CONSTRAINED ESTIMATION

Phoebus J. Dhrymes[1]

June 2006

1.1 Introduction

1.1.1 Preliminaries

In a previous chapter we have derived the uniformly most powerful test (UMP) for testing the hypothesis
$H_0 : \theta = 0$, as against the alternative
$H_1 : \theta > 0$,
where θ is the parameter of an appropriate probability measure characteristic of the data generating process (DGP). In this chapter we shall consider the problem of estimating the unknown parameter(s) subject to (inequality) constraints and derive appropriate test statistics. Just to give an example, suppose

$$X = \{x_1, x_2, \ldots, x_n\},$$

where X is a sequence and the x_i are (random elements, specifically vectors) defined on the probability space (Ω , $\mathcal{A}$, $\mathcal{P}$). If the $x's$ are independent identically distributed (i.i.d.) we may obtain, see Dhrymes (1989), the probability distribution induced by them, say P , which depends on (the true value of) an underlying parameter $\theta \in \Theta$, where Θ is the space of admissible parameters. If θ represents the mean, a typical problem may be to estimate it by minimizing the function in Eq. (1.1) below

$$S = (x - \theta)'(x - \theta), \quad \text{subject to } \theta \geq 0, \tag{1.1}$$

where x refers to the vector of the realization of X .

In the context of the general linear model (GLM), as exhibited in Eq. (1.2) below, let X be the $T \times m$ matrix of observations on the (m) explanatory variables and let u be a T-element column vector containing the realizations of an i.i.d. variable u_i with mean zero and variance σ^2 . Thus, consider

$$y = X\beta + u, \quad \text{where now } \theta = (X\beta, \sigma^2)\,. \tag{1.2}$$

The vector y has mean $X\beta$ and covariance matrix $\sigma^2 I_T$; the analog of θ in Eq. (1.1), for the i^{th} realization, is now $(x_{i\cdot}\beta, \sigma^2)$ which is more complex than that in Eq. (1.1) because of the introduction of the (m) explanatory variables $x_{i\cdot}$.

A not uncommon problem in the context of the GLM is to estimate the parameters of Eq. (1.2) **subject to the constraint**

$$R\beta = r \tag{1.3}$$

where R, r are, respectively, a known matrix and vector; R is of dimension $k \times m$, $k < m$ and of **full row rank**, i.e. there are no **redundant constraints**.

An important variant of this problem is to state the constraints as $R\beta \geq r$, or $R\beta \geq 0$, under the same conditions as above. This problem which has not received sufficient attention in the literature, shall be discussed extensively below.

1.1.2 Review of Estimation subject to $R\beta = r$

Traditionally, this problem is solved by means of Lagrange Multipliers, and the solution is obtained, e.g. in Dhrymes (1978), p. 52ff., as

$$\tilde{\beta} = \hat{\beta} + A(r - R\hat{\beta}), \quad A = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}, \quad \hat{\beta} = (X'X)^{-1}X'y, \tag{1.4}$$

so that the **restricted** estimator $\tilde{\beta}$ consists of the **unrestricted** least squares estimator (under H_1) plus a correction factor that depends on the departure of the (unrestricted) least squares estimator from the constraints, i.e. $r - R\hat{\beta}$.[2] The validity of the constraints can be tested by means of the Likelihood Ratio (LR) test statistic; in the case of (or *mutatis mutandis* asymptotic) normality, the LR is given by

$$\lambda = \frac{\max_{\theta|H_0} L(\theta)}{\max_{\theta|H_1} L(\theta)} = \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{-(n/2)}, \tag{1.5}$$

where $L(\theta)$ is the **joint likelihood function of the observations**, $\theta = (\beta', \sigma^2)'$ and $\tilde{\sigma}^2, \hat{\sigma}^2$ are the estimators of the variance, σ^2, respectively, under the null ($H_0 : R\beta = r$) and the alternative ($H_1 : R\beta \neq r$). A "large" value

[2] Traditionally in this context $H_0 : R\beta = r$, and the alternative is $H_1 : R\beta \neq r$, i.e. β is unrestricted.

of λ is considered evidence[3] in favor of the null, so that, for example, if the difference $\tilde{\sigma}^2 - \hat{\sigma}^2$ is sufficiently large it will lead to the rejection of H_0.

We observe that, under the null, we obtain the estimator

$$\tilde{\sigma}^2 = \frac{1}{T}(y - X\tilde{\beta})'(y - X\tilde{\beta}) = \frac{1}{T}(y - X\hat{\beta})'(y - X\hat{\beta}) + \frac{1}{T}u'Bu, \tag{1.6}$$

where

$$B = X(X'X)^{-1}R'DR(X'X)^{-1}X', \quad D = [R(X'X)^{-1}R']^{-1}. \tag{1.7}$$

Thus, we may rewrite the LR as

$$\lambda = \left(1 + \frac{u'Bu}{u'Mu}\right)^{-(T/2)}, \quad \hat{\sigma}^2 = \frac{1}{T}u'Mu, \quad M = I_T - X(X'X)^{-1}X. \tag{1.8}$$

The distribution of the LR above is rather difficult to obtain; however, a one to one transformation of it yields a statistic with an easily determined distribution. In particular

$$\lambda^{-(2/T)} = 1 + \frac{u'Bu}{u'Mu}, \tag{1.9}$$

so that[4]

$$-2\ln\lambda \approx \frac{u'Bu}{u'Mu/T} \sim \frac{k(T-m)}{T}F_{k,T-m}. \tag{1.10}$$

When the error process is not normal, but satisfies conditions so that an appropriate central limit theorem (CLT) holds, then employing the same method, i.e. pretending that the likelihood function above is the appropriate function to maximize we find, asymptotically, that

$$\frac{1}{\sqrt{T}}[-2\ln\lambda] \approx \frac{u'Bu/\sqrt{T}}{u'Mu/T} \xrightarrow{\mathrm{d}} \chi^2_k. \tag{1.11}$$

Evidently, the LR method examined above can also be carried out without reference to normality or likelihoods. This is done by the device employed in elementary courses, by looking at the restricted and unrestricted sum of squared residuals and making a judgment based on the difference between them.[5] Precisely, imposing the restrictions we shall obtain a certain sum of squares, say, $\tilde{u}'\tilde{u}$, while not imposing the restrictions we obtain the unrestricted sum of squared

[3] Note that by construction the LR as we have defined it above is **less than one**.

[4] When there are no restrictions, a multiple of the leftmost member of the entity in Eq. (1.10) is the statistic printed in computer programs for the test of the "significance" of the regression, with $k = 1$, corresponding to the constant term.

[5] Somewhat inappropriately, this is called in the econometrics literature the Chow test.

residuals, say $\hat{u}'\hat{u}$. Considering the difference, normalized by the unrestricted sum of squared residuals, we find

$$d = \frac{\tilde{u}'\tilde{u} - \hat{u}'\hat{u}}{\hat{u}'\hat{u}} = \lambda^{-(2/T)} - 1, \tag{1.12}$$

so that the Chow test statistic is essentially the LR test statistic applied in situations of pseudo maximum likelihood estimation. Note, in addition, that this approach, in principle, requires us to obtain **both the restricted and the unrestricted estimator**.

On the other hand, the very method of obtaining the restricted estimator affords us the opportunity to devise another test based entirely on the restricted estimator. This is the so called Lagrange Multiplier (LM) test. Formulating the problem as a minimization under constraint (or, if doing ML, maximization under constraint), we obtain the Lagrange multiplier estimator

$$\hat{\lambda} = D(R\hat{\beta} - r) = D(R\beta - r) + DR(X'X)^{-1}X'u, \tag{1.13}$$

which shows the estimated LM to be a random element with mean $D(R\beta - r)$ and covariance matrix $\sigma^2 D$. If normality is assumed, we can define, under the null, the test statistic

$$T_{LM} = \hat{\lambda}'(\sigma^2 D)^{-1}\hat{\lambda} \sim \chi^2_k, \tag{1.14}$$

provided σ^2 is known; if not, T_{LM} **is not a statistic** and we need to modify it by substituting for σ^2 therein some consistent estimator, say σ^{*2}. Thus, if we take $\sigma^{*2} = \hat{\sigma}^2$, then this statistic is identical to that produced earlier in the context the LR procedure, because under the null

$$\hat{\lambda}'(\hat{\sigma}^2 D)^{-1}\hat{\lambda} = \frac{T}{u'Mu}u'Bu \sim \frac{Tk}{T-m}F_{k,T-m}. \tag{1.15}$$

Under these circumstances the two statistics are **numerically** identical. [6]

Another test of the restrictions is the **conformity test**; this is carried our by first obtaining the **unrestricted** estimator, $\hat{\beta} = (X'X)^{-1}X'y$ and thereafter asking whether it conforms to the constraints. The test statistic is then based on $R\hat{\beta} - r$ which, under normality, has the distribution, under the null, $N(0, \sigma^2 D^{-1})$, and is given by

$$T_c = \left(\frac{1}{u'Mu}\right) u'X(X'X)^{-1}R'DR(X'X)^{-1}X'u = \frac{u'Bu}{u'Mu}. \tag{1.16}$$

[6] A number of papers in the 70s sought to establish a magnitude hierarchy between LR, LM and conformity tests. For this particular case the claimed result is brought about by using the restricted estimator of σ^2. However, if we do so under normality this estimator will not be **independent** of the numerator of the fraction and, under the alternative, the statistic will not be **noncentral** F, unless we rely **exclusively** on asymptotics.

Thus, if done properly, all three tests, LR, LM and Conformity are numerically identical.

The preceding is the standard, or the prevailing, approach to estimation subject to (linear equality) parameter restrictions. The method, however, becomes quite involved and indeed inoperable, if an explicit solution for the estimator is not available. Thus, in the case where the restrictions appear in the form of inequalities the preceding is inapplicable.

It is then desirable, before we proceed, that we should employ some alternative procedure in dealing with the problem posed above. This alternative is based, in the first instance, on the projection theorem dealt with in the appendix. The problem at hand may be formulated as: find the projection of the vector y on the linear (vector) space spanned by the (linearly independent) columns of X, say $\mathcal{M}$. This is given by the vector $\tilde{y}$, which is the vector in $\mathcal{M}$ closest in distance to y. By the projection theorem, $y - \tilde{y}$ is **orthogonal** to $\tilde{y}$.

A projection, P, is an idempotent operator which carries the vector $y \in R^T$ to the vector space above, which is of course the column space of X, and may be denoted by $\tilde{y} = Py \in \mathcal{M}$ or more succinctly $\tilde{y} = P(y|\mathcal{M}$. Again by the projection theorem $y - \tilde{y}$ lies in the **orthogonal complement** of $\mathcal{M}$, $\mathcal{M}^{\perp}$, and is obtained through the idempotent operator $I - P$; thus we have

$$\tilde{y} = Py \in \mathcal{M}, \quad y - \tilde{y} = \tilde{u} = (I-P)y \in \mathcal{M}^{\perp}, \quad \text{so that} \quad \tilde{y}'(y-\tilde{y}) = 0, \quad \tilde{y}'y = \tilde{y}'\tilde{y}. \tag{1.17}$$

It may be shown that the projection operator P can be represented, in this case (a linear space), by the (idempotent) matrix $X(X'X)^{-1}X'$ and the operator $I - P$ can be represented by the (idempotent) matrix $I - X(X'X)^{-1}X'$. These matrices are **symmetric idempotent** and mutually orthogonal. This therefore demonstrates that every vector $y \in R^T$ can be written, uniquely,

$$y = \tilde{y} + \tilde{u}, \quad \tilde{y}'\tilde{u} = 0, \quad \tilde{y} \in \mathcal{M}, \ \tilde{u} \in \mathcal{M}^{\perp} \text{ or } \tilde{u} = y - P(y|\mathcal{M}) = y - \tilde{y}. \tag{1.18}$$

We may apply the projection theorem to the constrained parameter situation.[7] In formulating this problem, we note that the presence of parameter constraints restricts the space onto which the projection is made. If we establish the restricted space, then evidently precisely the same procedure as above can be employed. Initially we shall consider the case $r = 0$. Now note that

[7] I have not seen anywhere the derivation to follow. But given its rather simple structure I do not claim originality. I assume that it was done somewhere else in the past–only I do not know it.

since R is $k \times m$ of rank k, it contains a nonsingular sub-matrix of order k. Without loss of generality, let this consist of the first k columns, so that we can write

$$R = (R_1, R_2), \quad R_1 \text{ being } k \times k, \quad \text{and nonsingular.} \tag{1.19}$$

We then have

$$\beta_{(1)} = -R_1^{-1} R_2 \beta_{(2)}, \quad \beta = \begin{bmatrix} -R_1^{-1} R_2 \\ I_{m-k} \end{bmatrix} \beta_{(2)}, \tag{1.20}$$

which evidently satisfies the condition $R\beta = 0$. It follows therefore that

$$X\beta = X \begin{bmatrix} -R_1^{-1} R_2 \\ I_{m-k} \end{bmatrix} \beta_{(2)} = (X_1, X_2) \begin{pmatrix} -R_1^{-1} R_2 \\ I_{m-k} \end{pmatrix} \beta_{(2)} = X^* \beta_{(2)}. \tag{1.21}$$

Thus, applying the projection theorem, we are seeking the projection of the vector y on the space spanned by

$$X^* = X_2 - X_1 R_1^{-1} R_2, \tag{1.22}$$

where X^* is a matrix of order $T \times (m-k)$ and $\text{rank}(X^*) = m - k$. By the discussion earlier

$$\tilde{y} = X^* (X^{*\prime} X^*)^{-1} X^{*\prime} y, \tag{1.23}$$

so that the implied estimator of $\beta_{(2)}$ is $(X^{*\prime} X^*)^{-1} X^{*\prime} y$. It follows then that the appropriately restricted estimator of β is

$$\tilde{\beta} = \begin{bmatrix} -R_1^{-1} R_2 \\ I_{m-k} \end{bmatrix} (X^{*\prime} X^*)^{-1} X^{*\prime} y, \tag{1.24}$$

which obeys

$$R\tilde{\beta} = 0,$$

as required. Note further that

$$\tilde{u} = y - \tilde{y}, \quad \tilde{u}'\tilde{y} = 0, \quad \tilde{u} = [I - X^*(X^{*\prime} X^*)^{-1} X^{*\prime}] y, \tag{1.25}$$

so that everything we obtained directly by the method of Lagrange multipliers is also duplicated if we apply the projection theorem to the case of parameter constraints.

