Chapter 3

Regression models

Some suggested texts for regression
Regression is covered in many textbooks, including the following:

- J. Johnston and J. Dinardo, *Econometric Methods*, 4th Edition, Mcgraw-Hill, 1997
- R.S. Pindyck and D.L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th Edition, McGraw-Hill International, 1998
- H. Theil, *Principles of Econometrics*, Wiley, 1971
- T. Amemiya, *Advanced Econometrics*, Harvard University, Press, 1985
- A. Spanos *Statistical Foundations of Econometric Modelling*, Cambridge University Press, 1986
- A. Harvey, *The Econometric Analysis of Time Series*, 2nd Edition, Philip Allan, 1981

For applied regression books:

- N.R. Draper and H. Smith, *Applied Regression Analysis*, 3rd Edition, Wiley Series in Probability and Statistics, 1998
- E.R. Berndt, *The Practice of Econometrics Classic and Contemporary*, Addison Wesley, 1991

3.1 What is a model?

The focus of econometrics is to measure a relationship, (often motivated through economic theory), or analyse and describe an actual phenomenum or process. The theory or process can be represented by a *model*, which in practice can be summarized by data. The model represents a simplification of the actual phenomenum to be explained or predicted. Statistical thory can then be used to evaluate the model and test relationships between (economic) variables.

Consider, as an example, a deterministic model of consumption behaviour,

$$\log C = a \log R + b, \text{ where } a, b \in \mathcal{R}.$$

where a is the *consumption elasticity* with respect to *income*.[1] The model can be made stochastic with the addition of an *error* term u, so that,

$$\log C_t = a \log R_t + b + u_t, \text{ for } t = 1, ..., T \text{ observations.}$$

The disturbance term represents the difference between the observation $\log C$ and the approximate mean, $a \log R + b$. A large discrepency could indicate:

- the existence of a non-linear relationship between $\log R$ and $\log C$;
- varying coefficients a and b, over time;
- omission of other variables;
- measurement error in $\log R$ and/or $\log C$.

An explanatory model is used to assess the effect some variables may have on another set of variables. A causal relation is implied, as a distinction is made between the variables that are explained and others that affect them. A descriptive model does not make this distinction – an example of this is the prediction of an election result from survey data.

[1] This example is taken from Gourieroux and Monfort (1995, Chapter 1).

3.1.1 General formulation

In the preceding section, y_t has been i.i.d. Here we extend this to the case where the observations depend linearly on some known (fixed) $(k \times 1)$ vector of regressors X_t. The basic structure will be that

$$y_t = X_t\beta + u_t, \qquad \text{where } u_t \sim \text{ i.i.d.}$$
$$\text{and } \mathrm{E}(u_t) = 0, \ \ \mathrm{var}(u_t) = \sigma^2,$$

for $t = 1, ..., T$. This model is indexed by the parameters β and σ^2.

Interpretation

We can think of the regression model algebraically as simply a k dimensional plane in the space of $y_t \times X_t'$. However, it is also useful to work with a distributional understanding via the normal density. Suppose $X_t' = (1, X_t^*)'$ and we regard y_t and X_t^* as being jointly normally distributed with

$$\begin{pmatrix} y_t \\ X_t^* \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \right\}.$$

Then

$$\begin{aligned} y_t | X_t^* &\sim N\left\{ \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(X_t^* - \mu_x\right), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \right\} \\ &= N\left\{ \mu_{yx} + \Sigma_{yx}\Sigma_{xx}^{-1} x_i^*, \sigma^2 \right\} \\ &= N(X_t'\beta, \sigma^2). \end{aligned}$$

So a regression model with Gaussian errors can be thought of as the correct conditional model when the data and regressors are jointly Gaussian.

3.1.2 Econometric models

An econometric model consists of the following:

(1) A set of *behavioural equations* derived from an economic model or statistical theory, involving some observed variables and some *disturbances*.
(2) An indication of whether the random variables are observed with error.
(3) A specification of the probability distribution of the disturbances and error of measurement.

The choice of a particular model depends on many factors, including its plausibility, ease of estimation, goodness of fit and ability to forecast. The validity of the model depends on its ability to dominate other models.

3.1.3 Types of models

Macro/microeconometric models, in general, are based on economic theory yielding statis/long-run relationships. Examples include modelling cross sectional and/or time-series data. Data can be quantitative or qualitative; in the latter case, these can often be expressed numerically through the use of *dummy* variables, (taking one of two possible values).

Time series data

These measure a particular variable during successive time periods or at different dates. The observations are often successive and at equi-distant time intervals.

Cross-sectional data

These measure a particular variable at a given time point, for different identities, (e.g. over different countries). Examples include analysing households or firms at a given date, where the data has been obtained from surveys.

Pooling cross-sectional and time series data can be interpreted as a cross section of a time series or a time series of cross sections, so that

$$y_{it} = \alpha_i + \beta_i X_{it} + u_{it}$$

for $i = 1, ..., n$ and $t = 1, ..., T$. These are often termed *panel* data.

3.1.3.1 Statistical models

A Statistical model is a pair $(\mathcal{Y}, \mathcal{P})$, where $\mathcal{Y}$ is the set of possible observations and $\mathcal{P}$ is a family of probability distributions. In practice, assumptions are imposed on $\mathcal{P}$ to simplify the model.

3.1.3.2 Linear models

Suppose we specify a model through $\mathrm{E}(Y_t|X_t) = \alpha + \beta X_t$,[2] where $\theta = (\alpha, \beta)'$ are the parameters of the model, so that

$$Y_t = \alpha + \beta X_t + u_t, \text{ where } \begin{cases} \text{for } t = 1, ..., T & \text{and} \\ \mathrm{E}(u_t|X_t) = 0 \end{cases}$$

This assumes $\mathrm{E}(Y_t|X_t)$ belongs to a linear subspace in $\mathbb{R}^T$. In the consumption example,

$$\begin{pmatrix} \log C_1 \\ \vdots \\ \log C_T \end{pmatrix} = a \begin{pmatrix} \log R_1 \\ \vdots \\ \log R_T \end{pmatrix} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} b + \begin{pmatrix} u_1 \\ \vdots \\ u_T \end{pmatrix},$$

so that the linear subspace is generated by the vectors

$$X_1 = \begin{pmatrix} \log R_1 \\ \vdots \\ \log R_T \end{pmatrix} \text{ and } X_2 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

3.1.3.3 Exponential models

Suppose we let $\theta \in \Theta$ represent the vector $(p \times 1)$ of parameters. A statistical model $(\mathcal{Y}, \mathcal{P})$, where $\mathcal{P} = \{\mathcal{P}_\theta, \theta \in \times\}$ is exponential if the distributions P_θ have densities of the form

$$c(\theta)\, h(y) \exp\left\{ \sum_{j=1}^{p} w_j(\theta) t_j(y) \right\},$$

where w_j and t_j are real-valued functions, (see Section 2.2). $\mathcal{P}$ is an exponential family and $t = \{t_1(y), ...t_p(y)\}'$ is known as the *canonical*[3] statistic.

