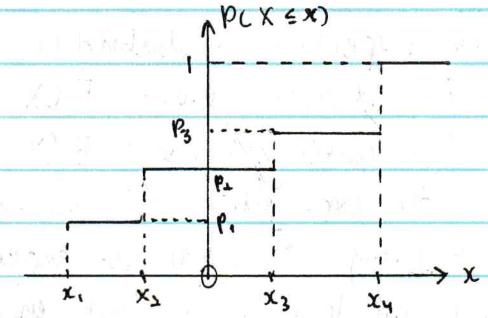
- As $x \to -\infty$, $F(x) \to 0$
- As $x \to \infty$, $F(x) \to 1$.

- for discrete r.v.:

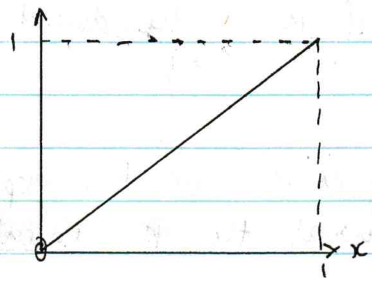
  $F_X(x_1) = p_1$

  $F_X(x_2) = p_1 + p_2$

  ... and so on



- for continuous r.v.:



in the uniform distribution, $F_U(x) = x$

Probability density function (PDF):

A PDF exists only when the r.v. is continuous, and its CDF is differentiable. For a discrete r.v., the derivative does not exist since the CDF is discontinuous. We can generally say it is not useful to consider the density of a discrete r.v.

The PDF is denoted $f(x) \equiv F'(x)$.

↳ It is non-negative, since it is the derivative of a weakly increasing function

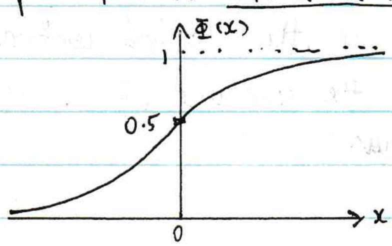
↳ $x > y \Rightarrow F(x) \geq F(y)$

- $\Pr(a \leq X \leq b) = F(b) - F(a)$

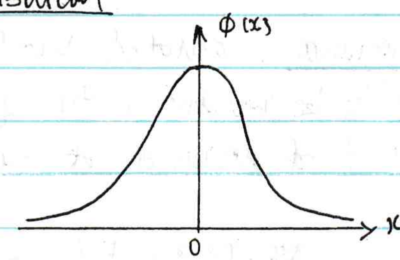
  $= \int_a^b f(x)\,dx$

example: PDF of standard normal distribution

CDF:



PDF:



PDF of standard normal denoted $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$

Moments of r.v.s

A fundamental property of a r.v. is its expectation. The expectation is the first moment of an r.v., it is an abstraction of the notion of "average": a weighted average of different realizations, based on their respective probabilities.

- for discrete r.v.: $E(X) \equiv \sum_{i=1}^{n} p_i x_i$
- for continuous r.v.: $E(X) \equiv \int_{-\infty}^{\infty} x f(x)\, dx$

(the expectation is defined analogously using the PDF.)

Not every r.v. has an expectation! $E(X)$ could diverge if f does not tend to 0 fast enough, or if n in the discrete expectation is infinite.

Higher moments of an r.v., if they exist, are the expectations of the r.v. raised to a power. In general, the k^{th} <u>uncentered moment</u> of X is $E(X^k)$.

$$m_k(X) \equiv \int_{-\infty}^{\infty} x^k f(x)\, dx$$

If a distribution possesses a k^{th} moment, it also possesses all moments of order $< k$. These moments are called "uncentered" as X does not have a mean of 0.

We prefer to look at <u>centered / central moments</u>, which are defined as the ordinary moments of the difference between the r.v. and its expectation, ie $E(X - E(X))^k$.

- $\mu_k \equiv E(X - E(X))^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\, dx$, $\mu = E(X)$.
  if X is a continuous r.v.
- if X is discrete, $\mu_k \equiv E(X - E(X))^k = \sum_{i=1}^{n} p_i (x_i - \mu)^k$

<u>Variance</u>, denoted $Var(X) = \sigma^2$, is the second central moment. A variance cannot be negative; the square root of the variance, σ, is called the standard deviation of the distribution.

$$\begin{aligned} Var(X) &= E(X - E(X))^2 \\ &= E(X^2 - 2X E(X) + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

example: $Y \sim N(m, \sigma^2)$

$Z \sim N(0,1)$ is the standard normal distribution.

Hence we can express $Y = m \pm \sigma Z$

$$E(Y) = E(m + \sigma Z) = m + \sigma \underbrace{E(Z)}_{=0} = m$$

$$\mathrm{Var}(Y) = E(Y - E(Y))^2 = E(m \pm \sigma Z - m)^2 = E(\sigma Z)^2 = \sigma^2$$

CDF of Y? $F_Y(y) = P(Y \leq y)$
$= P(m + \sigma Z \leq y)$
$= P(Z \leq \frac{y-m}{\sigma})$
$= \Phi(\frac{y-m}{\sigma})$

CDF of Z denoted by Φ.