Anticipating further work we also note that if we proceeded exclusively by means of Lagrange multipliers in this problem we have obtained the projection of y on the space spanned by X subject to the restriction $R\beta = 0$, as

$$\tilde{y} = X\tilde{\beta} = My, \quad M = M_1 - M_2, \; M_1 = X(X'X)^{-1} X'$$

$$M_2 = X(X'X)^{-1} R' [R(X'XR']^{-1} R(X'X)^{-1} X'. \tag{1.26}$$

We note that both M_1 and M_2 are **symmetric idempotent** matrices such that $M_1 M_2 = M_2$, which shows that M is also **symmetric idempotent**. For future reference we also note that

$$\text{rank}(M) = \text{tr}M_1 - \text{tr}M_2 = m - k. \tag{1.27}$$

The problem is more complicated if $r \neq 0$, because when this is so the construction of Eq. (1.20) is replaced by

$$\beta_{(1)} = R_1^{-1}[r - R_2\beta_{(2)}], \quad \beta = \begin{pmatrix} R_1^{-1}r \\ 0 \end{pmatrix} + \begin{bmatrix} -R_1^{-1}R_2 \\ I_{m-k} \end{bmatrix} \beta_{(2)}. \tag{1.28}$$

The problem is how we treat the extra term

$$\begin{pmatrix} R_1^{-1}r \\ 0 \end{pmatrix},$$

which appears in Eq. (1.27). We may proceed as follows; write

$$y = X\beta + u = X\left[\begin{pmatrix} R_1^{-1}r \\ 0 \end{pmatrix} + \begin{bmatrix} -R_1^{-1}R_2 \\ I_{m-k} \end{bmatrix} \beta_{(2)}\right], \quad \text{or} \quad y^* = X^*\beta_{(2)} + u, \tag{1.29}$$

where

$$y^* = y - X_1 R_1^{-1} r, \quad X^* = X_2 - X_1 R_1^{-1} R_2 \quad \text{as in Eq. (1.22);} \tag{1.30}$$

then following the same procedure as before, we obtain

$$\tilde{y}^* = N^* y^*, \quad \tilde{u}^* = [I_T - N^*] y^*, \quad N^* = X^*(X^{*\prime}X^*)^{-1}X^{*\prime}. \tag{1.31}$$

It is evident that $\tilde{u}^*$ and $\tilde{y}^*$ are mutually orthogonal. At this stage it may be useful to ask whether we can define an analogous relationship between $\tilde{y} = X\tilde{\beta}$ and $\tilde{u} = y - \tilde{y}$. We shall do so for the case $r = 0$; for the general case $r \neq 0$ the situation becomes considerably more complex and will not be pursued here. Nonetheless we shall give a procedure which will, if followed, will lead to the appropriate conclusion. Using Eq. (1.4) we may write, for the case $r = 0$,

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'D\hat{\beta}, \quad D = [R(X'X)^{-1}R']^{-1}, \quad \hat{\beta} = (X'X)^{-1}X'y,$$

so that

$$\tilde{y} = X\tilde{\beta} = (M_1 - M_2)y$$

$$\tilde{u} = y - \tilde{y} = (I_T - M_1 + M_2)y; \quad \text{thus}$$

$$(y - X\tilde{\beta})'X\tilde{\beta} = y'[I_T - M_1 + M_2][M_1 - M_2]y = 0,$$

which produces the desired result. To produce an analogous result when $r \neq 0$, we resort to a device, to be used at a later stage as well, of writing the objective function

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta})'(y - X\hat{\beta} + (\hat{\beta} - \beta)X'X(\hat{\beta} - \beta),$$

and noting that the estimator constrained by $R\beta = r$ can be obtained from the second term of the right member above. The underlying model may then be framed: estimate the parameters of the regression model $S_2\hat{\beta} = S_2\beta + S_2(X'X)^{-1}X'u$, where S_2 is a non-singular (could also be lower or upper triangular) decomposition of $X'X$. For notational simplicity put $w == S_2(X'X)^{-1}X'u$ and note that $\mathrm{Cov}(w) = \sigma^2 I_m$. To get rid of r in the constraint equations we make the transformation

$$\alpha = S_2\beta + a, \quad a \text{ to be determined.}$$

The transformed system then becomes

$$\hat{\alpha} = \alpha + w, \quad \text{subject to} \quad RS_2^{-1}(\alpha - a) = r.$$

Taking $a = -S_2R'(RR')^{-1}r$, the system may be finally written as

$$\hat{\alpha} = \alpha + w, \quad \text{subject to} \quad RS_2^{-1}\alpha = 0,$$

and the issue of orthogonaliity may be discussed from this perspective.

If

$$u \sim N(0, \sigma^2 I_T)$$

then

$$\frac{1}{\sigma^2}\tilde{u}^{*\prime}\tilde{u}^* = \frac{1}{\sigma^2}u'[(I_T - X^*(X^{*\prime}X^*)^{-1}X^{*\prime}]u \sim \chi^2_{T-m-k}. \tag{1.32}$$

We further note that the least squares procedure which projects on $\mathcal{M}$, the column space of X, yields

$$\hat{u} = [I - X(X'X)^{-1}X']u, \tag{1.33}$$

so that

$$\frac{1}{\sigma^2}\hat{u}'\hat{u} = \frac{1}{\sigma^2}u'[I_T - X(X'X)^{-1}X']u \sim \chi^2_{T-m}. \tag{1.34}$$

Thus, the transformed LRT statistic of Eq. (1.12), obtained earlier, is equal to

$$\frac{\tilde{u}'\tilde{u} - \hat{u}'\hat{u}}{\hat{u}'\hat{u}} = \frac{u'[X(X'X)^{-1}X' - X^*(X^{*\prime}X^*)^{-1}X^{*\prime}]u}{u'[I_T - X(X'X)^{-1}X']u}. \tag{1.35}$$

Since the matrix in the numerator is idempotent, and orthogonal to the one in the denominator, we conclude, see Dhrymes (1978), pp. 41-43, that the numerator and denominator entities are mutually independent (central) χ^2 entities, the numerator with k degrees of freedom and the denominator with $T-m$ degrees of freedom; hence the ratio is proportional to a (central) $F_{k,T-m}$ variable. This is exactly the result obtained in the discussion surrounding Eq. (1.12), although the emphasis there was on asymptotics.

1.2 Maximum Likelihood when the true parameter is a point on the boundary of the admissible parameter space

All discussion above assumed, implicitly or explicitly, that the true parameter point, say θ_0, is an **interior** point of the admissible parameter space, Θ, i.e. that around θ_0 there exists a neighborhood, say $N(\theta_0, \epsilon)$, $\epsilon > 0$, which lies entirely in Θ. When we turn, however, to the problem involving inequality constraints, the inf or sup of the problem as well as the true parameter point **lie on the boundary of the admissible space**.

Central to the development of ML procedures in the context of parameter spaces Θ of unspecified form where the true parameter point θ_0 is on the boundary, is the approximation of the space, in a subset containing θ_0, by a cone with vertex at θ_0. What this means is made clear in

Definition 1. The admissible parameter space $\Theta \subset R^m$ is said to be approximated at the point $\theta_0 \in \Theta$ by a cone C with vertex at θ_0, if

i . $\inf_{x\in C} ||x-y|| = o(||y-\theta_0||)$ for all $y \in \Theta$ **and**

ii. $\inf_{y\in\Theta} ||x-y|| = o(||x-\theta_0||)$ for all $x \in C$.

The fact that $\theta_0 \in \partial\Theta$, i.e. it is a boundary point, does not create insurmountable problems, but the usual proofs of consistency would need to be revised. For example the part of the proof of consistency given in Dhrymes (1994), Proposition 3, part iii, pp. 282-285, is not allowable if θ_0 is on the boundary. Another difficulty is that we cannot routinely seek the ML estimator by examining the solution of the first order conditions.

Nonetheless a proof of consistency (in the sense of convergence in probability) and of $\sqrt{n}$-consistency can be had. The original such proof is given in Chernoff (1954), which is the one we shall follow in our discussion, even though it is quite terse and should be amplified. A more general careful discussion is given in Andrews (1999), but we shall not deal with it because it is so general and, thus, more complicated; this makes it more difficult to derive specifically what we need for the special situation at hand. Finally, we shall use elements in the proof of Proposition 3 and Corollary 1 in Dhrymes (1994), pp. 282-285.

In dealing with the problem in the context we defined above, we employ the following assumptions:

i. the observations $\{x_i : i = 1, 2, 3, \dots T\}$ are i.i.d. with common density $f(\cdot;\theta)$ where $\theta \in \Theta \subset R^m$;

ii. the derivatives of f, to order at least three, exist a.c. in the closure of a neighborhood, $N(\theta_0, \epsilon)$ of the true parameter point θ_0;

iii. the density has continuous derivatives, at least to order three, which are majorized by the integrable function H, i.e. $||Df(x;\theta)|| < H(x)$, $||D^2 f(x;\theta)|| < H(x)$ etc; moreover, for some constant M $EH(x) < M$, independently of θ, so that in effect the expected values of all derivatives are **uniformly bounded**;

iv. let $L(\theta; X)$ be the log-likelihood function of the problem, put

$$L_T(\theta) = \frac{1}{T} L(\theta; X)] = \frac{1}{T}\sum_{t=1}^{T} \ln f(X_t;\theta); \tag{1.36}$$

define

$$\lim_{T\to\infty} \frac{1}{T} E[L(\theta; X)] = L^*(\theta) = E[\ln f(x;\theta)], \tag{1.37}$$

and assume that the convergence is **uniform** in θ.

v. let ϕ be an ϵ-neighborhood of Θ, whose closure contains θ_0, and the latter is a limit point of ϕ; then

$$\sup_{\theta\in\Theta\backslash\phi} L^*(\theta) < L^*(\theta_0), \tag{1.38}$$

and we assume that $L^*(\theta) = L^*(\theta_0)$ implies $\theta = \theta_0$;

vi. the matrix $J = E[(Df)'(Df)]$ is positive definite for $\theta \in N$.

The first four conditions are technical, while conditions v and vi are identification conditions. The notation Df denotes the (row) vector of first order partial derivatives, i.e. $Df = \partial \ln f / \partial \theta$ and is thus a $1 \times m$ vector.

Since

$$\frac{1}{T} L(\theta; X) = \frac{1}{T} \sum_{i=1}^{T} \ln f(x_i; \theta), \tag{1.39}$$

we may expand the log-likelihood function, about θ_0 as

$$\begin{aligned} \frac{1}{T} L(\theta; X) &= \frac{1}{T} L(\theta_0, X) + A(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' B(\theta - \theta_0) + ||\theta - \theta_0||^3 O_p(1) \\ A &= \frac{1}{T} \frac{\partial L(\theta_0; X)}{\partial \theta}, \\ B &= \frac{1}{T} \frac{\partial^2 L(\theta_0; X)}{\partial \theta \partial \theta}. \end{aligned} \tag{1.40}$$

The derivatives evaluated at θ_0 are **left or right derivatives**, as the case requires, which also exist. By the conditions of this problem B is the sum of i.i.d. entities (matrices), whose expected value exists by item iii above; consequently by a law of large numbers, see e.g. Dhrymes (1989), p. 188, it converges to its expected value, in view of condition iii. By the properties of ML estimation, this is equal to the negative of the matrix J noted in item vi.

We thus obtain

Proposition 1. Let $\{X_i : i = 1, 2, 3, \ldots, T\}$ be a sequence of i.i.d. random elements with common density $f(\cdot; \theta)$, $\theta \in \Theta \subset R^k$ and that Θ is **bounded**. Suppose that the true parameter point $\theta_0 \in \partial \Theta$, i.e. it lies on the boundary of the admissible space, and conditions i through vi hold. The following statements are true where, in general, we define the ML estimator to be

$$L_T(\hat{\theta}_{ML}) = \sup_{\theta \in \Theta} L_T(\theta; \theta) + o_p(1). \tag{1.41}$$

In particular for $\phi \subset \Theta$ such that θ_0 is a limit point of ϕ, the ML estimator, $\hat{\theta}_\phi$, is determined by the condition

$$L_T(\hat{\theta}_\phi; X) = \sup_{\theta \in \phi} L_T(\theta; X). \tag{1.42}$$

The following statements are true:

i. $\hat{\theta}_\phi \xrightarrow{\mathrm{P}} \theta_0$;

ii. given i, $\hat{\theta}_\phi - \theta_0 = O_p(T^{-(1/2)})$, i.e. $\sqrt{T}(\hat{\theta}_\phi - \theta_0)$ remains bounded as $T \rightarrow \infty$, which is the definition of $\sqrt{n}$-consistency

Proof. First we clarify the conditions in item iv, in the discussion preceding Proposition 1. The right member of Eq. (1.37) is (the sum of) a sequence of i.i.d. random variables with finite mean and variance and, thus, see Dhrymes (1989), pp. 188-190, it converges **at least in probability to** its expected value $E[\ln f(x;\theta)] = L^*(\theta)$.

The same reasoning will also show that the matrix B of Eq. (1.41) converges, at least in probability to $-J$, as claimed earlier, and A of Eq. (1.41) converges to zero at least in probability because $E[D\ln f(x;\theta_0)] = 0$.

To prove consistency we note that the maximum likelihood estimator here is given by

$$L_T(\hat{\theta}_\phi) = \sup_{\theta \in \phi} L_T(X;\theta). \tag{1.43}$$

Moreover, it is well-known, see Dhrymes (1994) p. 209, that since $L^*(\theta) = E[\ln f(\cdot;\theta)$,

$$L^*(\theta) = E_{\theta_0}[\ln f(\cdot;\theta] \leq E_{\theta_0}[\ln f(\cdot;\theta_0) = L^*(\theta_0) \tag{1.44}$$

By assumption iv, we have that, at least,

$$L_T(X;\theta) \xrightarrow{\mathrm{P}} L^*(\theta), \quad \text{uniformly in } \theta. \tag{1.45}$$

Thus, for any subset, say, $S \subset \Theta$

$$\sup_{\theta \in S} L_T(X;\theta) \xrightarrow{\mathrm{P}} \sup_{\theta \in S} L^*(\theta). \tag{1.46}$$

Because of the smoothness of the likelihood function and uniform convergence

$$\hat{\theta}_\phi \xrightarrow{\mathrm{P}} \bar{\theta}_\phi, \tag{1.47}$$

where $\bar{\theta}_\phi$ is defined by

$$L^*(\bar{\theta}_\phi) = \sup_{\theta \in \phi} L^*(\theta). \tag{1.48}$$

Since θ_0 is a limit point of ϕ, and $\sup_{\theta \in \phi} L^*(\theta) \leq L^*(\theta_0)$, we conclude that

$$L^*(\bar{\theta}_\phi) = L^*(\theta_0) \tag{1.49}$$

which, by the identification condition, implies $\bar{\theta}_\phi = \theta_0$, so that we can write,

$$\hat{\theta}_\phi = \theta_0 + o_p(1), \tag{1.50}$$

which shows consistency.

To show $\sqrt{n}$-consistency return to Eq.(1.41) and substitute $\hat{\theta}_\phi$ for θ to obtain

$$\begin{aligned}\frac{1}{T}L(\hat{\theta}_\phi;X) &= \frac{1}{T}L(\theta_0,X) + A(\hat{\theta}_\phi-\theta_0) + \frac{1}{2}(\hat{\theta}_\phi-\theta_0)'B(\hat{\theta}_\phi-\theta_0) + ||\hat{\theta}_\phi-\theta_0||^3 O_p(1) \\ &= \frac{1}{T}L(\theta_0,X) + A(\hat{\theta}_\phi-\theta_0) - \frac{1}{2}(\hat{\theta}_\phi-\theta_0)'J(\hat{\theta}_\phi-\theta_0) \\ &\quad + \frac{1}{2}(\hat{\theta}_\phi-\theta_0)'(B+J)(\hat{\theta}_\phi-\theta_0) + ||\hat{\theta}_\phi-\theta_0||^3 O_p(1). \qquad (1.51)\end{aligned}$$

Since over ϕ, $\frac{1}{T}L(\hat{\theta}_\phi;X) - \frac{1}{T}L(\theta_0,X) \geq 0$, we obtain

$$\begin{aligned}0 &< A(\hat{\theta}_\phi-\theta_0) + \frac{1}{2}(\hat{\theta}_\phi-\theta_0)'B(\hat{\theta}_\phi-\theta_0) + ||\hat{\theta}_\phi-\theta_0||^3 O_p(1) \\ &< -\frac{1}{2}(\hat{\theta}_\phi-\theta_0)'J(\hat{\theta}_\phi-\theta_0) + m_T^*\left(\frac{||\hat{\theta}_\phi-\theta_0||}{\sqrt{T}} + ||\hat{\theta}_\phi-\theta_0||^2\right). \qquad (1.52)\end{aligned}$$

The right member of Eq. (1.53) is justified as follows:

i. by the consistency of $\hat{\theta}_\phi$, there exists a constant m_{1T} such that $||\hat{\theta}_\phi - \theta_0|| < m_{1T}$ with probability arbitrarily close to one;

ii. because $A/\sqrt{T} \xrightarrow{\text{d}} N(0,J)$ there exists a constant m_{2T} such that $||A|| < \frac{m_{2T}}{\sqrt{T}}$ with probability arbitrarily close to one;

iii. similarly, because $B \xrightarrow{\text{P}} -J$, there exists a constant m_{3T} such that $||B+J|| < m_{3T}$ with probability arbitrarily close to one;

iv. there exists a constant m_{4T} such that $||\hat{\theta}_\phi-\theta_0||^3 O_p(1) < m_{4T}||\hat{\theta}_\phi-\theta_0||^3$.