[2] The conditional set-up will be motivated in Section 3.2.3.

[3] Expressed in its simplest form, which is often unique.

Example: Poisson

Let $\boldsymbol{Y} = (Y_1, ..., Y_T)' \sim$ POI(λ), where $\lambda \in \mathbb{R}_+$. The distribution of $\boldsymbol{Y}$ has the density

$$\begin{aligned} f(\lambda, y) &= \prod_{t=1}^{T} \exp(-\lambda) \frac{\lambda^{y_t}}{y_t!} \\ &= \left(\prod_{t=1}^{T} \frac{1}{y_t!} \right) \exp(-T\lambda) \exp\left\{ (\log \lambda) \sum_{t=1}^{T} y_t \right\}, \end{aligned}$$

that is, the family is exponential with $p = 1$, $t(y) = \sum_{t=1}^{T} y_i$ and $w_1(\lambda) = \log \lambda$.

3.1.3.4 Count data models

The dependent variable, y, takes integer values. Specific examples include binary dependent variables models, where $y = 0$ or 1; the binary nature of y in a *linear probability model* can cause estimation problems , (see Johnston and Dinardo 1997, Chapter 13).

Probit and Logit models

Alternatives to linear probability models are given by transforming $\mathrm{E}(y|X)$ into a probability, so that for some function F,

$$P(y_t = 1) = F(X_t \beta),$$

where F is often chosen as a distribution function. Choosing F as the standard normal gives the *Probit* model; whereas choosing the logistic distribution yields the *Logit* model.

3.1.3.5 Latent variable models

Suppose that an observable random variable, y, can take the values 0 and 1 and that we define a *latent* variable, y^*, as

$$y_t^* = X_t \beta + \epsilon_t$$

which is not observed. Often, y takes on the values 0 and 1, according to

$$y_t = \begin{cases} 1 & \text{if } y_t^* > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Assuming ϵ_t to be normal yields the Probit. If ϵ follows an *extreme value distribution*,[4] e.g. the Weibull distribution, which has heavier tails than the normal, then it is termed the Logit model.

3.1.4 Building a model (identification, estimation and diagnostics)

Three stage iterative procedure based on identification, estimation and diagnostic tests. See for example, Hendry (1995). Aim is to find a model that is 'approximately' correct.

(1) *Identification*: use of data and economic theory (& other information) on how series is generated in order to suggest *parsimonious* class of models to be considered. For example, one could formulate a model, look at scatter plots, residuals, correlograms, etc.
(2) *Estimation*: efficient use of data to make inferences about parameters conditional on adequacy of model being considered. Are the coefficients reasonable; is the relationship *statistically significant*?
(3) *Diagnostic checking*: compare fitted model to data with intent to reveal model inadequacies; improve model. Look at predictive performance etc.

[4] relating to the modelling of extreme events, in order to explain events that occur with small probability but have a significant effect on the model.

3.2 What are regressors?

A model represents a set of stochatic relations between the random variables of the process. The model determines the values of the variables which need to be explained – the *endogenous* or *dependent* variables, which are determined by the relations of the model. The model also includes *regressors*, (or *exogenous* / *independent* variables), which affect the endogenous variables, but are determined outside the system. Given data on the variables, statistical techniques are used to estimate the parameters of the model.

The model has, in general, two components: a *deterministic* component and a *random* component. The random component includes stochastic variables, which have probability distributions associated with them. The deterministic component includes variables which, given a model and dataset, are fixed or pre-determined.

3.2.1 Regression vs. correlation

Correlation analysis is closely linked to *regression* analysis. The main objective in the former is to measure the degree of linear association (i.e. strength) between two variables. In regression analysis, the aim is to estimate or predict the average value of one variable, given fixed values of the other variables.[5] These are often represented by analytical models (possibly derived from theoretical ones) of explanatory variables, $X \to Y$, the dependent variables.

Suppose we have (X_i, Y_i), for $i = 1, ..., n$ observations representing a sample from some population N. Direction and closeness of the linear association between two variables can be measured by the *correlation coefficient*, R. Let $\bar{X}$ and $\bar{Y}$ be the sample means of X and Y, respectively, then the data in *deviation* form is:

$$\begin{aligned} x_i &= X_i - \bar{X} \\ y_i &= Y_i - \bar{Y}, \end{aligned}$$

so that the sample covariance of X and Y is

$$\widehat{cov}(X,Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n} = \sum_{i=1}^{n} \frac{x_i y_i}{n}.$$

We can express the deviations using *standardized* variables, x_i/s_x, y_i/s_y to obtain the *correlation coefficient* as the covariance between these standardized quantities as

$$R = \sum_{i=1}^{n} \frac{x_i y_i}{n\, s_x\, s_y},$$

where $s_x = \sqrt{\sum_{i=1}^{n}(x_i^2/n)}$, $s_y = \sqrt{\sum_{i=1}^{n}(y_i^2/n)}$ are the standard deviations and $-1 \leq R \leq 1$. Notice that both X and Y are treated as random variables and no distinction is made between the dependent and explanatory variables.

3.2.2 Fixed regressors with errors

In regression analysis, the dependent variable is assumed to be stochastic and has a probability distribution associated with it. The regressors are assumed to be fixed in repeated sampling, that is, X is the same over various samples. In most regression models, the assumption of fixed regressors is made.

Suppose that the relationship between X and Y is expressed as

$$Y_i = g(X_i, u_i) \tag{3.1}$$

where the function g depends on the unknown parameters and u_i is the *discrepancy* or *disturbance*.

The X_i's are assumed to be fixed, while the disturbance u_i is assumed random. This implies that the family of distributions for Y is also random, determined by the relation given in (3.1) and the family of distributions chosen for u.

[5] Regression analysis is typically defined as the study of the dependence of one variable (the dependent variable) on one or more expanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

3.2.3 Regression via conditioning

An equivalent formulation is to suppose that the X_i's are random and that the family of distributions for X is the family of conditional distributions for u, given X. The relation given in (3.1) then determines the family of conditional distributions for Y, given X. A statement on the distribution of u in this case is thus a statement about the conditional distribution of Y, given $X = x$.

The joint distribution can be expressed as $f(X,Y) = f(X)\, f(Y|X)$ where we are usually interested in $f(Y|X)$, depending on the *exogeneity* status of X in relation to Y. Suppose economic theory suggests $\mathrm{E}(Y|X) = g(X)$, and that we consider a linear relationship, $\alpha + \beta X$. For each i,

$$\begin{aligned} \mathrm{E}(Y|X_i) &= \alpha + \beta X_i \\ u_i &= Y_i - \mathrm{E}(Y_i|X_i) = Y_i - \alpha - \beta X_i. \end{aligned}$$

Conditional models A *conditional model* is a set of possible values for a random variable Y and a family of conditional distributions for Y, $f(Y|X = x)$.