PDF of Y? $f_Y(y) = F_Y'(y)$
$= \Phi'(\frac{y-m}{\sigma})$
$= \frac{1}{\sigma}\phi(\frac{y-m}{\sigma})$

Multivariate Distributions

So far, we've only considered univariate distributions, where the r.v.s map to the real number line. A vector-valued r.v. takes on values that are vectors. It can be thought of as several scalar r.v.s that have a single joint distribution.

Consider a bivariate r.v. (vector length = 2), (X_1, X_2).

- CDF of (X_1, X_2): $F_{X_1,X_2}(x_1, x_2) = \underbrace{Pr[(X_1 \leq x_1) \cap (X_2 \leq x_2)]}_{\text{joint probability that both } X_1 \leq x_1 \text{ and } X_2 \leq x_2}$
- PDF of (X_1, X_2): $f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$ ("joint density function")

X_1 and X_2 are said to be statistically independent if the joint CDF is the product of $\mathrm{CDF}(X_1)$ and $\mathrm{CDF}(X_2)$.

$$F_{X_1,X_2}(x_1, x_2) = \underbrace{F_{X_1}(x_1, \infty)}_{\text{marginal CDF of } X_1} F_{X_2}(\infty, x_2)$$

The definition of <u>marginal density</u> follows from the marginal CDF:

$f(x_1) \equiv F_1(x_1, \infty)$, F_1 denotes the partial derivative of F wrt to x_1.

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \quad \text{(in terms of joint density)}$$

Another condition for independence is $f(x_1, x_2) = f(x_1) \cdot f(x_2)$.

Intuition for marginal distribution:

$$F_{X_1}(x_1) = \lim_{x_2 \to \infty} F_{X_1, X_2}(x_1, x_2)$$

Since if $x_2 \to \infty$, $Pr[(X_1 \leq x_1) \cap (X_2 \leq x_2)] \approx Pr(X_1 \leq x_1)$ as X_2 will always be $\leq \infty$.

<u>Conditional Probability</u>

Suppose A and B are 2 events. The probability of A conditional on B, or given B, is denoted $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

$$P(A \cap B) = P(A \mid B)\, P(B)$$

The idea is that, if we somehow know that B has been realized, then we can use that knowledge to know if A has also been realized:

example: A and B are disjoint, $P(A \cap B) = 0$. If B is realized, we know that A has not been, ie $P(A \mid B) = 0$.

example: $B \subseteq A$, $P(A \cap B) = P(B)$. If B is realized, then $P(A \mid B) = 1$.

- conditional densities: Say we have 2 r.v.s X_1 and X_2. The distribution of X_1 is conditional on some specific realized value of X_2. Hence, the conditional distribution gives us the probabilities of events in X_1, given that the realization of $X_2 = x_2$.

$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$

Using Bayes' Theorem, we also know that $f(x_1 \mid x_2) f(x_2) = f(x_2 \mid x_1) f(x_1) = f(x_1, x_2)$

Conditional Expectations

The conditional expectation is the ordinary expectation computed using the conditional distribution.

For a given x_2, $E(X_1 | x_2)$ is a deterministic (ie non-random) quantity. We can consider $E(X_1 | x_2)$ $\forall x_2$ to construct a new r.v., $E(X_1 | X_2)$, with realization $E(X_1 | x_2)$. This r.v. is a deterministic function of the r.v. X_2.

Some useful properties:

- Law of Iterated Expectations: $E(E(X_1 | X_2)) = E(X_1)$
- any deterministic function of a conditioning variable X_2 is its own conditional expectation: eg $E(X_2 | X_2) = X_2$
  $E(X_2^\alpha | X_2) = X_2^\alpha$, $\alpha \in \mathbb{R}$.
- conditional on X_2, the expectation of a product of another r.v. X_1 and a deterministic function of X_2 is the product of that deterministic function and the expectation of X_1 conditional on X_2:
  $E(X_1 \cdot h(X_2) | X_2) = h(X_2) \cdot E(X_1 | X_2)$, $h(\cdot)$ is any deterministic fn.

example: $E(X_1 | X_2) = 0$

$$\begin{aligned} E(X_1 \cdot h(X_2)) &= E[E(X_1 \cdot h(X_2) | X_2)] \quad \text{by Law of It. Exp.} \\ &= E(h(X_2) \cdot E(X_1 | X_2)] \quad \text{by property \#2} \\ &= E(h(X_2) \cdot 0) \\ &= E(0) = 0 \end{aligned}$$

SPECIFICATION OF REGRESSION MODELS

A key assumption is that $E(u_t | X_t) = 0$. With this assumption, we can obtain from the simple linear regression model:

$$\begin{aligned} E(y_t | X_t) &= \beta_1 + \beta_2 X_t + E(u_t | X_t) \\ &= \beta_1 + \beta_2 X_t \end{aligned}$$

In general, we want to condition on exogenous variables, not endogenous ones. An exogenous variable has its origins outside the model under consideration (eg X); the mechanism generating the endogenous variable is captured in the model. In the linear regression model, y is endogenous.