Hence, for another constant m_T^* which is equal to or greater than the maximum of $m_{iT}, i=1,2,3,4$ we obtain the rightmost member of Eq.(1.53). Next, rewrite it as

$$\frac{1}{2}(\hat{\theta}_\phi-\theta_0)'J(\hat{\theta}_\phi-\theta_0) < m_T^*\left(\frac{||\hat{\theta}_\phi-\theta_0||}{\sqrt{T}} + ||\hat{\theta}_\phi-\theta_0||^2\right), \qquad (1.53)$$

and note that J is a **positive definite** matrix so that

$$\lambda_{\min}(J)[\hat{\theta}_\phi-\theta_0]'[\hat{\theta}_\phi-\theta_0] \leq (\hat{\theta}_\phi-\theta_0)'J(\hat{\theta}_\phi-\theta_0). \qquad (1.54)$$

Consequently,

$$||\hat{\theta}_\phi-\theta_0||^2 < \frac{1}{\lambda_{\min}(J)}\ m_T^*\left(\frac{||\hat{\theta}_\phi-\theta_0||}{\sqrt{T}} + ||\hat{\theta}_\phi-\theta_0||^2\right). \qquad (1.55)$$

Cancelling out a term $||\hat{\theta}_\phi - \theta_0||$ from both sides there exists a constant m_T such that

$$\hat{\theta}_\phi - \theta_0 < \frac{m_T}{\sqrt{T}}, \tag{1.56}$$

which demonstrates $\sqrt{n}$-consistency.

q.e.d.

Remark 1. We should note that the ML estimator here is not obtained *in the usual way* over ϕ. What we do is to approximate ϕ by **a cone with vertex at** θ_0, and the estimation is treated as a nonlinear program.

1.3 The GLM with Inequality Constraints

The problem to be examined below is the estimation of the GLM parameters in the case of inequality constraints; precisely, we estimate the parameter θ in the model

$$y = X\theta + u, \text{ subject to } R\theta \geq 0, \ R \text{ is } k \times m, \text{ of rank } k \leq m. \tag{1.57}$$

The set of problems arising in the context of a model with inequality constraints of the form, say $\theta \geq 0$, or $R\theta \geq 0$, was largely developed in the context of testing hypotheses in the biomedical area of research. What sets it apart from the standard GLM with equality constraints examined in a previous section is the nature of the space on which we project observations. If we have no constraints, the space on which we project is a linear space, precisely the column space of the matrix of explanatory variables, X, which is of dimension m. When we have only equality constraints, the space on which we project is also a linear space, but of reduced dimension, more precisely of dimension $m-k$, where k is the number of equality constraints. When there are **inequality constraints** the space on which we project is usually a **closed convex polyhedron**. In such cases, by necessity, the inf and sup of the objective or criterion function, as the case may be, **lies on the boundary of the space** and seems reasonable to assert that the true parameter point also lies on the boundary. For example, if we have to test the efficacy of two or more, say s, treatments, as compared with the standard treatment, we examine the appropriate test statistics for evidence of non-improvement, i.e. evidence that $\theta_i \leq 0$, or of improvement, $\theta_i > 0$, where θ_i is the difference between the i^{th} new treatment and the standard treatment. The parameter space here is Θ, and the space we are interested in is the non-negative orthant, with the exception of the origin. Evidently, the test procedure has to take into account the possibility that the test statistics will fall in any one of several orthants, and there are 2^m orthants in all.

Similarly in the case $R\theta \geq 0$, the parameter space in question is a polyhedral cone. As we know from the appendix a polyhedral cone is covered by the collection of the relative interior(s) of its faces. The latter are entities of the form

$$\mathrm{ri}(F_J) = \{\theta : R_J\theta = 0, \quad R_J^*\theta > 0\} \tag{1.58}$$

where R_J is a sub-matrix of R containing J columns, while R_J^* is the complement of R_J in R, i.e. it contains its remaining columns.

As is clear from the appendix a convex polyhedral cone contains at most 2^k faces, where k is the number of restrictions imposed by R. Since the projection may be on any one of these faces, (properly their relative interiors) it is clear that distributional issues, not to mention estimation issues, are considerably more complicated than in the earlier GLM context.

An important feature of such problems is that the estimators, of necessity, lie on the **boundary** of the admissible space, and a consequence of this is that the LR test statistic does not have the classical central χ^2 distribution.[8]

Before we proceed with the discussion of issues of estimation and testing, however, we shall present an alternative to the procedures followed above which will yield a result that will be very important in the current context.

This is the analog of the result given, e.g. in Dhrymes (1989), pp. 121-125, which is: Let X be a r.v. defined on the probability space (Ω, $\mathcal{A}$, $\mathcal{P}$) and let $\mathcal{G}_i, i = 1, 2$ be two σ-(sub)algebras such that $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{A}$, then

$$E[E(X|\mathcal{G}_2)|\mathcal{G}_1] = E(X|\mathcal{G}_1). \tag{1.59}$$

The analog for projections, as given in the appendix, reads: Let y be a point in R^n, the n-dimensional Euclidean space; let $C_i, i = 1, 2$ be subspaces of R^n such that $C_1 \subseteq C_2$, then

$$P[P(y|C_2)|C_1] = P(y|C_1), \tag{1.60}$$

where the notation $P(x|C)$ denotes the projection of a point $x \in R^n$, on the subspace $C \subseteq R^n$, where n is the number of observations.

[8] The fact that the LR test statistic does not have the classical central χ^2 distribution is given in Chernoff (1954) who obtained the proper distribution in a special case, but did not examine the general case.

Remark 2. In Propositions 6 and 7 of the appendix we have used the notation $P(x)$, and $P^{\circ}(x)$ to denote the projection of a point $x \in R^n$ onto a space V and the projection of x onto its orthogonal complement $V^{\perp}$, respectively. Our usage here, and in all discussion outside the appendix, will be $P(x|V)$ and $P(x|V^{\perp})$, respectively. Notice also that, by definition, $P^{\circ} = I - P$, in the notation of the appendix, and $P(x|V^{\perp}) = (I-P)(x|V)$, or $x - P(x|V)$. The usage in the appendix is dictated by the notation of the sources cited.

The application of these propositions to the problem we considered in previous sections leads to the conclusion that if we wish to estimate the regression

$$y = X\beta + u, \quad \text{subject to} \quad R\beta = r, \tag{1.61}$$

we can proceed in two steps: first, project y on the m-dimensional subspace of R^T generated by the (m linearly independent) columns of X, say C_2, and **then project the result on the subspace of** C_2, given by

$$C_1 = \{\gamma \in C_2 : \gamma = Xb, \ b \in \{b : Rb = r\}\}. \tag{1.62}$$

Carrying out the operation, the first step yields, $\hat{y} = X\hat{\beta}$, where $\hat{\beta}$ is the unrestricted least squares estimator; the second step then involves

$$\min_{b \in C_1} (\hat{\beta} - b)' X' X (\hat{\beta} - b), \tag{1.63}$$

which yields, using the method of Lagrange multipliers,

$$\tilde{b} = \hat{\beta} - (X'X)^{-1} R' [R(X'X)^{-1} R']^{-1} (R\hat{\beta} - r). \tag{1.64}$$

A comparison with Eq. (1.4) **shows that the two estimators are the same!** Note that when $r = 0$ we may write the equation above

$$\tilde{b} = [I_m - (X'X)^{-1} R' [R(X'X)^{-1} R']^{-1} R] \hat{\beta},$$

where $I_m - (X'X)^{-1} R' [R(X'X)^{-1} R']^{-1} R$ is a **non-symmetric idempotent matrix** of rank $m - k$.

Another way of looking at the problem is to note that

$$\inf_{\beta \in C_1} (y - X\beta)'(y - X\beta) = y'[I - X(X'X)^{-1} X'] y + \inf_{\beta \in C_1} (\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta), \tag{1.65}$$

and it is apparent that the **minimizer is the same** whether obtained according to the left member of Eq. (1.65) or according to the second term of the right member.

Since the problem of this section is to minimize

$$(y - X\beta)'(y - X\beta), \quad \text{subject to} \quad \beta \in C_3, \quad C_3 = \{\beta : R\beta \geq 0\}, \tag{1.66}$$

and utilizing the result above, we may reformulate the problem as

$$\min_{b \in C_3} (\hat{\beta} - b)'(X'X)(\hat{\beta} - b). \tag{1.67}$$

This is a quadratic program which does not have a closed form solution, and it involves the projection of a vector in R^m, $\hat{\beta}$, on the polyhedral cone C_3.[9] In contrast to the previous problem where C_2 was an (m-dimensional) hyperplane in R^T, here C_3 is the intersection of k half spaces generating a (closed convex) polytope. [10]

Because the context of this operation is non-standard in econometrics we give first a theorem that deals with these issues.

Following the discussion at the beginning of the previous section, we shall first show that the estimator of β in this context is consistent, indeed $\sqrt{n}$-consistent, and then derive the LR-like test statistic for testing the null hypothesis, $H_0 : R\beta = 0$, as against the alternative $H_1 : R\beta \geq 0$.

Set as the objective function the least squares expression

$$S_T(\beta) = \frac{1}{T}(y - X\beta)'(y - X\beta) \tag{1.68}$$

and note that

$$S_T(\beta) = \frac{1}{T}\sum_{t=1}^{T}[u_t - x_{t\cdot}(\beta - \beta_0)]^2, \tag{1.69}$$

i.e., **conditional on the** $x's$, it is (the average of the squares of) a sequence of independent non-identically distributed random variables with mean $x_{t\cdot}(\beta-\beta_0)$ and variance σ^2; incidentally, this incorporates the typical GLM assumption that the $x's$ are independent on the $u's$. We would expect then that $S_T(\beta)$ will converge at least in probability. Indeed, it may be shown that, under certain assumptions on $x_{t\cdot}$, it converges a.c., i.e. with probability one, **uniformly in** β. To see that rewrite Eq.(1.69) as

$$S_T(\beta) = \frac{1}{T}(u - X(\beta - \beta_0))'(u - X(\beta - \beta_0)) \tag{1.70}$$

[9] More accurately it is a projection of $X\hat{\beta}$ on the polyhedral cone XC_3.

[10] Often the literature refers to this construct as a polyhedron or polyhedral surface; but, by convention, the term polyhedron is a term in solid geometry and thus is defined in **three** dimensions. Nonetheless to avoid using unfamiliar and seldom used terms we shall henceforth refer to such constructs as closed convex polyhedral cones.

$$= \frac{1}{T}[u'u + (\beta - \beta_0)'X'X(\beta - \beta_0) - 2u'X(\beta - \beta_0)].$$

Since ϕ of the previous discussion is bounded, we can write

$$\begin{aligned}
||S_T(\beta) - S^*(\beta)|| &\leq |\left\|\frac{u'u}{T} - \sigma^2\right\| + ||\beta - \beta_0||^2 \left\|\frac{X'X}{T} - M_{xx}\right\| + 2||\beta - \beta_0|| \left\|\frac{u'X}{T}\right\| \\
&\leq K\left(\left\|\frac{u'u}{T} - \sigma^2\right\| + \left\|\frac{u'X}{T}\right\| + \left\|\frac{X'X}{T} - M_{xx}\right\|\right) \\
&\xrightarrow{\mathrm{P}} 0,
\end{aligned} \tag{1.71}$$

where

$$S^*(\beta) = \sigma^2 + (\beta - \beta_0)'M_{xx}(\beta - \beta_0), \quad M_{xx} = \operatorname*{plim}_{T\to\infty} \frac{X'X}{T} > 0, \quad X'X > 0, \tag{1.72}$$

and K is related to the bound on β, or at least on $\beta \in \phi$. It is seen then that conditions i through vi asserted in connection with Proposition 1, hold in this context as well. (Note that the condition $E[(Df)'(Df)] = J > 0$ holds here as well except that it is much simpler, it being $X'X/T > 0$. Thus, although we can derive consistency as well as $\sqrt{n}$-consistency as a by product of Proposition 1, we shall state the matter as a separate proposition because the proofs in this case are much simpler.

Proposition 2. Consider the context of Proposition 1, except that the likelihood function is replaced by the least squares expression

$$S_T(\beta) = \frac{1}{T}(y - X\beta)'(y - X\beta) \tag{1.73}$$

and the problem is to estimate β subject to $R\beta \geq 0$. Let the admissible parameter space be given by Θ; let the true parameter point be β_0, and $\phi \subset \Theta$ such that $\beta_0 \in \mathrm{cl}\ \phi$; moreover, it is assumed that β_0 is a limit point of ϕ, i.e. every neighborhood of β_0, however small, contains infinitely many points in ϕ. Let

$$S_T(\hat{\beta}_\phi) = \inf_{\beta\in\phi} \frac{1}{T}(y - X\beta)'(y - X\beta), \tag{1.74}$$

then the following statements are true:

i. $\hat{\beta}_\phi \xrightarrow{\mathrm{P}} \beta_0$;

ii. $\hat{\beta}_\phi$ is $\sqrt{n}$-consistent.

Proof. We recall that

$$S_T(\beta) \xrightarrow{\mathrm{P}} S^*(\beta) = \sigma^2 + (\beta - \beta_0)'M_{xx}(\beta - \beta_0); \tag{1.75}$$

put

$$S_T(\hat{\beta}_\phi) = \inf_{\beta \in \phi} S_T(\beta), \tag{1.76}$$

and note that the sequence of estimators $\hat{\beta}_\phi$ is bounded because ϕ is bounded, and thus has at least one limit point; recall also that β_0 is a limit point of ϕ and that $S^*(\beta_0) \leq S^*(\beta)$ for any $\beta \in \phi$. Since, by construction,

$$S_T(\hat{\beta}_\phi) \leq S_T(\beta_0), \tag{1.77}$$

taking probability limits on both sides, and noting that convergence is uniform, we find

$$S^*(\bar{\beta}_\phi) \leq S^*(\beta_0), \quad \text{which implies } \bar{\beta}_\phi = \beta_0, \tag{1.78}$$

where $\bar{\beta}_\phi = \mathrm{plim}_{T\to\infty} \hat{\beta}_\phi$.

This concludes the proof for consistency.