3.2.4 The simple linear regression model

The classical (simple) linear regression model satisfies the following assumptions:

$$X_i \text{ is non-stochastic} \tag{3.2}$$

$$\mathrm{E}(u_i) = 0 \qquad \forall\, i \tag{3.3}$$

$$\mathrm{var}(u_i) = \mathrm{E}(u_i^2) = \sigma^2 \qquad \forall\, i \tag{3.4}$$

$$\mathrm{cov}(u_i, u_j) = \mathrm{E}(u_i u_j) = 0 \qquad \forall\, i \neq j. \tag{3.5}$$

In (3.2), the $\{X_i\}$ are treated as fixed, so one can treat the properties of u_t unconditionally (without conditioning on $X_1, ..., X_n$). Assumptions (3.3), (3.4) and (3.5) can be represented by $u_i \sim$ iid$(0, \sigma^2)$.[6] There are three parameters to estimate: α, β and σ^2. Both α and β are taken as a pair, using some numerical estimate, $\bar{\alpha}$ and $\bar{\beta}$ to fit a line, $\hat{Y}_i = \bar{\alpha} + \bar{\beta} X_i$; the residuals from the line are then used to form an estimate of σ^2. Suppose we consider a possible sequence of Y populations given in the top panel of Figure 3.1. The assumptions above imply certain assumptions about the population. The bottom panel of Figure 3.1 show the assumptions imposed to analyse the problem: the probability distributions $P(Y_i|X_i)$ have the same variance for all X_i; the means $\mathbb{E}(Y_i) = \mu_i$ lie on a straight line and that the random variables Y_i are uncorrelated. The assumption of normality has also been imposed on the residuals.

Let the residuals from any fitted straight line be denoted by

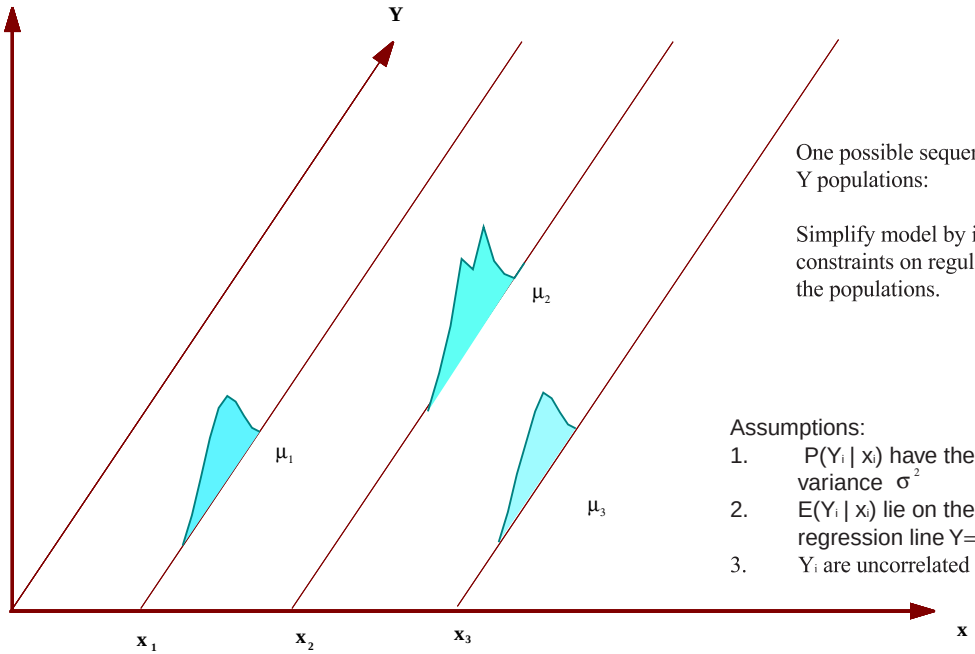
$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\alpha} - \widehat{\beta} X_i \text{ for } i = 1, ..., n,$$

and are depicted in Figure 3.2. Since Y is measured vertically and our objective is to minimize the error in explaining Y, the vertical distance is used as a measure of error. Each pair of $(\bar{\alpha}, \bar{\beta})$ values define a different line; hence a different set of $\{e_i\}_{i=1}^{n}$. The 'Least Squares' principle is to select one pair $\hat{\alpha}, \hat{\beta}$ to minimize the *Residual Sum of Squares* (RSS), that is,

$$\min_{\alpha,\beta} \text{ RSS } = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{3.6}$$

and is denoted the *least squares criterion*. We will see that a theoretical justification is provided by the Gauss Markov theorem and the maximum likelihood criterion for the normal regression model, (and produces an analogy to Pythagoras'

[6] Suppose we had a *time series*: $Y_{it} = \alpha + \beta X_{it} + u_{it}$, for $t = 1, ..., T$. Then for the model $Y_t = N\alpha + \beta X_t + U_t$, where X_t, Y_t and U_t are the sum over i of X_{it}, Y_{it} and u_{it} respectively, then assumptions (3.3), (3.4) and (3.5) mean that the disturbances, U_t, are uncorrelated, each having zero mean and a constant, finite variance, σ^2. If we make the additional assumption of independence over time, then $U_t \sim$ iid$(0, N\sigma^2)$. Taking them to be uncorrelated is a weaker assumption than independence, although if we were to assume normality, independence would immediately follow.



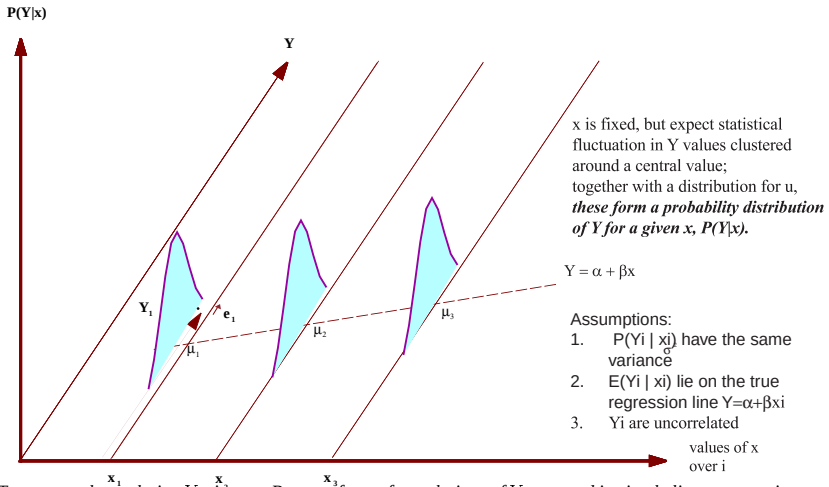


Figure 3.1 *Top: general population Y given x; Bottom: form of populations of Y assumed in simple linear regression..*