Error terms

When we specify a regression model, it is essential to make assumptions about the properties of error terms. The simplest assumption is that all of the error terms have mean 0, come from the same distribution, and are independent of each other.

A very strong assumption often made about error terms is that they are independently and identically distributed (IID). This means the error terms are mutually independent, and are realizations from the same identical probability distribution.

When are error terms not IID?

- serial correlation: when successive observations are ordered by time, an error term might be correlated to the neighbouring error terms if there is correlation across time periods of ~~add~~ random factors that influence the DV but are not accounted for in the regression function.
- heteroskedasticity: the variance of the error terms may be systematically larger for some observations than for others.

A complete specification of an econometric model is one that provides an unambiguous recipe for simulating the model to generate simulated data.

- deterministic specification: β
- stochastic specification: u

$u \perp\!\!\!\perp X \Rightarrow E(u \mid X) = 0$ (ie u is statistically independent to X)

By Law of Iterated Expectations, $E[E(u \mid X)] = E(u) = 0$.

Is the model correctly specified? $\exists\, \beta$ such that $u(\beta) = y - X\beta$

↳ true DGP $\in$ model

Non-linear regression model

Non-linearity refers to the parameters (ie β) and not the r.v.s!
For example:

- $y_t = \beta_1 + \beta_2 X_t + \beta_3 X_t^2 + u_t$ is a multiple linear regression model. The r.v. X is non-linear, $E(y_t | X_t)$ varies quadratically with X_t when $\beta_3 \neq 0$. However, it is linear with regard to the parameter β_i; when $\beta_3 = 0$, it reduces to the simple linear regression model.
- $y_t = \delta_1 + \delta_2 \frac{1}{X_t} + u_t$ is a linear regression ~~model~~ model. Even though $E(y_t | X_t)$ may depend nonlinearly on X_t, it still depends linearly on the unknown parameters of the regression function.
- $y_t = e^{\beta_1} \cdot X_{t2}^{\beta_2} \cdot X_{t3}^{\beta_3} + u_t$ is a nonlinear regression model. The regression function is multiplicative, and is not linear in the parameters β_2 and β_3.
- $y_t = \alpha + \beta X_{t1} + \frac{1}{\beta} X_{t2} + u_t$ is a nonlinear regression model.

1.5 METHODS OF MOMENTS ESTIMATION

We can estimate parameters by replacing population means by sample means; this technique is called the method of moments. In general, the method of moments estimates population moments by the corresponding sample moments. In order to apply this method to regression models, we must use the facts that population moments are expectations, and that regression models are specified in terms of the conditional expectations of the error terms.

For the linear regression model: $y_t = X_t \beta + u_t$.
Suppose that u_t are IID $\Rightarrow E(u_t) = 0$.
We have a sample size n, so we have n error terms. We consider the sample mean of the error terms ~~s~~ and set it to 0; the population mean.

$$E(u_t) = \frac{1}{n} \sum_{t=1}^{n} u_t = \frac{1}{n} \sum_{t=1}^{n} (y - X_t \beta) = 0 \quad (\text{MM: sample mean} \leftarrow \mu.)$$

We can use the fact that our model specifies that the mean of u_t is 0 conditional on the explanatory variable X_t. Not only is $E(u_t) = 0$, $E(X_t u_t) = 0$ too.

Proof that $E(X_t u_t) = 0$:

$$E(X_t u_t) = E[E(X_t u_t | X_t)] \quad \text{by Law of It. Exp.}$$
$$= E(X_t \cdot E(u_t | X_t))$$
$$= E(X_t \cdot 0)$$
$$= 0$$

We can supplement $E(u_t)$ as such to obtain the sample mean:

$$E(X_t u_t) = \frac{1}{n} \sum_{t=1}^{n} \underbrace{X_t}_{\text{instrumental variable}} \underbrace{(y - X_t\hat{\beta})}_{\text{zero function}} = 0$$

In matrix algebra: $\rightarrow \quad E(X^T u) = 0$

$$X^T (y - X\hat{\beta}) = \underset{\sim}{0}$$
$$X^T y = X^T X \hat{\beta}$$
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where $\hat{\beta}$ is known as the ordinary least squares estimator.

<u>Least squares estimation</u>

The expression $y_t - X_t\beta_0$ is equal to the error term for the t^{th} observation, u_t, where β_0 is the correct value of β. If the same expression is thought of as a function of β, with β allowed to vary arbitrarily, it is called a <u>residual</u>.

The n-vector $y - X\beta$ is the vector of residuals.

The sum of the squares of the components of the vector of residuals is known as the <u>sum of squared residuals</u> (SSR).

$$SSR(\beta) = \sum_{t=1}^{n} (y_t - X_t\beta)^2$$

The idea of least squares estimation is to minimize the SSRs associated with a regression model. The parameter vector that minimizes SSR is the same as the estimator derived through the method of moments.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{n} (y_t - X_t\beta)^2, \quad \hat{\beta} = (X^T X)^{-1} X^T y$$