To show $\sqrt{n}$-consistency, first consider the expansion of $S_T(\hat{\beta}_\phi)$ about β_0 and **note that it is an exact quadratic**, i.e.

$$\begin{aligned} S_T(\hat{\beta}_\phi) &= \frac{1}{T}(u - X(\hat{\beta}_\phi - \beta_0))'(u - X(\hat{\beta}_\phi - \beta_0)) \\ &= S_T(\beta_0) + \frac{\partial S_T}{\partial \beta}(\beta_0)(\hat{\beta}_\phi - \beta_0) + \frac{1}{2}(\hat{\beta}_\phi - \beta_0)' \frac{\partial^2 S_T}{\partial\beta\partial\beta}(\beta_0)(\hat{\beta}_\phi - \beta_0). \end{aligned} \tag{1.79}$$

Since by construction $S_T(\hat{\beta}_\phi) - S_T(\beta_0) \leq 0$ we may write

$$0 \geq -2\frac{u'X}{T}(\hat{\beta}_\phi - \beta_0) + (\hat{\beta}_\phi - \beta_0)'\frac{X'X}{T}(\hat{\beta}_\phi - \beta_0), \tag{1.80}$$

and, canceling $||\hat{\beta}_\phi - \beta_0||$ from both sides, we find

$$||\hat{\beta}_\phi - \beta_0|| \leq \frac{2}{||X'X/T||} \left\| \frac{X'u}{\sqrt{T}} \right\| \frac{1}{\sqrt{T}}. \tag{1.81}$$

Because of the convergence in probability of $X'X/T$ and the convergence in distribution of $X'u/\sqrt{T}$, there exists a constant K, such that with probability greater than $1-\epsilon$, for arbitrary $\epsilon > 0$,

$$\hat{\beta}_\phi - \beta_0 \leq \frac{K}{\sqrt{T}}, \tag{1.82}$$

which shows $\sqrt{n}$-consistency.

q.e.d.

1.4 Distributional Aspects

1.4.1 The case $H_0 : \theta = 0$, as against $H_1 : \theta > 0$

Although there are many references to the distributional issues in such models, we shall generally follow the presentatiion in Shapiro (1985), (1988), which we have found more complete and lucid than other presentations.

Before we proceed with the topic of this section we give two results that we will find useful in the ensuing discussion.

Proposition 3. Let $x \in R^n$ and C, K closed convex cones in R^n, $C \subset K$. Then

i. $||x||^2 = ||P(x|C)||^2 + ||x - P(x|C)||^2$;

ii. if either C or K is a linear space

$$||x - P(x|C)||^2 = ||x - P(x|K)||^2 + ||P(x|K) - P(x|C)||^2.$$

Proof. For i, we note from the projection theorem that $P(x|C)$ is **orthogonal to** $x - P(x|C)$, so that

$$x = P(x|C) + (x - P(x|C)), \text{ and thus } ||x||^2 = ||P(x|C)||^2 + ||(x - P(x|C))||^2.$$

As for ii, we note that since $C \subset K$, adding and subtracting $P(x|K)$, we find

$$x - P(x|C) = (x - P(x|K)) + (P(x|K) - P(x|C)), \text{ which implies}$$

$$||x - P(x|C)||^2 = ||x - P(x|K)||^2 + ||P(x|K) - P(x|C)||^2,$$

essentially because $P(x|C)'P(x|K) = ||P(x|C)||^2$. That the latter is so may be demonstrated as follows: Let $x \in R^n$ and P, Q be the projection matrices onto C and K, respectively, i.e. $P(x|C) \in C$ and $Q(x|K) \in K$; because C is a linear space P is symmetric and idempotent. Now for all $x \in C$

$$(P - Q)(x) = 0, \text{ which implies } P(x|C)'[P(x|C) - Q(x|K)] = 0, \text{ or}$$

$$P(x|C)'P(x|C) = P(x|C)'Q(x|K)] = ||P(x|C)||^2.$$

q.e.d.

We now examine the case where we are testing $H_0 : \theta = 0$ as against $H_1 : \theta > 0$.

Let the relevant observations be denoted by $x_{t\cdot} = (x_{t1}, x_{t2}, \ldots, x_{tm})$, $t = 1, 2, \ldots T$. It is assumed that the observations are i.i.d. with distribution $N(\theta, I_m)$.

The LR-like statistic is proportional to the sum of squared residuals under the restricted model minus the sum of squared residuals under the unrestricted model. For this case we obtain

$$LR \sim \frac{1}{T} X'X - \inf_{\theta > 0} \frac{1}{T} (X - e\theta)'(X - e\theta), \tag{1.83}$$

where θ is taken here to be an m-element **row vector**, and e a T-element **column** vector all of whose elements are one. Often minimizations as in the second term of the right member above can be reduced to a minimization involving a single variable only. This occurs, for example when the minimization is over a convex cone which is contained in a linear space. In particular, consider

$$\begin{aligned}\frac{1}{T}(X - e\theta)'(X - e\theta) &= \frac{1}{T}\left((X - e\bar{x})'(X - e\bar{x}) + [e(\bar{x} - \theta)]'[e(\bar{x} - \theta)]\right)\\ &= \frac{1}{T}(X - e\bar{x})'(X - e\bar{x}) + [(\bar{x} - \theta)]'[(\bar{x} - \theta)],\end{aligned}$$

where $\bar{x} = e'X/T$, so that we need operate only with the second component of the right member since the first component does not contain θ . Notice that the second component depends **only on a single observation of the random variable** $\bar{x}$. Moreover, the estimator minimizing the rightmost member of Eq.(1.83) is precisely the estimator minimizing the second term of the rightmost member of the equation above.

Denoting the transform of the LRT statistic by $-2\ln\lambda$,

$$-2\ln\lambda \sim ||\bar{x}||^2 - ||\bar{x} - \tilde{\theta}||^2 = ||\tilde{\theta}||^2 = ||P(x|C)||^2, \quad C = \{\theta : \theta > 0\}. \tag{1.84}$$

This is so because $\tilde{\theta}$ is a projection and thus $\bar{x} - \tilde{\theta}$ is orthogonal to $\tilde{\theta}$. Notice also that the minimizing process assigns to $\tilde{\theta}_i$ the value zero when $\bar{x}_i \leq 0$, and the value $\bar{x}_i$ when $\bar{x}_i > 0$; thus $\bar{x}'\tilde{\theta} = \tilde{\theta}'\tilde{\theta}$. A more direct way of deriving this test statistic is to note that

$$\begin{aligned}-2\ln\lambda &= \frac{1}{T}X'X - \min_{\beta \in C} \frac{1}{T}(X - e\theta)'(X - e\theta) = \frac{1}{T}X'X - \frac{1}{T}(X - e\tilde{\theta})'(X - e\tilde{\theta})\\ &= \frac{1}{T}X'X - \left[\frac{1}{T}X'X + \tilde{\theta}'\tilde{\theta} - 2\bar{x}'\tilde{\theta}\right] = \tilde{\theta}'\tilde{\theta}.\end{aligned} \tag{1.85}$$

The distribution of the statistic in Eq. (1.85), however, is not invariant with respect to the locus of the projection, i.e. it has one distribution if it lands on the positive orthant and a different distribution if it lands in another orthant. Since we have m entities, and each of the $\bar{x}_i$, can assume a non-positive or a positive value, there are 2^m orthants, a constructive representation of which was given in the appendix. To particularize it to this case, S_m is the positive orthant, where **all** the $\bar{x}_i$ are positive. Note that the index, (m), refers to the number of positive $\tilde{\theta}_i$ (or $\bar{x}_i$); S_{m-1} represents the orthants in which all but one of the $\bar{x}_i > 0$, and the others are non-positive. There are m of these; S_{m-2} represents the orthant(s) in which $m-2$ of the $\bar{x}_i > 0$ and all the others are non-positive. There are $\binom{m}{m-2}$ and so on until S_0, which contains $\bar{x}_i$, which are all **negative** (non-positive). Thus, there are

$$(1+1)^m = \sum_{k=0}^{m} \binom{m}{k} = 2^m,$$

relevant orthants as noted earlier. Because of the symmetry of the normal, the probability of being in any one them is $1/2^m$. Thus, the probability of being in S_m or S_0 is in both cases $1/2^m$. The probability of being in S_{m-1} is $m/2^m$ and, more generally, the probability of being in S_j is

$$\Pr(\tilde{\theta} \in S_j) = \frac{\binom{m}{j}}{2^m}, \quad \text{so that} \quad \sum_{j=0}^{m} \Pr(\tilde{\theta} \in S_j) = \sum_{j=0}^{m} \frac{\binom{m}{j}}{2^m} = 1, \tag{1.86}$$

as was to be expected. Now, in the region denoted by S_j, there are j positive and $m-j$ non-positive $\bar{x}_i$. The latter are then mapped to the origin, and the positive ones are mapped onto a positive orthant. Hence, if $\bar{x} \in S_j$, j of the elements of $\tilde{\theta}$ are **positive** and the remaining are zero, so that, under the null, $\tilde{\theta}'\tilde{\theta}$ is the sum of squares of j **independent** $N(0, 1/T)$ variables; hence its distribution is χ^2_j, i.e. it is proportional to a central chi squared with j degrees of freedom, because $\sqrt{T}\tilde{\theta}_i \sim N(0,1)$, when it is not equl to zero.

It follows therefore that

$$Pr(\lambda \le \tau) = \sum_{j=0}^{m} Pr(\tilde{\theta} \in S_j) Pr(\chi^2_j \le \tau). \tag{1.87}$$

In deriving this representation we have made use of the **convention** that $\chi^2_0 \equiv 0$. Thus, for example in S_0, $\tilde{\theta}'\tilde{\theta} = 0^2 \equiv \chi^2_0$.

Example 1. Suppose $m = 2$; then there are $2^2 = 4$ orthants, each with probability of $.25$ due to the properties of the normal; S_2 corresponds to the case where $\tilde{\theta}_i = \bar{x}_i, i = 1,2$ are both positive; note that $\sqrt{T}\bar{x}_i, i = 1,2,$ are independently $N(0,1)$. Thus

$$T(\bar{x}_1^2 + \bar{x}_2^2) \sim \chi_2^2;$$

S_1 corresponds to the case where one of the $\tilde{\theta}_i, i = 1,2$ is non-positive, i.e. in the first component of S_1, $\bar{x}_1 \leq 0$ and $\bar{x}_2 > 0$, or in the second component $\bar{x}_2 \leq 0$ and $\bar{x}_1 > 0$. So

$$\Pr(\tilde{\theta} \in S_1) = .5 \tag{1.88}$$

and the distribution of $\tilde{\theta}_i^2 | S_1$ is that of $(\sqrt{T}\bar{x}_i)^2$ for $i = 1$ **or** $i = 2$; in either case it is a central χ_1^2.

Finally, for S_0, $\bar{x}_1, \bar{x}_2 \leq 0$, and thus $\tilde{\theta} = 0$; by convention, $\tilde{\theta}'\tilde{\theta} \sim \chi_0^2 \equiv 0$. Thus, the LR test statistic obeys

$$\Pr(\lambda \leq \tau) = \sum_{j=0}^{2} Pr(\tilde{\theta} \in S_j) Pr(\tilde{\theta}'\tilde{\theta} \leq \tau | \tilde{\theta} \in S_j) \tag{1.89}$$

$$= .25\Pr(\chi_0^2 \leq \tau) + .5\Pr(\chi_1^2 \leq \tau) + .25\Pr(\chi_2^2 \leq \tau).$$

This distribution is commonly denoted by $\bar{\chi}^2$, and is referred to as the **chi bar squared** distribution, it being a mixture of central chi squared distributions. Since for $\tau \geq 0$, $\Pr(\chi_0^2 \leq \tau) = 1$, the equation above can be simplified to

$$\Pr(\lambda \leq \tau) = \sum_{j=0}^{2} Pr(\tilde{\theta} \in S_j) Pr(\tilde{\theta}^2 \leq \tau | \tilde{\theta} \in S_j) \tag{1.90}$$

$$= .25 + .5\Pr(\chi_1^2 \leq \tau) + .25\Pr(\chi_2^2 \leq \tau).$$

1.4.2 The case $H_0 : R\beta \geq 0$, as against $H_1 : \beta \in R^m$

What is involved in this section is the test of the hypothesis $H_0 : R\beta \geq 0$, as against the alternative $H_1 : \beta \in R^m$ where, in a vector context, the notation $\geq$ means that the elements of the vector are **nonnegative** and there is at least one element which is strictly positive. To be exact the model is, under the null $H_0 : \beta \in C$, $C = \{\beta : R\beta \geq 0\}$,

$$y = X\beta + u, \quad \text{subject to} \quad R\beta \geq 0. \tag{1.91}$$

Maximum likelihood estimation, under normality with known variance, involves the minimization of the sum of squared errors. Under the alternative, $H_1 : \beta$ **unrestricted**, the problem is

$$\min_{\beta \in R^m} \frac{1}{T}(y-X\beta)'(y-X\beta) = \frac{1}{T}(y-X\hat{\beta})'(y-X\hat{\beta}), \;\; \hat{\beta} = (X'X)^{-1}X'y = P(y|\mathcal{M}), \tag{1.92}$$

where $\mathcal{M}$ is the linear space generated by the columns of X.

Since $C \subseteq \mathcal{M}$, the discussion in the appendix implies that

$$P(y|C) = P[P(y|\mathcal{M})|C], \; C = \{\beta : R\beta \geq 0\}. \tag{1.93}$$

Thus, the problem under consideration may be stated as

$$\min_{\beta \in C} \frac{1}{T}(y - X\beta)'(y - X\beta) = \frac{1}{T}(y - X\hat{\beta})'(y - X\hat{\beta}) + \min_{\beta \in C} \frac{1}{T}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta). \tag{1.94}$$

Now, given normality with known variance, the likelihood ratio behaves like

$$\begin{aligned} -2\ln\lambda \sim & \min_{\beta \in C} \frac{1}{T}(y - X\beta)'(y - X\beta) - \min_{\beta \in \mathcal{M}} \frac{1}{T}(y - X\beta)'(y - X\beta) \\ = & \min_{\beta \in C} \frac{1}{T}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) = \frac{1}{T}(\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}), \end{aligned} \tag{1.95}$$

and the LR test rejects the null for "large" values of the statistic in the right member of Eq. (1.95). Since the fraction $1/T$ plays no role, except in asymptotics, it will henceforth be omitted.

One complicating circumstance is that in the case where the admissible parameter space is a **polyhedral cone**, by necessity, the maximum **must occur** on its boundary and thus the true parameter point **must also be asserted to be on the boundary**. We have dealt with properties of estimators in this context at an earlier section and we have shown that such estimators are consistent and $\sqrt{n}$-consistent. Thus in the following discussion we shall take this as given and we shall be concerned exclusively with the distributional aspects of the LR (or LR-like) test statistics.

In view of the discussion above we need to find the distribution of

$$\bar{\chi}^2 = \min_{\beta \in C}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta); \tag{1.96}$$

without loss of generality we write Eq. (1.96) as

$$\bar{\chi}^2 = \min_{\beta \in C} [(\hat{\beta} - \beta_0) - (\beta - \beta_0)]' X'X [(\hat{\beta} - \beta_0) - (\beta - \beta_0)], \tag{1.97}$$

because in this case $C = C - \beta_0$ due to the fact that C contains the linear space $\{\beta : R\beta = 0\}$. Finally, the equation above may be rewritten as

$$\bar{\chi}^2 = \min_{\zeta \in C} (z - \zeta)'[X'X](z - \zeta), \quad z \sim N(0, (X'X)^{-1}). \tag{1.98}$$

Minimization in Eq.(1.98) involves the **projection** of z onto the convex polyhedron C. By the discussion in the appendix, C is covered by the relative interiors of its faces

$$ri(F_J) = \{\beta : R_J \beta = 0, \;\; R_J^* \beta > 0\}, \quad J = 0, 1, 3, \ldots, k. \tag{1.99}$$

Note that for $J = 0$ the relative interior of the face $R\beta \geq 0$, is $R\beta > 0$. The projection in question may be represented by a **symmetric idempotent matrix** because it involves the projection of z on the **linear space generated by the face**. Because it is somewhat awkward in this context to deal with a $N(0, (X'X)^{-1})$ variable, we use the decomposition $S_2' S_2 = X'X$, where S_2 is a nonsingular matrix, so that the minimand of Eq. (1.98) may be rendered as

$$\bar{\chi}^2 = \min_{\zeta \in C} (z^* - \zeta)'(z^* - \zeta), \quad z^* = S_2 z \sim N(0, I_m), \tag{1.100}$$

which is justified by the fact that the transformation $S_2 : R^m \rightarrow R^m$ is one to one. Observe that if the projection is on the linear space generated by the face F_J noted above, then the estimator is of the form

$$\tilde{\zeta} = P_J z^*, \;\; \text{so that} \;\; \bar{\chi}^2 | \bar{\chi}^2 \in F_J = z^{*\prime}[I_m - P_J] z^* \sim \chi^2_{m - \dim F_J}, \tag{1.101}$$

where P_J is a symmetric idempotent matrix of rank dim (F_J).