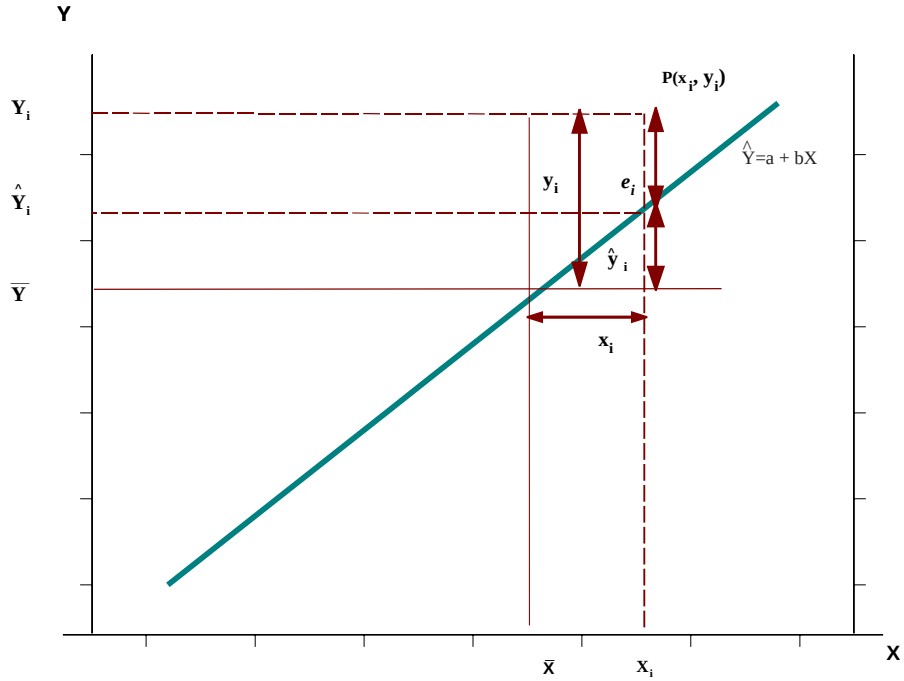


Figure 3.2 *Least-Squares estimators: Residuals from a fitted line..*

Geometric theorem). It also avoids the drawbacks of other distance measures such as $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$ and $\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$, known as the L_1 (least absolute deviation) or minimum absolute deviation (MAD) regression estimator since it minimizes the absolute deviations. An other example is the least medium absolute deviation (high breakdown) estimator, given by,

$$\widetilde{\beta} = \underset{\beta}{\arg\min}\, medium\, \{|y_1 - x_1'\beta|, ..., |y_n - x_n'\beta|\}\,.$$

The least squares line minimizes the sum of squared residuals, passes through the point $(\bar{X}, \bar{Y})$ and the residuals have zero correlation with the sample values of X, as will be shown later. For a geometrical interpretation, see Johnston and Dinardo (1997). The necessary conditions for a stationary value are

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \alpha} = -2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i) = -2\sum_{i=1}^{n} u_i = 0 \tag{3.7}$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta} = -2\sum_{i=1}^{n} X_i(Y_i - \alpha - \beta X_i) = -2\sum_{i=1}^{n} X_i u_i = 0 \tag{3.8}$$

called the *normal equations* for the linear regression of Y on X. Solving we obtain,

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{3.9}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = R\frac{s_y}{s_x}. \tag{3.10}$$

Figure 3.3 shows the true population regression and *one regression estimate* $Y = \hat{\alpha} + \hat{\beta}x$.

Suppose we consider a possible sequence of Y populations given in the top panel of figure 3.1. The assumptions above imply certain assumptions about the population. The bottom panel of figure 3.1 show the assumptions imposed to analyse the problem: the probability distributions $P(Y_i|X_i)$ have the same variance for all X_i; the means $\mathrm{E}(Y_i) = \mu_i$ lie on a straight line and that the random variables Y_i are uncorrelated. The assumption of normality has also been imposed on the residuals.

3.2.5 Matrix notation

Suppose there are now k explanatory variables: $X_1, X_2, ..., X_k$. So the data is now $(X_{i1}, X_{i2}, ..., X_{ik}, Y_i)$ for $i = 1, ...n$ observations. The model corresponding to

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + u_i.$$

could arise from the specification $\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\beta}\boldsymbol{X}$, written in matrix form, where

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ 1 & \vdots & \ddots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{pmatrix}.$$

Some notation and useful formulae on matrix algebra are given in the Appendix.

Classical assumptions

(1) $\boldsymbol{X}$ is non-stochastic and has full column rank

(2) $\boldsymbol{u}$ is a random vector with

- $\mathbb{E}(\boldsymbol{u}|\boldsymbol{X}) = 0$
- $\mathrm{var}(\boldsymbol{u}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$
- $\boldsymbol{u}|\boldsymbol{X}$ is normal

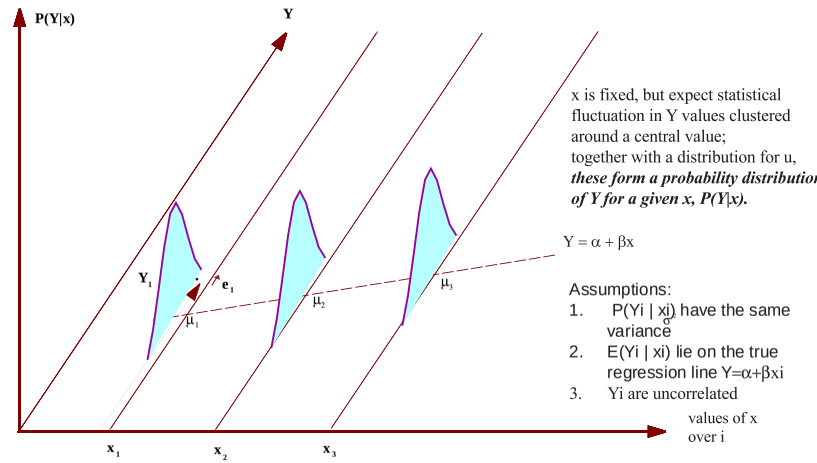


Figure 3.3 *True (population) regression and estimated (sample) regression..*

(3) $\rho(\boldsymbol{X}) = k$.
(4) $\text{var}(\boldsymbol{u}|\boldsymbol{X}) = \text{E}(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$
(5) $\boldsymbol{u}|\boldsymbol{X}$ is normal

The most common association with the above estimator is with the so called least squares principle, which determines an estimate $\hat{\beta}$ by finding the value of β which minimises the squared vertical distances between y and $X\beta$.