To complete this discussion we need to determine the dimension of the linear space spanned by the face F_J. Proceeding in a somewhat more general fashion let the face be denoted by

$$F_{k_1} = \{x : R_{k_1} x = 0, \; R_{k_2} x \geq 0\}, \tag{1.102}$$

where, without loss of generality we have taken the zero (equality) restrictions to correspond to the first k_1, and the positive (inequality) restrictions to correspond to the last k_2 restrictions, $k_1 + k_2 = k = \text{rank}(R)$. Note that if $k_2 = 0$

we have the problem of previous sections, and the distribution of the test statistic in Eq.(1.101) is **central** χ^2_{m-k} ; if $k_1 = 0$, then $F_{k_1} = C$, and the case will be dealt in the context of

Proposition 4. The dimension of the linear space spanned by F_{k_1} is $m - k_1$.

Proof. The face is given by

$F_{k_1} = \{x : R_1 x = 0,\ R_2 x \geq 0\}$, where $R_i,\ i = 1, 2$ are, respectively, $k_1 \times m,\ k_2 \times m$.

Partition

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}, \tag{1.103}$$

such that R_{11} is $k_1 \times k_1$ and, without loss of generality, **nonsingular**. The face then is described by the equations

$$R_{11}x^{(1)} + R_{12}x^{(2)} = 0, \quad R_{21}x^{(1)} + R_{22}x^{(2)} \geq 0, \tag{1.104}$$

where we have partitioned x **conformably** so that the $x^{(i)}, i = 1, 2,$ are $k_1 \times 1$ and $m - k_1 \times 1$, respectively. But Eq. (1.104) implies

$$x^{(1)} = -R_{11}^{-1} R_{12} x^{(2)}, \quad (R_{22} - R_{21}R_{11}^{-1}R_{12})x^{(2)} > 0. \tag{1.105}$$

We now establish that the rank of $R_{22} - R_{21}R_{11}^{-1}R_{12}$ is k_2 ; note that the matrix is $k_2 \times m - k_1$, $k_2 \leq m - k_1$, and consider

$$BR = \begin{bmatrix} I_{k_1} & 0 \\ -R_{21}R_{11}^{-1} & I_{k_2} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix}. \tag{1.106}$$

Since B is **nonsingular** and R is for rank k , it follows that

$$R_{22.1} = R_{22} - R_{21}R_{11}^{-1}R_{12}, \quad \text{is of rank} \quad k_2. \tag{1.107}$$

Define

$$S = \begin{bmatrix} -R_{11}^{-1}R_1 \\ I_{m-k_1} \end{bmatrix}$$

so that it is $m \times m - k_1$ of rank $m - k_1$; thus, its $m - k_1$ linearly independent columns generate a linear space, L(S),of dimension $m - k_1$, which is a subspace of R^m . If $\beta \in L(S)$ then $\beta = S\alpha$, and thus

$$R\beta = RS\alpha = \begin{bmatrix} 0 \\ R_{22.1}\alpha \end{bmatrix}, \tag{1.108}$$

which, thus, describes the linear space of dimension $m - k_1$ generated by the face F_{k_1} .

q.e.d.

We conclude, therefore, that $\bar{\chi}^2$ of Eq. (1.101) obeys

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \leq \tau | \bar{\chi}^2 \in F_{k_1}) \Pr(\bar{\chi}^2 \in F_{k_1}) = \Pr(\chi^2_{k_1} \leq \tau) \Pr(\bar{\chi}^2 \in F_{k_1}).$$

For simplicity of exposition we shall always use the form of the last equation above. Finally, note that this formulation implies that C is cobverd by the relative interiors of F_{k_1}, $k_1 = 1, 2, 3, \ldots, k$.

1.4.3 The case $H_0 : R\beta = 0$, as against $H_1 : R\beta \geq 0$

For this problem, put

$$C = \{\beta : R\beta = 0\}, \; K = \{\beta : R\beta \geq 0\}, \; C \subset K, \text{ and note that } C \text{ is a linear space.} \tag{1.109}$$

Consequently

$$K = C \oplus K \cap C^{\perp}, \tag{1.110}$$

and we have that

$$P(y|K) = P(y|C) + P(y|K \cap C^{\perp}), \quad P(y|C) \perp P(y|K \cap C^{\perp}). \tag{1.111}$$

The LR or LR-like statistic is given by

$$\min_{R\beta=0} (y - X\beta)'(y - X\beta) - \min_{R\beta \geq 0} (y - X\beta)'(y - X\beta)$$

$$= ||P(y|K) - P(y|C)||^2 = ||P(y|K \cap C^{\perp})||^2. \tag{1.112}$$

Since $K \cap C^{\perp}$ is a convex polyhedral cone the entity in the last member of Eq. (1.112) is distributed as

$$\bar{\chi}^2 = ||P(y|K \cap C^{\perp})||^2, \tag{1.113}$$

and

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\bar{\chi}^2 \leq \tau | \bar{\chi}^2 \in F_{k_1})$$

$$= \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\chi^2_{m-k_1} \leq \tau), \tag{1.114}$$

where F_{k_1}, $k_1 : k_1 = 0, 1, 2, \ldots k$ are the (relative interiors of the) faces that cover $K \cap C^{\perp}$.

To verify this, note that $C = \{\beta : R\beta = 0\}$ is the **null space** of R, which is of dimension $m - k$, see Dhrymes (2000), p. 18. The vectors **orthogonal** to it (C) are given by

$$C^{\perp} = \{\beta : \beta = R'\alpha : \alpha \in R^k\}, \text{ so that if } \beta \in C^{\perp}, x \in C, \ \beta' x = \alpha' R x = 0. \tag{1.115}$$

Thus,

$$K \cap C^{\perp} = \{\beta : R_{k_1}\beta = 0, R_{k_2}\beta > 0\}, \quad 0 \leq k_1 < k. \tag{1.116}$$

This is so since $C^{\perp} = \{\beta : \beta = R'\alpha, \ \alpha \in R^k$, and there exist vectors α, such that

$$R\beta = RR'\alpha = \begin{bmatrix} R_1\beta = 0 \\ R_2\beta > 0 \end{bmatrix}, \tag{1.117}$$

which can be demonstrated as follows: to ease notational burden, partition

$$R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}, \ R_i, \ i = 1, 2, \text{ with rank } k_i, \ i = 1, 2, \quad k = k_1 + k_2.$$

Note that since R' is $m \times k$ of **rank** k, $C^{\perp}$ **does not contain the zero vector**, except for the trivial case $\alpha = 0$. To demonstrate the validity of Eq. (1.117), consider

$$R\beta = RR'\alpha = \begin{bmatrix} R_1R_1' & R_1R_2' \\ R_2R_1' & R_2R_2' \end{bmatrix} \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \end{bmatrix} = \begin{bmatrix} R_1R_1'\alpha^{(1)} + R_1R_2'\alpha^{(2)} \\ R_2R_1'\alpha^{(1)} + R_2R_2'\alpha^{(2)} \end{bmatrix}, \tag{1.118}$$

note that R_1R_1' is a **nonsingular** matrix of rank k_1, R_2R_2' is a **nonsingular** matrix of rank k_2, and set the first sub-vector in the rightmost member above to zero, to obtain

$$R\beta = RR' \begin{bmatrix} -(R_1R_1')^{-1}R_1R_2' \\ I_{k_2} \end{bmatrix} \alpha^{(2)} = \begin{bmatrix} 0 \\ R_{22.1}\alpha^{(2)} \end{bmatrix}, \tag{1.119}$$

where $R_{22.1} = R_2R_2' - R_2R_1'(R_1R_1')^{-1}R_1R_2'$. Since it may be shown that $R_{22.1}$ is **nonsingular**, $\alpha^{(2)}$ may be chosen so that $R_{22.1}\alpha^{(2)} > 0$. By the conditions of the problem $\alpha^{(2)} \neq 0$, so that we have produced an element of $C^{\perp}$, β, such that

$$R\beta = \begin{bmatrix} R_1\beta = 0 \\ R_2\beta > 0 \end{bmatrix}.$$

Hence the space $K \cap C^{\perp}$ consists of vectors of the form

$$\{\beta : R_{k_1}\beta = 0, R_{k_2}\beta > 0\}, \quad k_1 = 0, 1, \ldots, k - 1, \tag{1.120}$$

and is thus a **convex polyhedral cone**. Consequently, the entities in Eq. (1.120) describe the relative interiors of its faces. From previous discussion these faces are of dimension $m - k_1$, and we conclude that

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k-1} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\chi^2_{m-k_1} \leq \tau), \;\; k_1 + k_2 = k, \tag{1.121}$$

as required. The notation above requires some clarification which we shall not repeat in later discussion. Notice that if the sequence $(1, 2, 3, \ldots, k_1)$ is fixed one would gather by the notation that this is the only term involved in the summation. However, this would not be correct since a face of dimension $m-k_1$ can also be obtained in $\binom{k}{k_1}$ ways, so that the term $\Pr(\bar{\chi}^2 \in F_{k_1})$ refers to the probability of being **in any** of the $\binom{k}{k_1}$ faces of dimension $m - k_1$. The second term on the rightmost meber of Eq. (1.121) is correctly stated since the chi-squared therein is the appropriate one **for any face** of $dim(F) = m - k_1$. The more correct but cumbersome notation would be

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \left(\sum_{\text{over all faces with } dim(F)=m-k_1} \Pr(\bar{\chi}^2 \in F_{k_1}) \right) \Pr(\chi^2_{m-k_1} \leq \tau). \tag{1.122}$$

Remark 3. The reader should bear in mind that even though we use the notation F_{k_1} to designate a face of a polyhedron, the term is to be understood as an **equivalence class**. The notation for a face, e.g., $F_{k_1} = \{x : R_{k_1}x = 0, \; R_{k_2}x \geq 0\}$ denotes an entity with k_1 equalities and k_2 inequalities. But even if k_1 **is given** the number of equality restrictions can be taken in $\binom{k}{k_1}$ ways, so that there are $\binom{k}{k_1}$ such faces!

In econometrics, the restrictions are typically stated in affine form, for example, we put $H_0 : R\beta = r$, as against $H_1 : R\beta \geq r$. Or the null may be $R_1\beta = r_1, R_2\beta \geq r_2$ and the alternative that β is unrestricted. In the analytical context (Lagrange multipliers) employed earlier, the transition from r equal to zero to r known but unspecified in terms of sign occasioned a considerable adjustment to the procedure of estimating and/or testing the parameters of such models; this transition, by contrast, is nearly painless in the projection context as the following discussion will demonstrate.

1.4.4 $H_0 : R\beta = r$ as against, $H_1 : R\beta \geq r$

The LR or LR-like statistic is given by

$$-2\ln\lambda = \min_{R\beta = r}(y - X\beta)'(y - X\beta) - \min_{R\beta \geq r}(y - X\beta)'(y - X\beta), \tag{1.123}$$

and rejects the null for large values of the statistic above. As we have done before, add and subtract $X\hat{\beta}$ to render the right member of Eq. (1.123) as

$$-2\ln\lambda = \min_{R\beta = r}(z-\beta)'X'X(z-\beta) - \min_{R_2\beta \geq r_2}(z-\beta)'X'X(z-\beta), \;\; z \sim N(\beta_0, (X'X)^{-1}). \tag{1.124}$$

Writing $z - \beta = (z - \beta_0) - (\beta - \beta_0)$, and treating z as $z - \beta_0$, and β as $\beta - \beta_0$ we can finally represent the statistic of Eq.(1.123) as

$$-2\ln\lambda = \min_{R\beta = 0}(z-\beta)'X'X(z-\beta) - \min_{R\beta \geq 0}(z-\beta)'X'X(z-\beta), \;\; z \sim N(0, (X'X)^{-1}), \tag{1.125}$$

where now $z = z - \beta_0$ and $\beta = \beta - \beta_0$, so that in Eq. (1.125) $z \sim N(0, (X'X)^{-1})$ as claimed. The expression follows from the fact that $R\beta - R\beta_0 = 0$ under the null, and **equal to or greater than zero under the alternative**. Putting

$$C = \{\beta : R\beta = 0\}, \quad K = \{\beta : R\beta \geq 0\}, \; C \subset K, \tag{1.126}$$

and noting that C is a **linear space**, we have precisely the problem of the previous subsection.

1.4.5 The case $H_0 : \beta = 0$, as against $H_1 : R\beta \geq 0$

A closely related problem to the one dealt with in subsection 1.4.2 is whether the parameter θ merely obeys certain linear (inequality) constraints or whether the parameter itself is zero.

The likelihood ratio test rejects the hypothesis $\beta = 0$, for large values

$$\begin{aligned} -2\ln\lambda &= y'y - \min_{\beta \in C}(y - X\beta)'(y - X\beta) = y'y - (y - X\tilde{\beta})'(y - X\tilde{\beta}) \\ &= \tilde{\beta}'X'X\tilde{\beta}, \quad C = \{\beta : R\beta \geq 0\}. \end{aligned} \tag{1.127}$$

This is so because $X\tilde{\beta}$, being a projection, obeys

$$(y - X\tilde{\beta})'X\tilde{\beta} = 0, \quad (y - X\tilde{\beta})'y = y'y - \tilde{\beta}'X'X\tilde{\beta},$$

or, if one prefers, because of i of Proposition 3. From previous discussion we thus conclude that

$$-2\ln\lambda = P(y|C) = \bar{\chi}^2, \quad \text{and}$$

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\chi^2_{m-k_1} \leq \tau). \tag{1.128}$$

Remark 4. It is instructive to examine the similarity of the results obtained from (linear) **inequality** constraints with those obtained from **linear equality** constraints. The latter were obtained at an earlier section by analytical (Lagrange multiplier) methods while the former were obtained using projection theory. The comparison will make clear that whether we are using Lagrange multipliers or projection methods we get identical results, *mutatis mutandis* because we are carrying out precisely the same task using two different modalities or approaches; the projection approach is more general, or flexible, in that it is easily extended to inequality constraints, while the Lagrange multiplier approach entails far greater complexity.

In particular we wish to compare the results obtained in subsection 1.2, dealing with linear equality constraints and those of subsections 1.4.1 through 1.4.5.

We begin by noting that when we assume that the variance (of the

error process) is known the rightmost term of the likelihood ratio of Eq. (1.5) is irrelevant and we are dealing with

$$\lambda = \frac{\max_{\theta|H_0} L(\theta)}{\max_{\theta|H_1} L(\theta)},$$

so that for the problem of subsection 1.4.1 we have

$$\begin{aligned} \lambda &= e^{-\frac{1}{2}[(X'X/T) - \frac{1}{T}(X - e\tilde{\theta})'(X - e\tilde{\theta})]} = e^{-\frac{1}{2}[(X'X/T) - (X'X/T) - \tilde{\theta}'(e'e/T)\tilde{\theta}]} \\ &= e^{-(1/2)\tilde{\theta}'\tilde{\theta}}. \end{aligned} \tag{1.129}$$

Thus,

$$-2\ln\lambda = \tilde{\theta}'\tilde{\theta} \tag{1.130}$$

as claimed in Eq.(1.86). The same can be done in subsections 1.4.2 through 1.4.5, but in the interest of brevity we shall omit the demonstration.

Next we recall in summary form the results regarding the test statistics; note that the null is always the more restrictive hypothesis and the alternative the less restrictive:

i. $H_0 : R\beta \geq 0$, $H_1 : \beta$ unrestricted.

Test statistic: $(\hat{\beta} - \tilde{\beta})' X' X (\hat{\beta} - \tilde{\beta}) = \bar{\chi}^2$, where $\hat{\beta}$, $\tilde{\beta}$ are, respectively, the unrestricted and restricted estimators.

Its critical region is determined by

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\chi^2_{k_1} \leq \tau).$$

ii. $H_0 : R\beta = 0$, $H_1 : R\beta \geq 0$.

$$-2\ln\lambda = ||P(y|K \cap C^{\perp})||^2 = \bar{\chi}^2, \tag{1.131}$$

whose distribution and critical points are given by

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\chi^2_{m-k_1} \leq \tau). \tag{1.132}$$

This does not have a counterpart for the linear case since then it would read $H_0 : R\beta = 0$, $H_1 : R\beta = 0$.

iii. $H_0 : R\beta = r$, $H_1 : R\beta \geq r$. This case is exactly the same as the one in item ii.

iv. $H_0 : \beta = 0$, $H_1 : R\beta \geq 0$.

Likelihood ratio test statistic and distribution

$$-2\ln\lambda = P(y|C) = \tilde{\beta}' X' X \tilde{\beta} = \bar{\chi}^2$$

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{k_1=0}^{k} \Pr(\bar{\chi}^2 \in F_{k_1}) \Pr(\chi^2_{m-k_1} \leq \tau). \tag{1.133}$$

Although in subsection 1.2 for the standard GLM (with known error variance, which we normalize to one), the problem was set up with restrictions of the form $R\beta = r$, in the following discussion we shall take $r = 0$ for maximum compatibility. Notice first that items ii and iii do not have a counterpart in the context of **equality constraints**. Thus, our comparison with deal with cases i and iv. The analog of i dealt with in that subsection may be formulated as

$$\min_{R\beta=0} (y - X\beta)'(y - X\beta) = (y - X\hat{\beta})'(y - X\hat{\beta}) + \min_{R\beta=0} (\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta), \tag{1.134}$$

where $\hat{\beta}$ is the unrestricted estimate of β; thus we need only operate with the second term in the right member. Moreover the LR or LR-like statistic for testing the hypothesis $R\beta = 0$ **as against** $R\beta \neq 0$ **, i.e.** β **unrestricted**, is given by

$$-2\ln\lambda = \min_{R\beta=0}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta) - \min_{R\beta\neq 0}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta). \tag{1.135}$$

Notice that the second term in the right member of E q. (1.135) is zero! Thus, the LR test statistic is given by

$$-2\ln\lambda = (\tilde{\beta}-\hat{\beta})'X'X(\tilde{\beta}-\hat{\beta}) = u'X(X'X)^{-1}R'DR(X'X)^{-1}R'DR(X'X)^{-1}X'u$$

$$= u'X(X'X)^{-1}R'DR(X'X)^{-1}X'u \sim \chi^2_k. \tag{1.136}$$

This is so because solving the problem by Lagrange multipliers yields

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'DR\hat{\beta},\ D = [R(X'X)^{-1}R']^{-1},\ \text{or}$$

$$(\tilde{\beta}-\beta_0) = [I_m - (X'X)^{-1}R'D]R(\hat{\beta}-\beta_0).$$

This fully corresponds to the case in i, if we take $\Pr(\bar{\chi}^2 \in F_{k_1}) = 0$, for $k_1 \neq k$ and equal to one for $k_1 = k$!

To address the problem in iv note that

$$\tilde{\beta}-\beta_0 \sim N(0,\Phi),\ \Phi = (X'X)^{-1} - (X'X)^{-1}R'DR(X'X)^{-1}, \tag{1.137}$$

the matrix Φ obeys $\Phi R' = 0$, and is thus of rank $m-k$. Under the null of part iv,

$$X\tilde{\beta} \sim N(0, X\Phi X'),\quad X\Phi X' = M_1 - M_2,$$

$$M_1 = X(X'X)^{-1}X',\ M_2 = X(X'X)^{-1}R'DR(X'X)^{-1}X', \tag{1.138}$$

$M_1,\ M_2$ are both symmetric idempotent matrices as is $M_1 - M_2$; the rank of the latter is $m-k$. Hence, see Dhrymes (2000), pp. 73-78 there exists **an orthogonal matrix**, Q such that

$$X\Phi X' = Q\begin{bmatrix} I_{m-k} & 0 \\ 0 & 0 \end{bmatrix}Q', \tag{1.139}$$

and thus

$$Q'X\tilde{\beta} \sim N\left(0, Q'Q\begin{bmatrix} I_{m-k} & 0 \\ 0 & 0 \end{bmatrix}QQ'\right) = N\left(0, \begin{bmatrix} I_{m-k} & 0 \\ 0 & 0 \end{bmatrix}\right). \tag{1.140}$$

Consequently, a test of the hypothesis $H_0 : \beta = 0$ as against the alternative $H_0 : R\beta = 0$ can be carried out through the statistic

$$\eta = \tilde{\beta}'X'QQ'X\tilde{\beta} = \tilde{\beta}'X'X\tilde{\beta} \sim \chi^2_{m-k}. \tag{1.141}$$

If we compare this with the test statistic and its distribution in item iv, for the case $k_1 = k$, we see that the two are identical because in such a case

$$\Pr(\bar{\chi}^2 \in F_{k_1}) = 0, \text{ for } k_1 \neq k, \text{ and } \Pr(\bar{\chi}^2 \in F_{k_1}) = 1, \text{ for } k_1 = k.$$

1.4.6 The case $H_0 : R_1\beta = r_1, R_2\beta \geq r_2$, as against $H_1 : \beta$ unrestricted

.

In this context R_1 is $k_1 \times m$, R_2 is $k_2 \times m$, $k = k_1 + k_2$, and r has been partitioned conformably by $r_1,\ r_2$.

Put

$$C = \{\beta : R_1\beta = r_1,\ R_2\beta \geq r_2\}, \tag{1.142}$$

which is thus a closed convex polyhedral cone; consequently, the problem may be handled by the procedures developed above. There is, however, an additional difficulty **because the null is composite** and thus there is **no single distribution** under the null. In such cases we use the device we had used in an earlier chapter when we considered the problem of the UMP test in the scalar case $\theta \leq 0$ as against $\theta > 0$. Here we use the **least favorable null distribution** as explained in Lehmann (1959), p. 91. More specifically we behave as if the null is $R\beta_0 = r$. Speaking somewhat loosely, notice that if the null so stated is **accepted**, with level of significance α, then *a fortiori* any other null contained in the composite null will also be accepted. Also the power of this test, i.e. the probability of rejecting a false null, is highest within the class of tests with level of significance equal to or less than α. On the other hand it is possible that on some sample information the null so stated will be rejected but other nulls contained in the composite may well be accepted. Hence the term **least favorable**.

With the least favorable null stated as $R\beta_0 = r$ and proceeding as in the previous subsection(s) the LR test rejects the null for large values of

$$\begin{aligned} -2\ln\lambda &= \min_{\beta \in C}(y - X\beta)'(y - X\beta) - \min_{\beta \in R^m}(y - X\beta)'(y - X\beta) \\ &= \min_{\beta \in C}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \\ &= \min_{\beta \in C}((\hat{\beta} - \beta_0) - (\beta - \beta_0))'S_2'S_2(\hat{\beta} - \beta_0), \end{aligned} \tag{1.143}$$

where

$$C = \{\beta : R_1\beta = 0, R_2\beta \geq 0\}, \;\; S_2'S_2 = X'X, \quad X \;\; \text{nonsingular.} \tag{1.144}$$

Since the (relative interiors of the) faces of C are of the form

$$F_i = \{x : R_{k_1+i}\beta = 0, \;\; R_{k_2-i}\beta > 0, \;\; i = 0, 1, 2, \ldots, k_2, \tag{1.145}$$

the linear space generated by F_{k_1+i} is of **dimension** $m - k_1 - i$, and the test statistic becomes

$$-2\ln\lambda = \bar{\chi}^2 = \min_{R\beta\geq 0}(z-\beta)'X'X(z-\beta), \quad S_2(\hat{\beta}-\beta_0) = z \sim N(0, I_m), \;\text{and}$$

$$\Pr(\bar{\chi}^2 \leq \tau) = \sum_{i=0}^{k_2} \Pr(\bar{\chi}^2 \in F_{k_1+i})\Pr(\chi^2_{k_1+i} \leq \tau) \tag{1.146}$$

1.4.7 Asymptotics and Miscellaneous

Asymptotics

The developments above need not rely on the assumption that the $y's$ or $u's$ are normally distributed; we only need to assert that their properties are such that a central limit theorem (CLT) applies. We have already done or hinted at asymptotics for the case $H_0 : \theta = 0$, $H_1 : \theta > 0$, so here we confine our discussion to the general linear model

$$y = X\beta + u, \quad u_t : t = 1, 2, 3, \ldots, T, \;\text{i.i.d.} \tag{1.147}$$

with mean zero and variance one. All problems dealt with involve

$$\min_{\beta\in C}(y-X\beta)'(y-X\beta) = (y-X\hat{\beta})'(y-X\hat{\beta}) + \min_{\beta\in C}(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta), \tag{1.148}$$

where $\hat{\beta} = (X'X)^{-1}X'y$ and thus

$$\sqrt{T}(\hat{\beta}-\beta_0) = \left(\frac{X'X}{T}\right)^{-1}\frac{X'u}{\sqrt{T}}. \tag{1.149}$$

Using Proposition 45 in Dhrymes (1989), pp. 265-276, and noting that the Linderberg condition holds under the standard assumptions imposed on X in econometrics, we conclude

$$\sqrt{T}(\hat{\beta}-\beta_0) \xrightarrow{\text{d}} N(0, M_{xx}^{-1}), \;\; M_{xx} = \underset{T\to\infty}{\text{plim}}\frac{X'X}{T}, \tag{1.150}$$

which shows in the second term of the right member of Eq. (1.149) we could, for large samples, take $\hat{\beta}-\beta_0$ to be normal with mean zero and **known** covariance matrix!

Estimating Pr($\bar{\chi}^2 \in F_{k_1}$)

One important topic not discussed in the developments above is the determination of the term $\Pr(\bar{\chi}^2 \in F_{k_1})$. This issue was addressed *inter alia* by Shapiro (1985), (1988), (1989) but no general solution exists. On the other hand the availability of relatively inexpensive computer power enables us to fully implement the testing procedure we have developed earlier. To this end, consider any of the problems dealt with above. Carry out the optimizing, quadratic or nonlinear, program for a large number of times, say $N = 10,000$, or more if desired. Count the number of times the test statistic $\bar{\chi}^2$ yields k_1 zero constraints (i.e. $R_{k_1}\beta = 0$) and k_2 positive constraints, i.e. $R_{k_2}\beta > 0$ and divide by N ; this is the estimated $\Pr(\bar{\chi}^2 \in F_{k_1}$, which thus completes our discussion of this subject.

APPENDIX

Miscellaneous Mathematical Topics

Spaces

In this appendix we present a number of results that are useful in dealing with constrained parameter estimation and hypothesis testing, and which we have used repeatfedly in the discussion above.

Definition 1. A **metric** space is a set $\mathcal{S}$ equipped with a (global) distance function, d, such that for any two points $x, y \in \mathcal{S}$ $d(x, y)$ gives the distance between x and y as a positive real number. The function d, or **metric**, has the following properties: for any $x, y, z \in \mathcal{S}$:

i. $d(x, y) = d(y, x)$;

ii. $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$;

iii. the triangle inequality holds, i.e.

$$d(x, z) \leq d(x, y) + d(y, z).$$

In general we shall be dealing with real entities, so that unless otherwise indicted, the underlying space is R^n. Often we shall also use the notation $\mathcal{S}$.

Definition 2. A linear space (or vector space), over a field of scalars, $\mathcal{F}$, is a non-empty collection of elements, say $a \in R^n$, which is closed under addition and scalar multiplication. Notice that a non-empty vector space automatically contains the null set, $\emptyset$ and the sample space, i.e. the space to which the elements belong, here R^n.

Definition 3. A (non-empty) subspace of a vector space is iself a vector space.

Definition 4. Let V be the vector space above; the **dimension** of V, denoted by $dim(V)$ is the **maximal number of linearly independent**(a term to be defined below) elements it contains.

Definition 5. A **normed** linear space is a linear space on which we have defined a **norm**, $\| \cdot \|$, which obeys, for any elements x, y in the space and any scalar $a \in R$,

i. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$;

ii. $\|x + y\| \leq \|x\| + \|y\|$, the triangle inequality;

iii. $\|ax\| = |a|\|x\|$.

Definition 6. A normed linear space (or a metric space) is said to be complete if all Cauchy sequences converge. Such a space is to be a Banach space.

Definition 7. A Hilbert space, H, is a Banach space on which we have defined an additional operation $(\cdot, \cdot)$ termed the **inner product** which obeys: for any elements $x, y, z \in H$ and scalars $a, b \in \mathcal{F}$,

i. $(ax + by, z) = a(x, z) + b(y, z)$;

iii. $(x, y) = (y, x)$;

iii. $(x, x) = \|x\|^2$.

For more details see Dhrymes (1989), pp. 64-73.

Note that a normed linear space can be easily converted to a metric space by simply taking the metric to be

$$d(x, y) = \|x - y\|^2.$$

Example 1. Let $S = R^n$ and consider a (non-empty) subspace S^* such that the elements of S^* are n-tuples with elements in R^n (real vectors); under the usual definition for vector and matrix multiplication, S^* is certainly a linear space. If we define the usual Euclidean norm

$$\|x\| = \sqrt{\sum_{i=1}^{m} x_i^2},$$

then S^* becomes a normed linear space. It may be shown that it is a complete normed linear space or a Banach space; if it is a Banach space then the usual rules for multiplication of vectors and matrices render it a Hilbert space. To render it a **metric** space we can take $d(x, y) = \|x - y\|^2$.

Example 2. Consider S as above and attempt to define a norm by

$$\|x\|^2 = x'Dx, \quad D \text{ a positive definite matrix of order } n.$$

We shall term this the D-based norm and denote it by $\| \cdot \|_D$ It is easily seen that the entity above satisfies all requirements for a norm, except possibly for the triangle inequality (TI).