As before we wish to obtain the estimator $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squares of the residuals $e'e$, where $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$:

$$\text{Minimize } S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \text{ w.r.t. all the elements of } \boldsymbol{\beta}.$$

So the normal equation is now $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$, and so $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The form

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$

is an alternative form for the OLS estimator.[7]

3.2.5.1 Properties of the OLS estimators

$\hat{\boldsymbol{\beta}}$ is a random vector since it is a function of $\boldsymbol{y}$, and is:

(1) Linear
(2) Unbiased

[7] *An example*: the normal equations for the two-variable case.
Suppose we have $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, where the $\boldsymbol{X}$ matrix is:

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix},$$

so that,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}$$

and

$$\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n X_i y_i \end{pmatrix}$$

giving either
$n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^n X_i = \sum_{i=1}^n y_i$ or $\hat{\beta}_1 \sum_{i=1}^n X_i + \hat{\beta}_2 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i y_i$.
Note that the OLS equations for y on X in partitioned form, (see Appendix 3.5.5), are

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}$$

So

$$\begin{aligned} \hat{\beta}_1 &= \left\{(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1} - (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right\} \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{pmatrix} \\ &= (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y} \\ \hat{\beta}_2 &= (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}. \end{aligned}$$

If $\boldsymbol{y}$ is replaced by $\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{e}$ in $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$, then $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{e}) = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}'\boldsymbol{e}$ so that $\boldsymbol{X}'\boldsymbol{e} = 0$, implying that $\overline{e} = 0$; i.e. each regressor has zero sample correlation with the residuals, which implies $\hat{y}'\boldsymbol{e} = (\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{e} = \hat{\boldsymbol{\beta}}\,\boldsymbol{X}'\boldsymbol{e} = 0$.

(3) BLUE (Gauss Markov theorem) with variance $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
(4) Multivariate normal: $\hat{\boldsymbol{\beta}} \sim N\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\}$
(5) Maximum likelihood estimator.

3.3 Bias and variance

For $y = X\beta + \varepsilon$ and $E(\varepsilon) = 0$ while $Var(\varepsilon) = \sigma^2 I$, then

$$\widehat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon,$$

so $\widehat{\beta}$ is unbiased while $Var\left(\widehat{\beta}\right) = \sigma^2(X'X)^{-1}$.

3.3.1 MMSLE (Gauss-Markov)

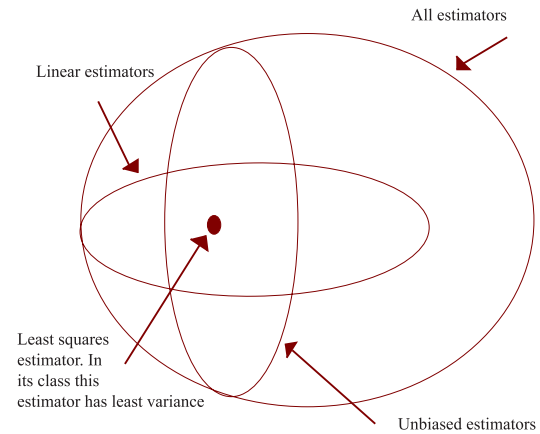


Figure 3.4 *The restricted class of estimators considered in the Gauss-Markov theorem.*

OLS estimators are *MMSLE* or *BLUE*: they have the minimum variance in the class of *L*inear *U*nbiased *E*stimators, see Figure 3.4.

Gauss Markov theorem

The least squares estimator is the *B*est *L*inear *U*nbiased *E*stimator of β.

The estimator is the *Best* in the sense that the var-cov matrix of any other estimator *exceeds* its var-cov matrix by a *positive semi-definite* matrix.

Proof

Let $\tilde{\beta} = \boldsymbol{A}'\boldsymbol{y}$, another linear unbiased estimator of $\tilde{\beta} = \boldsymbol{A}'\boldsymbol{X}\beta + \boldsymbol{A}'\boldsymbol{u}$. To ensure unbiasedness, we have

$$\begin{aligned} \mathrm{E}(\tilde{\beta}) &= \boldsymbol{A}'\boldsymbol{X}\beta + \boldsymbol{A}'\mathrm{E}(\boldsymbol{u}) \\ &= \boldsymbol{A}'\boldsymbol{X}\beta \end{aligned}$$

which implies $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}$. Then

$$\mathrm{var}(\tilde{\beta}) = \mathrm{E}(\boldsymbol{A}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{A}) = \sigma^2\boldsymbol{A}'\boldsymbol{A}.$$

Now let $\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and show that $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{A}'\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}$. We know that

$$\begin{aligned} \boldsymbol{A}'\boldsymbol{y} &= \{\boldsymbol{B} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\}'(\boldsymbol{X}\beta + \boldsymbol{u}) \\ &= (\boldsymbol{B}'\boldsymbol{X} + \boldsymbol{I})\beta + \{\boldsymbol{B}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{u} \end{aligned}$$

which is unbiased if and only if (iff) $\boldsymbol{B}'\boldsymbol{X} = 0$.

$$\begin{aligned} \mathrm{var}(\boldsymbol{A}'\boldsymbol{y}) &= \{\boldsymbol{B} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\}'\mathrm{E}(\boldsymbol{u}\boldsymbol{u}')\{\boldsymbol{B} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\} \\ &= \sigma^2\{\underbrace{\boldsymbol{B}'\boldsymbol{B}}_{\text{psd matrix}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\}. \end{aligned}$$

Hence $\mathrm{var}(\tilde{\beta}) \geq \mathrm{var}(\hat{\beta})$.

3.3.1.1 Regression as a decomposition

The OLS residuals are *orthogonal* to the explanatory variables:

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = 0.$$

So $\boldsymbol{y}$ can be written as $y = X\hat{\beta} + e$, so that

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \left\{\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right\} \\ &= \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$; both are symmetric idempotent matrices. Note that

(1) $\boldsymbol{M}\boldsymbol{X} = 0$
(2) $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}$
(3) $\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{u}$
(4) $\rho(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}) = n - k$

as $\mathrm{tr}(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\} = n - \mathrm{tr}\{\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\} = n - k$.

3.3.1.2 Decomposition of the sum of squares

Zero covariance between regressors and the residual underlie the decomposition. Decompose $\boldsymbol{y}$ into

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \underbrace{\boldsymbol{X}\hat{\beta}}_{\text{explained by regression}} + \underbrace{\boldsymbol{e}}_{\text{unexplained part}}$$

so that

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} = \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} + \boldsymbol{e}'\boldsymbol{e}.$$

However, since $\boldsymbol{y}'\boldsymbol{y} = \sum Y_t^2$, and we are normally interested in the *variation* of Y, (measured by sum of the squared deviations from the sample mean): $\sum(Y_t - \bar{Y})^2 = \sum Y_t^2 - n\bar{Y}^2$, we work with

$$\underbrace{(\boldsymbol{y}'\boldsymbol{y} - n\bar{Y}^2)}_{\text{TSS}} = \underbrace{(\hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} - n\bar{Y}^2)}_{\text{ESS}} + \underbrace{\boldsymbol{e}'\boldsymbol{e}}_{\text{RSS}}.$$

3.3.1.3 Goodness of fit

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

measures the proportion of the total variation in Y explained by the linear combination of the regressors. A related statistic is $\bar{R}^2$ which takes account of the number of regressors used in the equation and is defined as

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)} \\ &= \frac{1-k}{n-k} + \frac{n-1}{n-k}R^2. \end{aligned}$$

$\bar{R}^2$ is useful for comparing the fit of specifications that differ in the addition or deletion of explanatory variables. (Note that R^2 will never decrease with the addition of any variable. If the added variable is irrelevant in terms of explanatory power, ESS will remain constant; $\bar{R}^2$ may decrease with the addition of a variable with low explanatory power.