To show that it satisfies the TI as well and, consequently, can serve as a norm, recall from Dhrymes (2000), p. 86, that there exists a **nonsingular** matrix W such that $D = W'W$. Thus, consider the transformation of S onto S^* , such that

$$S^* = \{x^* : x^* = Wx, \quad x \in S\}. \tag{1.151}$$

Thus, this transformation is a **homeomorphism**. In the context of S^* the usual Euclidean inner product $(\cdot, \cdot)$ can evidently serve as a norm, i.e.

$$(x_1^*, x_2^*) = \sum_{j=1}^{n} x_{1j}^* x_{2j}^*, \quad \|x_1^*\|^2 = (x_1^*, x_1^*). \tag{1.152}$$

We shall now show that the D-based norm on S satisfies the TI. Since

$$\|x_1^* + x_2^*\|^2 \leq (\|x_1^*\| + \|x_2^*\|)^2,$$

we have

$$\|x_1^* + x_2^*\|^2 = \|x_1 + x_2\|_D^2 \leq (\|x_1^*\| + \|x_2^*\|)^2 = (\|x_1\|_D + \|x_2\|_D\|)^2, \tag{1.153}$$

which, upon taking square roots, demonstrates that the D-based norm on S is, indeed, a norm.

Sets

Definition 8. Let $\mathcal{S}$ be a space and $S \subset \mathcal{S}$, non-empty. The set S is said to be **closed** (in $\mathcal{S}$), if and only if for every $s \in S$ it is not true that there exists a neighborhood of s , however small, lying entirely in S . A set that is not closed is said to be open.

For example the set

$$S = \{(x, y) : ax + by \geq c\}$$

is **closed** because for $s = (x, \frac{c}{b} - \frac{a}{b}x\}$ **no neighborhood** of s, say $N(s, \epsilon)$ lies entirely in S, for arbitrarily small $\epsilon > 0$.

On the other hand $S^* = \{(x, y) : ax + by > c\}$ is **open** since for every $s \in S^*$ we can find a neighborhood of s that lies entirely in S^*.

Dfinition 9. A set $S \in \mathcal{S}$ is said to be a cone with vertex s_0 if $s_0 + c(s - s_0) \in S$, whenever $s \in S$, and $c > 0$. More discussion is given below.

Definition 10. Let $\mathcal{S} = R^2$; an ellipse in $\mathcal{S}$ with center at (x_0, y_0) is defined as the set of points $(x, y) \in R^2$ such that the sum of their distances from the foci (c_1, c_2) is constant. The distance between c_1 and c_2 is c. This property describes a set whose analytic representation is

$$\sqrt{x - c_1)^2 + (y - y_0)^2} + \sqrt{(x - c_2)^2 + (y - y_0)^2} = 2a. \tag{1.154}$$

Noting that $c_1 = (x_0 - c, y_0)$ and $c_2 = (x_0 + c, y_0)$, because c_1, c_2 are equidistant from the center (x_0, y_0), along the x-axis, Eq. (1.26) can also be written

$$\sqrt{(x - x_0 + c)^2 + (y - y_0)^2} + \sqrt{x - x_0 - c)^2)^2 + (y - y_0)^2} = 2a, \tag{1.155}$$

where a is a parameter determining the semimajor axis of the ellipse. Clearing the root signs and simplifying, we obtain the standard representation of an ellipse with center at (x_0, y_0) as

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} = 1, \tag{1.156}$$

where a is the semimajor axis, b is the semiminor axis, and $b^2 = a^2 - c^2$.

Definition 11. In three dimensions, the analog of an ellipse is termed an **ellipsoid**, whose canonical representation is

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} + \frac{(z - z_0)^2}{c^2} = 1. \tag{1.157}$$

Remark 1. The generalization to n dimensions is given in canonical form by

$$\sum_{i=1}^{n} a_i^2 (x_i - x_{i(0)})^2 = 1, \tag{1.158}$$

where now the axes are labeled $x_i, i = 1, 2, \ldots, n$ instead of x, y, z etc., and x_0 is the center of the ellipsoid.

By the usual convention we should call the construct of Eq. (1.61) a hyper-ellipsoid, but for simplicity we shall refer to it as the n-ellipsoid and, when the context prevents confusion, simply as ellipsoid.

It is interesting to note that ellipsoids are quite useful in (multi-dimensional) hypothesis testing and confidence intervals, when the underlying distribution is normal. For example, suppose we have an estimator of some n-dimensional parameter, say $\hat{\theta}$. Suppose further that

$$\sqrt{T}(\hat{\theta} - \theta_0) \to N(0, V^{-1}). \tag{1.159}$$

Confidence intervals and hypothesis testing are based on the ellipsoid (including its interior)

$$T(\hat{\theta} - \theta_0)' V (\hat{\theta} - \theta_0) \leq c, \tag{1.160}$$

where c is a suitable scalar, determined by the level of significance, and V is a known **positive definite** matrix. Consider then the collection

$$A(\theta) = \{\theta : T(\hat{\theta} - \theta)' V (\hat{\theta} - \theta) \leq c\}. \tag{1.161}$$

If the hypothesized parameter, say θ^*, is in A we accept the hypothesis that $\theta = \theta^*$; if not, we reject. But the set $A(\theta)$ is the ellipsoid (including its interior) $T(\hat{\theta} - \theta)' V (\hat{\theta} - \theta) = c$ with center $\hat{\theta}$. This is so because V, being a positive definite matrix, has a decomposition $V = Q \Lambda Q'$, see Dhrymes (2000), p. 86, where Q is the **orthogonal matrix** of the characteristic vectors and Λ is the **diagonal matrix** of the (positive) characteristic roots. Hence employing the orthogonal, norm preserving, transformation

$$\hat{\eta} - \eta = Q(\hat{\theta} - \theta), \tag{1.162}$$

we can write the boundary of the acceptance set as

$$\partial A(\eta) = (\hat{\eta} - \eta)' \Lambda (\hat{\eta} - \eta) = \sum_{i=1}^{n} \lambda_i (\hat{\eta} - \eta)^2 = c, \quad \lambda_i > 0, \tag{1.163}$$

which is in canonical form.

Definition 12. A set $S \subseteq \mathcal{S} = R^n$ is said to be an **affine set** if and only if $x, y \in S$ implies $z(\gamma) \in S$ for any $\gamma \in R$, where $z(\gamma) = \gamma x + (1 - \gamma) y$.

Remark 2. It is useful, at this stage, to point out a few facts about affine sets.

i. The subspaces of R^n are the affine sets that contain the origin.

iii. Every non-empty affine set can be expressed uniquely as

$M = L+a$, L a (linear) subspace and $a \in R$, i.e. M is a translate of L.

This is referred to as: M is parallel to the subspace L.

iv. $L = \{z : z = x - y,\ x, y \in M\}$.

v. The **dimension** of an affine set is the dimension of its associated linear subspace; in turn the latter is the number of linearly independent (a term to be defined below) elements contained in L.

For an elaboration of such results see Rockafellar (1970), p. 4-10.

Definition 13. A hyperplane in R^n is an affine set, for example

$$H_{sr} = \{x : s'x = r\},\ r \in R. \tag{1.164}$$

If we fix $s \in R^n$ and vary $r \in R$ we obtain a collection of hyper-planes which are simply translations of $H_{s0} = s^{\perp}$, because H_{s0} is the collection of points $x \in R^n$ which are **orthogonal to** s. Evidently, the H_{sr} are convex sets, for all r.

Definition 14. The term half space (of R^n) denotes collections of the form

$$\{x : x \in R^n,\ s'x \leq r\} \text{ closed subspace}, \ \ \{x : s'x < r\}, \text{ open subspace}. \tag{1.165}$$

We may of course define half spaces in terms of the operations $\geq$ and $>$. Note that a half space is a convex set; notice further that, by construction, the intersection of a finite number of convex sets is a convex set.

Definition 15. For $\{x_i : x_i \in R^n, \ \ i = 1, 2, 3, \ldots, m \leq n\}$, the entity $\sum_{i=1}^{m} a_i x_i$, $a_i \in R$ is said to be a **linear combination**.

Definition 16. Let $\{x_i : x_i \in R^n\}$, be as above; the entity $\sum_{i=1}^{m} a_i x_i$, $\sum_{i=1}^{m} a_i = 1$, is said to be an **affine combination**.

Definition 17. Let $\{x_i : x_i \in R^n\}$, be as above; the entity $\sum_{i=1}^{m} a_i x_i$, $\sum_{i=1}^{m} a_i = 1$, **and** $a_i \geq 0$ is said to be a **convex combination**.

Definition 18. Let $\mathcal{S} = R^n$ be a linear space, with the usual norm and metric definitions, and $S \in \mathcal{S}$; the **affine hull** of S, $AH(S)$, is the intersection of all affine manifolds[11] containing S; alternatively, given a set of points $\{p_i : p_i \in R^n, i = 1, 2, 3, \ldots, m \leq n\}$, their affine hull, AH, is given by the expression

$$AH(S) = \{p : p = \sum_{i=1}^{m} \lambda_i p_i, \quad \sum_{i=1}^{m} \lambda_i = 1\}.$$

The points $\{p_1, p_2, \ldots, p_m\}$ are said to be the **generators** of AH.

An important feature of affine sets is given by

Proposition 1. Given $a \in R^n$ and A an $m \times n$ real matrix the set

$$M = \{x : x \in R^n, Ax = a\}$$

is an affine set in R^n; moreover, every affine set in R^n can be represented in this fashion.

Proof. See Rockafellar (1970), pp. 5-6.

Definition 19. Let $\mathcal{S}$ be as in Definition 18, and $S \in \mathcal{S}$; the **convex hull** of S, $CH(S)$, is the intersection of all convex sets containing S; alternatively, given a set of points $\{p_i : p_i \in R^n, i = 1, 2, 3, \ldots, m \leq n\}$, their convex hull CH is given by the expression

$$CH(S) = \{p : p = \sum_{i=1}^{m} \lambda_i p_i, \quad \lambda_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{n} \lambda_i = 1\}.$$

The points $\{p_1, p_2, \ldots, p_m\}$ are said to be the **generators** of CH.

An interesting question that arises in connection with (affine and) convex sets is their **dimension**. This leads to

Definition 20. An affine set M is said to be parallel to another affine set L if and only if

$$M = L + a = \{x + a : x \in L\}, \ a \in R^n.$$

The notation means $M = \{y : y = x + a, x \in L\}$

[11] Given a set S the affine manifold associated with S is the set S and all affine combinations of its elements. In general an **affine set** is a set M such that $\lambda x + (1 - \lambda) \in M$, whenever $x, y \in M$ and $\lambda \in \mathcal{F}$. In most of our discussions $\mathcal{F} = R$. Some authors use the term affine cover, denoted by $affS$, for what we have denoted as $AH(S)$ and coS, for what we denoted as $CH(S)$.

Definition 21. The set $\{p_0, p_1, p_2, \ldots, p_m\}$ is said to be affinely linearly independent if and only if its affine hull is m-dimensional, i.e.

$$M = AH(p_0, p_1, p_2, \ldots, p_m) = L + p_0, \quad L = AH(0, p_1 - p_0, p_2 - p_0, \ldots, p_m - p_0).$$

Definition 22. Let the set $\{p_0, p_1, p_2, \ldots, p_m\}$ be affinely linearly independent; its convex hull is said to be an m-dimensional simplex, and the points p_i are said to be its vertices.

Proposition 2. Let C be a convex set; its dimension is the maximum of the dimensions of the various simplices contained in C.

Proof. See Rockafellar (1970), p 13.

Remark 3. Applying the framework above to the context of Propositions 1 and 2, we can write

$$M = AH(a, a_{\cdot 1}, a_{\cdot 2}, \ldots, a_m) = L + a, \tag{1.166}$$

where $L = AH(0, a_{\cdot 1} - a, a_{\cdot 2} - a, \ldots, a_m - a\}$. Then L is the smallest subspace containing $a_{\cdot 1} - a, a_{\cdot 2} - a, \ldots, a_n - a\}$, which are by assumption **linearly independent**. Now, the dimension of L is the largest set of **linearly independent** vectors in L. Thus, it follows that the **dimension** of C, denoted by $\dim C = \dim AH(C)$, is the number of (linearly) independent elements it contains.

Remark 4. If C is generated by $\{p_i : i = 1, 2, 3, \ldots, m \leq n\}$, and the latter are **linearly independent**, then $\dim AH(C) = \dim CH(S) = \dim C$, as is evidently implied by construction; a more general formulation of this result may be found, with proof, in Hiriart-Urruty and Lemarechal (HL) (1993), vol. I, pp. 102-106.

Definition 23. Let S be as in Definition 18 and let it contain the points $\{p_i : i = 1, 2, 3 \ldots m \leq n\}$

The set

$$C = \{p : p = \sum_{i=1}^{m} w_i p_i, \quad w_i \geq 0, \; m \leq n\} \tag{1.167}$$

is said to be the **cone** generated by the points $\{p_1, p_2, \ldots, p_m$. If, **in addition**, $\sum_{i=1}^{n} w_i = 1$, C is said to be a (closed) **convex cone**. Thus, the convex hull of a set of points in R^n is a (closed) convex cone. More simply a set, C, is said to be a cone if $kx \in C$ whenever, $x \in C$ and $k > 0$.

Definition 24. Let $\mathcal{S}, S$ be as in Definition 18. If the set S is of the form

$$S = \{x : x \in R^n, A'x \geq 0\}, \quad A \text{ a matrix of rank } k, \quad \text{and dimension } n \times k, \tag{1.168}$$

then S is said to be a (closed) polyhedral cone.

Note that S is not bounded, but it is closed and convex; note also that each restriction $a'_{\cdot i}x \geq 0$ defines a hyperplane anchored at the origin.

Remark 5. It is interesting to note that convex cones and polyhedral cones are very closely related. In fact we have

Proposition 3. Every closed convex cone with finitely many generators is a polyhedral cone. Moreover, if P_c is a polyhedral cone, there exist vectors, $a_{\cdot i} : a_{\cdot i} \in R^n, i = 1, 2, 3, \ldots, n$ such that

$$P_c = \{a : a = \sum_{i=1}^{n} w_i a_{\cdot i}, \quad \sum_{i=1}^{n} a_i = 1, \quad w_i \geq 0\}.$$

Proof: See Silvapulle and Sen (2005), p. 123.

Associated with any cone there is a **dual or polar** cone, as follows.

Definition 25. Let A be an $n \times k$ matrix of rank $k \leq n$ and consider the closed convex polyhedral cone

$$P_c = \{x : A'x \geq 0\}. \tag{1.169}$$

The **polar cone** associated with it is given by

$$P_c^\circ = \{Ay : y \in R^k, y \leq 0\}. \tag{1.170}$$

If the (closed) convex cone is given by

$$C_1 = \{Ax : x \geq 0, \ x \in R^k\}, \tag{1.171}$$

then its polar cone is given by

$$C_1^\circ = \{y : A'y \leq 0, \quad y \in R^n\}. \tag{1.172}$$

Remark 6. Note that if $x \in P_c$ and $y \in P_c^\circ$, then $y'A'x \leq 0$, because $A'x \geq 0$ and $y \leq 0$. Similarly, if $x \in C_1$ and $y \in C_1^\circ$ then $x'A'y \leq 0$.

Definition 26. Let P_c be a closed convex polyhedral cone. The collection

$$F_J = \{x : A'_J x = 0, A^{*\prime}_J x \geq 0\} \tag{1.173}$$

is said to be a **face** of P_c, where A_J is the sub-matrix of A (which is of dimension $n \times k$) consisting of $J \leq k$ of its columns and A^*_J is the complement in A, i.e. it consists of the remaining columns of A.

A more general definition of a face, given in HL (1993), p. 112, is: A non-empty convex set $F \subseteq C$ is said to be a face (of C) if every segment of C having in its relative interior an element of F is contained entirely in F, i.e. if $x_1, x_2 \in C$ and there exists $\gamma \in (0,1)$ such that $\gamma x_1 + (1-\gamma)x_2 \in F$ then the interval $[x_1, x_2] \subseteq F$.