Other statistics used are the *Schwartz* criterion (SIC) or (BIC) and the *Akaike information* criterion (AIC) defined as

$$\begin{aligned} \text{SIC} &= \ln\frac{\boldsymbol{e}'\boldsymbol{e}}{n} + \frac{k}{n}\ln n \\ \text{AIC} &= \ln\frac{\boldsymbol{e}'\boldsymbol{e}}{n} + \frac{2k}{n}. \end{aligned}$$

Both look for specifications that reduce the RSS, but each criterion penalizes the addition of regressors.

3.3.1.4 Estimation of σ^2

Since $\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{u}$,

$$\begin{aligned} \mathrm{E}(\boldsymbol{e}'\boldsymbol{e}) &= \mathrm{E}(\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}) \\ &= \mathrm{E}\{\,\mathrm{tr}(\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u})\} \\ &= \mathrm{E}\{\,\mathrm{tr}(\boldsymbol{u}\boldsymbol{u}'\boldsymbol{M})\} \\ &= \sigma^2\,\mathrm{tr}(\boldsymbol{M}) \\ &= \sigma^2\,\mathrm{tr}(\boldsymbol{I}) - \sigma^2\,\mathrm{tr}\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\} \\ &= \sigma^2 n - \sigma^2\,\mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\} \\ &= \sigma^2(n-k). \end{aligned}$$

So, $\hat{\sigma}^2 = \boldsymbol{X}'\boldsymbol{X}/(n-k)$ defines an unbiased estimator of σ^2 which can be used for estimating the variance of $\hat{\beta}$, since $\mathrm{var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

A numerical example

Suppose

	X	Y	XY	X^2	$\hat{Y}$	e	Xe
	2	4	8	4	4.5	−0.5	−1
	3	7	21	9	6.25	0.75	2.25
	1	3	3	1	2.75	0.25	0.25
	5	9	45	25	9.75	−0.75	−3.75
	9	17	153	81	16.75	0.25	2.25
Sums	20	40	230	120	40	0	0

The normal equations give

$$\begin{aligned} 40 &= 5\hat{\alpha} + 20\hat{\beta} \\ 230 &= 20\hat{\alpha} + 120\hat{\beta} \end{aligned}$$

with solution $\hat{y} = 1 + 1.75X$. In devaiation form:

	x	y	xy	x^2	y^2	$\hat{y}$	e	xe
	-2	-4	8	4	16	-3.5	-0.5	1
	-1	-1	1	1	1	-1.75	0.75	-0.75
	-3	-5	15	9	25	-5.25	0.25	-0.75
	1	1	1	1	1	1.75	-0.75	-0.75
	5	9	45	25	81	8.75	0.25	1.25
Sums	0	0	70	40	124	0	0	0

$$\begin{aligned}\hat{\beta} &= \frac{\sum x_i y_i}{\sum x_i^2} = \frac{70}{40} = 1.75 \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} = 8 - 1.75(4) = 1\end{aligned}$$

and the ESS and RSS are

$$\begin{aligned}\text{ESS} &= \hat{\beta}\sum x_i y_i = 1.75(70) = 122.5 \\ \text{RSS} &= \text{TSS} - \text{ESS} = 124 - 122.5 = 1.5 \\ R^2 &= \frac{\text{ESS}}{\text{TSS}} = \frac{122.5}{124} = 0.9879.\end{aligned}$$

The estimate of σ^2 and the estimated variances of $\hat{\alpha}$ and $\hat{\beta}$ are given by

$$\begin{aligned}s^2 &= \text{RSS}/(n-2) = 1.5/3 = 0.5 \\ \widehat{\text{var}}(\hat{\beta}) &= s^2/\sum x_i^2 = 0.5/40 = 0.0125 \\ \widehat{\text{var}}(\hat{\alpha}) &= 0.5\left(\frac{1}{5} + \frac{16}{40}\right) = 0.3.\end{aligned}$$

The estimated standard errors of the regression coefficients are thus

$$\text{s.e.}(\hat{\alpha}) = \sqrt{0.3} = 0.5477 \quad \text{and} \quad \text{s.e.}(\hat{\beta}) = \sqrt{0.0125} = 0.1118.$$

A preselected critical value from the t distribution with 3 d.o.f. is $t_{0.025} = 3.182$. So a 95% Confidence Interval for α is

$$-1 \pm 3.182(0.5477), \text{ i.e. } (-0.74, 2.74)$$

and β is:

$$1.75 \pm 3.182(0.1118), \text{ i.e. } (1.39, 2.11).$$

A 95% C.I. for σ^2 is

$$\frac{1.5}{9.35} \text{ to } \frac{1.5}{0.216},$$

i.e. $(0.16, 6.34)$, since $\chi^2_{0.025}(3) = 0.216$, $\chi^2_{0.975}(3) = 9.35$ and $\sum e_i^2 = 1.5$.

3.4 Central limit theorems for regression*

If the data is not Gaussian, then the exact distributional behaviour of $\widehat{\beta}$ is not immediately known. It is thus interesting to study large sample approximations which can give us rough but general results. Unfortunately when we move away from iid models we cannot use such simple limit theorems as the Lindeburg-Levy CLT we have exploited so far in these lectures. Instead we need to refer to more general results which are slightly more intricate. I refer you to White (1984) and White (1994).

3.5 Appendix

Some notation and useful formulae on matrix algebra, taken from lecture notes by P.M. Robinson,[8] can be found below. As it is primarily intended to cover matrix formulae relating to standard economics problems, the list is not exhaustive and Magnus and Neudecker (1988) and Lütkepohl (1996) should be consulted for further information. A comprehensive review of matrix algebra can also be found in Johnston and Dinardo (1997, pp.455-484). Since parts should have been covered in the maths crash course, only a brief exposition on some of the key formulae will be given here. Additional references are Harvey (1981, pp. 359-361) and Magnus and Neudecker (1988).

3.5.1 Basic definitions and axioms

(1) $A = A_{p\times q} = (a_{ij})$ is a $p \times q$ matrix, (p rows, q columns) with a_{ij} as its (real) element in the i'th row, j'th column, where $i = 1, ..., p$ and $j = 1, ..., q$.
(2) $A' = (a_{ji})$ is the *transpose* of A.
(3) $cA = (c_{aij})$, if c is scalar.
(4) $A + B = (a_{ij} + b_{ij})$ if B is also $p \times q$.
(5) $AB = (\sum_{k=1}^{q} a_{ik}b_{kj})$, if B has q rows.
(6) $(AB)C = A(BC) = ABC$ *associative property*. In general, matrices do not *commute*, $AB \neq BA$.
(7) $(AB)' = B'A'$.
(8) $\rho(A)$ is the *rank* of A, i.e. the number of linearly independent rows (columns).
(9) If $\rho(A)$ is equal to the number of columns (rows), we say A is of *full column (row) rank*.
(10) $\rho(AB) \leq \min\{\rho(A), \rho(B)\}$, i.e. rank of a matrix cannot be increased by multiplying it with another matrix.
(11) If $a_{ij} = 0$, for all i, j, then $A = 0$.

3.5.2 Square matrix, $A_{p\times p}$

(1) $\text{tr}(A) = \sum_{i=1}^{p} a_{ii}$, the *trace* of A.
(2) If $a_{ij} = 0$ for all $i \neq j$, write $A = \text{diag}(a_{11}, ..., a_{pp})$.
(3) $\text{tr}(BC) = \text{tr}(CB)$ if BC, (and therefore CB), are square.
(4) $\text{tr}(A') = \text{tr}(A)$.
(5) The *determinant* of A is

$$|A| = \sum_{i=1}^{p} a_{ij}(-1)^{i+j}A_{ij}$$

where A_{ij} is the (i, j)'th *minor*, i.e. the determinant of the $(p-1) \times (p-1)$ matrix formed by deleting the i'th row and j'th column of A.
(6) $|A'| = |A|$.
(7) $|AB| = |A||B|$ if B is also $p \times p$.
(8) $|cA| = c^p|A|$ if c is scalar.
(9) If $|A| = 0$, then A is *singular* and so $\rho(A) < p$.
(10) If $|A| \neq 0$, A is said to be *nonsingular*. Then $\rho(A) = p$ and the inverse of A, denoted A^{-1} exists, such that $AA^{-1} = I_p$, where $I_p = \text{diag}(1, ..., 1)$ is the $p \times p$ *identity* matrix.
(11) The elements of A^{-1} are continuous in elements of A, except at $|A| = 0$.
(12) $(AB)^{-1} = B^{-1}A^{-1}$ if $|A| \neq 0$ and $|B| \neq 0$.
(13) $|A^{-1}| = |A|^{-1}$ if $|A| \neq 0$.
(14) $\rho(AB) = \rho(A)$ if $|B| \neq 0$; (also true for rectangular A).
(15) The zeros $\lambda_1, ..., \lambda_p$ of the polynomial

$$|A - \lambda I_p| = \sum_{i=1}^{p} \lambda^p \quad \text{(where } \lambda \text{ is scalar}$$

[8] Results are taken from P.M. Robinson's (1996) lecture notes in Advanced Econometric Theory, MSc. Econometrics and Mathematical Economics, London School of Economics.

are called the *eigenvalues* of A.

(16) The zeros of $|A - B\lambda|$ are the eigenvalues of $B^{-1}A$, (and also of AB^{-1}), when $|B| \neq 0$.

(17) Since $A - \lambda_i I_p$ is singular, for λ_i an eigenvalue, there exists a $p \times 1$ vector x_i, called an *eigenvector* of A, such that $(A - \lambda_i I_p)x_i = 0$

(18) If $A'A = I_p$, then A is said to be *orthogonal* and $A^{-1} = A'$.

(19) If $A = A'$, then A is said to be *symmetric*.

(20) $\text{tr}(A) = \sum_{i=1}^{p} \lambda_i$.

(21) $|A| = \prod_{i=1}^{p} \lambda_i$.

(22) If $\lambda_i \geq 0$, for all i, then A is said to be *non-negative definite*, written $A \geq 0$. (This does not mean that all elements of A are non-negative).

(23) If $\lambda_i > 0$, for all i, then A is said to be *positive-definite* and written $A > 0$.

3.5.3 Traces and idempotent matrices

The trace is the sum of the elements on the principle diagonal,

$$\text{tr}(\boldsymbol{A}) = \sum_i a_{ii}.$$

If the matrix $\boldsymbol{A}$ is of order $(m \times n)$ and $\boldsymbol{B}$ is of order $(n \times m)$ then $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are both square matrices,

$$\text{tr}(\boldsymbol{BA}) = \text{tr}(\boldsymbol{AB})$$

and

$$\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{CAB}) = \text{tr}(\boldsymbol{BCA})$$

provided the products exist as square matrices.

An *Idempotent* matrix is one that however many times it is multiplied by itself, remains in its original form. So

$$\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{A}^3 = \boldsymbol{...}$$

3.5.4 Partitioned matrices

Suppose

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

and

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}$$

then

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \boldsymbol{A}_{12} + \boldsymbol{B}_{12} \\ \boldsymbol{A}_{21} + \boldsymbol{B}_{21} & \boldsymbol{A}_{22} + \boldsymbol{B}_{22} \end{pmatrix}$$

provided $\boldsymbol{A}$ and $\boldsymbol{B}$ are of the same dimension and each pair $\boldsymbol{A}_{ij},\ \boldsymbol{B}_{ij}$ are of the same order.

$$\boldsymbol{AB} = \begin{pmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{pmatrix}$$

where columns of $\boldsymbol{A}$ are equal to the rows of $\boldsymbol{B}$.

3.5.5 Inverting partitioned matrices

If

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

then, either

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{B}_{11} & -\boldsymbol{B}_{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11}\boldsymbol{A}_{21}\boldsymbol{A}_{22}^{-1} \end{pmatrix}$$

where $\boldsymbol{B}_{11} = (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}$, and $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are square and nonsingular; alternatively

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ -\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{B}_{22} \end{pmatrix}$$

where $\boldsymbol{B}_{22} = (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}$.

An important special case is when a data matrix is partitioned as

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix},$$

then

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}.$$

¿From the formulae above

$$\boldsymbol{B}_{11} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 - \boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1 \end{pmatrix}^{-1} = (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}$$

with

$$\boldsymbol{M}_2 = \boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'.$$

Similarly,

$$\boldsymbol{B}_{22} = (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}$$

with

$$\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'.$$

Note that the $\boldsymbol{M}_1$ are symmetric and idempotent. These can be used in evaluating the OLS equations for $\boldsymbol{y}$ on $\boldsymbol{X}$ in partitioned form.

3.5.6 Definite and semidefinite quadratic forms

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$

where the matrix of the quadratic form is assumed to be positive. The quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ and associated symmetric matrix $\boldsymbol{A}$ are said to be

positive definite	if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ holds for any $\boldsymbol{x} \neq 0$
positive semidefinite	if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$ holds for any $\boldsymbol{x}$
negative definite	if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} < 0$ holds for any $\boldsymbol{x} \neq 0$
negative semidefinite	if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \leq 0$ holds for any $\boldsymbol{x}$
indefinite	if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} < 0$ and $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} > 0$ for two vectors $\boldsymbol{x}$and $\boldsymbol{y}$, where $\boldsymbol{x} \neq \boldsymbol{y}$.

3.5.7 Distribution of quadratic forms with normal variates

Let $\boldsymbol{x} = (X_1, \boldsymbol{...}, X_k)'$, so that $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ is a quadratic form. If $\boldsymbol{A}$ is *idempotent* of rank r and if the elements of $\boldsymbol{x}$ are independent standardized normal variates, the quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ is distributed as χ_r^2.

3.5.8 Kronecker products and vec notation

(1) The *Kronecker product* of (possibly rectangular) matrices $A_{p\times q}$ and $B_{r\times s}$ is the $pr \times qs$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{pmatrix}.$$

(2) $(A + B) \otimes C = A \otimes C + B \otimes C$.
(3) $(A \otimes B) \otimes C = A \otimes (B \otimes C) = A \otimes B \otimes C$.
(4) $(A \otimes B)(C \otimes D) = AC \otimes BD$ if AC and BD are defined.
(5) $(A \otimes B)' = A' \otimes B'$.
(6) tr $(A \otimes B) =$ {tr (A)}{tr (B)} if $p = q$ and $r = s$.
(7) $|A \otimes B| = |A|^r|B|^p$ if $p = q$ and $r = s$.
(8) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ if $|A| \neq 0$ and $|B| \neq 0$.
(9) $\rho(A \otimes B) = \rho(A)\rho(B)$.
(10) Denoting by a_j the j'th column of A,

$$\text{vec}(A) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{pmatrix}.$$

(11) vec$(ABC) = (C' \otimes A)$ vec(B).
(12) tr $(ABCD) =$ vec$'(C)(D \otimes B')$ vec(A').

3.5.9 Matrix differentiation

(1) $\frac{\partial}{\partial x}x'Ay = Ay$.
(2) $\frac{\partial}{\partial x}x'Ax = (A + A')x$.
(3) Let A be a function of a scalar θ. Then
(a)

$$\frac{\partial \ln|A|}{\partial\theta} = \text{tr}\left\{(A')^{-1}\frac{\partial A}{\partial\theta}\right\}.$$

(b)

$$\frac{\partial A^{-1}}{\partial\theta} = A^{-1}\frac{\partial A}{\partial\theta}A^{-1}.$$

(c)

$$\frac{\partial^2 \ln|A|}{\partial\theta^2} = \text{tr}\left\{(A')^{-1}\frac{\partial^2 A}{\partial\theta^2} - (A')^{-1}\frac{\partial A'}{\partial\theta}(A')^{-1}\frac{\partial A}{\partial\theta}\right\}.$$

(d)

$$\frac{\partial^2 A^{-1}}{\partial\theta^2} = A^{-1}\left(\frac{\partial A}{\partial\theta}A^{-1}\frac{\partial A}{\partial\theta} - \frac{\partial^2 A}{\partial\theta^2}\right)A^{-1}.$$

(4) Let A be a function of a vector θ. Then
(a)

$$\frac{\partial \ln|A|}{\partial\theta} = \left\{\frac{\partial \text{vec}'(A)}{\partial\theta}\right\}\text{vec}\left(A^{-1}\right).$$

(b)

$$\frac{\partial \text{vec'}(A^{-1})}{\partial\theta} = -\left\{\frac{\partial \text{vec'}(A)}{\partial\theta}\right\}\left\{A^{-1} \otimes (A')^{-1})\right\}.$$

(5) Let A be a function of a vector θ, but B and C be functionally independent of θ. Then
(a)

$$\frac{\partial \text{tr}\,(AB)}{\partial\theta} = \left\{\frac{\partial \text{vec}'(A)}{\partial\theta}\right\}\text{vec}\,(B').$$

(b)

$$\frac{\partial\,\text{tr}\,(A'BAC)}{\partial\theta} = 2\frac{\partial \text{vec}\,'(A)}{\partial\theta}(C \otimes B')\,\text{vec}\,(A).$$

(c)

$$\frac{\partial^2\,\text{tr}\,(A'BAC)}{\partial\theta\partial\theta'} = 2\frac{\partial\,\text{vec}\,'(A)}{\partial\theta}(C \otimes B')\frac{\partial\,\text{vec}\,(A)}{\partial\theta'} + M,$$

where M has (i, j)'th element

$$2\,\text{tr}\,\left(A'B\frac{\partial^2 A}{\partial\theta_i\partial\theta_j}C\right).$$

For a vector $\boldsymbol{A}$, $(n \times 1)$ and vector $\boldsymbol{x}$ also $(n \times 1)$ such that $\boldsymbol{A}'\boldsymbol{x} = \sum A_i x_i$,

$$\frac{\partial(\boldsymbol{A}'\boldsymbol{x})}{\partial\boldsymbol{x}} = \boldsymbol{A}$$

and

$$\frac{\partial(\boldsymbol{A}'\boldsymbol{x})}{\partial\boldsymbol{x}'} = \boldsymbol{a}'.$$

For the quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is a matrix:

$$\begin{aligned} \frac{\partial('\boldsymbol{A}\boldsymbol{x})}{\partial\boldsymbol{x}} &= (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x} \\ \frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial\boldsymbol{x}\partial\boldsymbol{x}'} &= \boldsymbol{A} + \boldsymbol{A}' \\ \frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial\boldsymbol{A}} &= \boldsymbol{x}\boldsymbol{x}' \end{aligned}$$

3.6 Bibliography

- Gourieroux, C. and Monfort, A. (1995) *Statistics and Econometric Models: General Concepts, Estimation, Prediction and Algorithms, (Themes in modern econometrics) Cambridge: Cambridge University Press. Volume I.*
- *Harvey, A. C. (1981) The Econometric Analysis of Time Series, Philip Allan, New York, Second Edition.*
- *Hendry, D. F. (1995) Dynamic Econometrics, Oxford University Press, First Edition.*
- *Johnston, J. and Dinardo, J. (1997) Econometric methods, McGraw-Hill, Fourth Edition.*
- *Lütkepohl, H. (1996) Handbook of Matrices, John Wiley and Sons, Chichester.*
- *Magnus, J. R. and Neudecker, H. (1988) Matrix Differential Calculus with Applications in Statistics and Econometrics, John Wiley & Sons, Chichester.*
- *White, H. (1984) Asymptotic Theory for Econometricians, Academic Press.*
- *White, H, (1994) Estimation, Inference and Specification Analysis, Cambridge: Cambridge University Press, Econometric Society Monographs.*