Definition 27. Let $C \in R^n$ be a convex set. Its relative interior, denoted by riC, consists of points $x \in AH(C)$, such that if $y \in AH(C)$ and $\|y - x\| \leq \epsilon$ then $y \in C$. In other words it is the subset of C which consists of points in its $AH(C)$ whose ϵ-neighborhoods lie entirely in C.

Remark 7. Of interest in this problem is whether the constraints are satisfied in the form $a'_{\cdot i} x = 0$ or in the form $a'_{\cdot i} x > 0$. Since there are k constraints and each constraint may be satisfied in 2 possible ways, the problem then admits of 2^k choices so that there are 2^k possible faces, viz.,

$$\begin{array}{ccccc} 0 & 0 & 0 & \ldots & 0 \\ \geq & 0 & 0 & \ldots & 0 \\ 0 & \geq & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \geq \end{array}$$

and continue by taking two at a time non-negative, three at a time non-negative, up to all k non-negative.

Example 3. Consider the closed convex cone $P_c = \{x : Ax \geq 0\}$, where $x \in R^2$ and A is two by two non-singular. The faces of the closed convex cone P_c are,

$$F_1 = \{x : Ax = 0\}$$

$$F_2 = \{x : a_{1\cdot} x = 0, \quad a_{2\cdot} x > 0\}$$

$$F_3 = \{x : a_{2\cdot} x = 0, \quad a_{1\cdot} x > 0\}$$

$$F_4 = \{x : Ax > 0\}. \tag{1.174}$$

Note that **because** A is **non-singular**, F_1 is a single point, i.e. $F_1 = (0,0)$. F_2 is given by the line $0 = a_{1\cdot}x = a_{11}x_1 + a_{12}x_2$, provided the set of points satisfying the latter is contained in the set of points satisfying $a_{2\cdot}x = a_{21}x_1 + a_{22}x_2 > 0$. Similarly, F_3 is given by the line $0 = a_{2\cdot}x = a_{21}x_1 + a_{22}x_2$, provided the latter is contained in the set of points satisfying $a_{1\cdot}x = a_{11}x_1 + a_{12}x_2 > 0$. Finally $F_4 = P_c$, or more properly the interior of P_c, denoted by $int\ P_c$.

Remark 8. Let P_c be a closed convex cone in R^n defined by

$$P_c = \{x : x \in R^n, A'x \geq 0, \quad A \text{ being } \ n \times k, \quad \text{of rank } k.$$

and let F_J be a face of P_c as defined in Definition 26. The **relative interior** of F_J, denoted by $ri(F_J)$ (or by $ri\ F_J$) is given by

$$ri(F_J) = \{x : A'_J x = 0, \quad A^{*\prime}_J x > 0\}. \tag{1.175}$$

Remark 9. Note that we could put, by convention, $riF_0 = int\ P_c$, since $ri\ F_0 = \{x : A'x > 0\}$.

The face F_k is the set $\{x : A'x = 0\}$. But in contrast to the case examined in Example 3, F_k is not the (vector) singleton $\{0\}$. This is so because A is $n \times k$ of rank $k \leq n$. By a theorem in linear algebra, see Dhrymes (2000), pp. 18-19, the **dimension of the column null space** of A' is $n - k \geq 0$!

Another interesting fact to note that F_J denotes an **equivalence class**, i.e. $\binom{k}{J}$ faces of the form $\{x : A_J x = 0, \quad A^*_J x > 0\}$ because J can be chosen in $\frac{k!}{(k-J)!J!}$ ways.

Remark 10. An interesting aspect of closed convex sets like P_c is that they are "covered" by the union of the relative interiors of their faces, i.e.

$$P_c \subseteq \bigcup_{J=1}^{k} ri(F_J), \text{ where } J \text{ is the number of columns in } A_J. \tag{1.176}$$

A crucial aspect of constrained estimation (and unconstrained estimation as well) is the concept of projection.

Projections and their Properties

Definition 28. Let $R^n = \mathcal{S}$ and let $S \subseteq \mathcal{S}$ be a (linear) subspace. Let $x \in \mathcal{S}$. The **projection** of $x \in \mathcal{S}$ onto the set S, say $s^* \in S$, is defined by the condition

$$\|x - s^*\| = \inf_{s \in S} \|x - s\|, \tag{1.177}$$

so that it is the point in S whose distance from x is minimal.

We give below a rather general projection result, which may be specialized to the matters under current discussion. To minimize notational ambiguity, we shall **henceforth denote convex polyhedral cones as well as other convex sets, as needed, by** C and we shall reserve P to **denote projections**.

Proposition 4. (Projection Theorem) Let $\epsilon = \{e_n : n \in \mathcal{N}\}$ be an orthonormal sequence of random elements in $\mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P})$, (the Hilbert space of square integrable random elements), and $\mathcal{M}$ be the closed linear manifold generated by ϵ,[12] so that $\mathcal{M} \subseteq \mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P})$. Given any element $X \in \mathcal{H}^2$, the following statements are true:

i. There exists a **unique** element, $Y \in \mathcal{M}$, such that

$$\|X - Y\| = \inf_{\zeta \in \mathcal{M}} \|X - \zeta\|. \tag{1.178}$$

ii. $Y = \sum_{j=1}^{n} c_j \epsilon_j$.

iii. for every element $\xi \in \mathcal{M}$, $X - Y \perp \xi$.

iv. $\mathcal{H}^2 = \mathcal{M} \oplus \mathcal{M}^{\perp}$, i.e. every element, $X \in \mathcal{H}^2$ can be written **uniquely** as $X = X_1 + X_2$, such that $X_1 \in \mathcal{M}$ and $X_2 \in \mathcal{M}^{\perp}$.

Proof. See Dhrymes (1998), pp. 46-47.

For ease of exposition write the projection Y as $Y = P(X|\mathcal{M})$, and $X - Y = P(X|\mathcal{M}^{\perp})$, which we shall also denote, occasionally, by P°.

A number of other properties of projections are useful.

Proposition 5. Let $\mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P})$ and the context be as in Proposition 4. Define the operators

$$P : \mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P}) \to \mathcal{M}, \quad P^{\circ} : \mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P}) \to \mathcal{M}^{\perp}. \tag{1.179}$$

Then the following statements are true:

[12] In this context, a closed linear manifold generated by ϵ consists of $M_n = \{X : X = \sum_{j=1}^{n} c_j \epsilon_j\}$ together with the limits of sequences lying in M_n and $\mathcal{M} = \bigcap_{n=-\infty}^{\infty} M_n(\epsilon)$.

i. the projection operators, P and P° are linear, i.e. given any $X_i \in \mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P})$ and $a_i \in \mathcal{C}$ (the field of complex numbers), $i = 1, 2$,

$$P(a_1 X_1 + a_2 X_2) = a_1 P(X_1) + a_2 P(X_2), \quad \text{and similarly for } P^\circ;$$

ii. for every $X \in \mathcal{H}^2(\Omega, \mathcal{A}, \mathcal{P})$ ther exists a unique decomposition

$$X = P(X) + P^\circ(X), \quad P(X) \in \mathcal{M}, \quad P^\circ(X) \in \mathcal{M}^\perp; \tag{1.180}$$

iii. for every $X \in \mathcal{H}^2$

$$\|X\|^2 = \|P(X)\|^2 + \|P^\circ(X)\|^2;$$

iv. let $X_n \in \mathcal{H}^2, n \geq 1$, such that $\|X_n - X\| \to 0$; then

$$\|P_n(X) - P(X)\| \to 0;$$

v. $X \in \mathcal{M}$ if and only if $P(X) = X$, and $X \in \mathcal{M}^\perp$ if and only if $P(X) = 0$;

vi. let $\mathcal{M}_i, P_i, i = 1, 2,$ be the closure of two linear manifolds and their associated projections; then, $\mathcal{M}_1 \supseteq \mathcal{M}_2$ if and only if, for every $X \in \mathcal{H}^2$

$$P_1 \circ P_2(X) = P_1(X), \quad \text{where} \quad P_1 \circ P_2(X) = P_1[P_2(X)].$$

Proof. See Dhrymes (1998), pp. 48-50, particularly Remark 4.

Remark 11. When the space on which we project is a (linear) vector space the projection operator, P, may be represented by an idempotent matrix; the matrix of the projection onto the orthogonal complement, P°, is also idempotent .

For example, in the GLM $y = X\beta + u$, X a matrix of full column rank and β **unrestricted**, estimating β involves the projection of the vector y on the vector space spanned by the **columns** of X, regarded as a subspace of R^T. In this case the operators P and P° may be represented, respectively, by the idempotent matrices

$$P \sim X(X'X)^{-1}X', \quad P^\circ \sim I - X(X'X)^{-1}X'.$$

We give below two results which, although may be derived from Proposition 4, deal directly with projections on convex sets and thus are more closely related to the material in this appendix. We have

Proposition 6. Let $C \subseteq R^n$ be non-empty, and $a \in R^n \setminus C$, i.e. a is in R^n but a is not in C. The following statements are true:

i. There exists a point $\tilde{a} \in C$ such that

$$\|a - \tilde{a}\| = \inf_{\zeta \in \mathcal{S} \setminus C} \|a - \zeta\|.$$

ii. The minimizer in i is unique.

iii. (Characterization) The point $\tilde{a} \in C$ is the unique point in C closest to $a \in R^n \setminus C$, if and only if

$$(a - \tilde{a})'(z - \tilde{a}) \leq 0,$$

for all $z \in C$.

iv. (Separating hyperplane) There exists a vector, v and a scalar c, such that

$$v'a > c, \quad v'z < c, \quad \forall z \in C.$$

Proof. The proof of i essentially rests on the continuity of the norm function; thus, define

$$f : C \to R, \quad f(z) = \|a - z\|;$$

let $z_0 \in C$ and define

$$B = \{x : x \in R^n,\ f(x) \leq f(z_0)\}.$$

Clearly the set $B \cap C$ is **non-empty and compact**; since f is continuous on $B \cap C$ (because it is continuous on C and thus on $(B \cap C)$), it attains a global minimum (on $B \cap C$), say at $\tilde{a}$. Since $C = (C \cap B) \cup (C \cap \bar{B})$, we need only show that $\tilde{a}$ is the minimum over $\bar{B}$. Taking $z_0 = \tilde{a}$ we see that, for $x \notin B$, $f(x) \geq f(\tilde{a})$, thus showing that $\tilde{a}$ is the **global minimizer** over B.

To prove ii, i.e. show uniqueness, suppose another point exists, e.g. $a^* \in C$, such that

$$\|a - \tilde{a}\| = \|a - a^*\|.$$

Using the identity

$$\frac{1}{2}\|x_1 + x_2\|^2 = \|x_1\|^2 + \|x_2\|^2 - \frac{1}{2}\|x_2 - x_1\|^2,$$

with $x_1 = \tilde{a}$, $x_2 = a^*$, we find

$$-\frac{1}{2}\|\tilde{a} - a^*\|^2 = 0,$$

which shows uniqueness.

To prove iii, suppose $\tilde{a} \in C$ and that $(a - \tilde{a})'(z - \tilde{a}) \leq 0$ for any $z \in C$, we show that $\|a - \tilde{a}\| \leq \|a - z\|$, for any $x \in C$ and conversely. Since $a - z = (a - \tilde{a}) + (\tilde{a} - z)$, we have

$$\|a - z\|^2 = \|a - \tilde{a}\|^2 + \|\tilde{a} - z\|^2 - 2(a - \tilde{a})'(z - \tilde{a});$$

but by assumption $(a - \tilde{a})'(z - \tilde{a}) \leq 0$ and thus, for any $z \in C$

$$\|a - z\| \geq \|a - \tilde{a}\|,$$

which shows that $\tilde{a} = P(a|C)$ and is thus the unique point in C closest to a . Conversely, suppose $\tilde{a} = P(a|C)$; then, for $\lambda > 0$, the point $a^* = \tilde{a} + \lambda(z - \tilde{a}) \in C$, by convexity. Because $a^* \neq \tilde{a}$, it follows that

$$\|a - \tilde{a}\|^2 \leq \|a - a^*\|^2 = \|a - \tilde{a}\|^2 + \lambda^2\|z - \tilde{a}\|^2 - 2\lambda(a - \tilde{a})'(z - \tilde{a}).$$

This shows that

$$0 \leq \lambda^2\|x - \tilde{a}\|^2 - 2\lambda(a - \tilde{a})'(x - \tilde{a}), \text{ or } 2\lambda(a - \tilde{a})'(x - \tilde{a}) \leq \lambda^2\|x - \tilde{a}\|^2.$$

Dividing both sides by $\lambda > 0$ and letting the latter approach zero from above, we conclude that $(a - \tilde{a})'(z - \tilde{a}) \leq 0$, for every $z \in C$.

To prove iv, we note from iii that $\tilde{a} = P(a|C)$ and $\tilde{a} \in C$ and, moreover, that

$$(a - \tilde{a})'(z - \tilde{a}) \leq 0 \quad \text{for any} \quad z \in C.$$

Put $v = a - \tilde{a}$, add and subtract a , in the second member above, to obtain

$$0 \geq (v, z - a + v) = v'z - v'a + v'v.$$

Since this holds for any $z \in C$, we may write

$$v'a \geq \sup_{z \in C} + v'v > \sup_{z \in C} = 2c - v'a,$$

where $2c = \sup_{z \in C} v'z + v'a$. Thus,

$$v'a > c.$$

Moreover, for an $z \in C$ we have

$$v'z \leq v'a - v'v < v'a = 2c - \sup_{z \in C} v'z < 2c - v'z,$$

or, for any $z \in C$,

$$v'z < c,$$

which completes the proof.

q.e.d.

Remark 12. Notice that $v'x = c$, is an affine hyperplane. Notice further that $v'a > c$, so that the point a lies **above** the hyperplane or at any rate entirely on one side of it; in addition $v'z < c$, for **any** $z \in C$. Thus, it may be said that C lies entirely on the **other side**, so that the hyperplane $v'x = c$, **separates** the two convex sets C and $\{a\}$. This result may be generalized to the case neither of the two convex sets **is a singeton**.

For completeness of the exposition pertinent to cones or, more generally, convex sets, we give

Proposition 7. Let C be a closed convex cone and $x \in \mathcal{S}\backslash C$, where $\mathcal{S} = R^n$. The following statements are true:

i. Let $x \in \mathcal{S} \setminus C$, $y = \tilde{x} \in C$, $z \in C^\circ$, $y'z = 0$, and $x = y + z$. Then

$$y = P(x|C), \quad z = P(x|C^\circ).$$

ii. Conversely, suppose $\tilde{x} = P(x|C)$, $z = P(x|C^\circ)$, then $x = y + z$ and $y'z = 0$.

Proof. For the proof of i, let $\tilde{x} = P(x|C)$ and consider $y = (1+\delta)\tilde{x}$, for $|\delta| < 1$; notice that $y \in C$ because $1 + \delta \geq 0$ and $\tilde{x} \in C$. By part ii of Proposition 3, $(x-\tilde{x})'(y-\tilde{x}) \leq 0$, or $\delta(x-\tilde{x})\tilde{x} \leq 0$. But since $\delta \in (-1, 1)$ and otherwise arbitrary, we conclude that $(x-\tilde{x})'\tilde{x} = 0$.

For the proof of ii, let $v \in C$, an arbitrary point and consider $y = \tilde{x}+v \in C$. By part ii of Proposition 4 above we have

$$(x-\tilde{x})'(y-\tilde{x}) \leq 0, \quad \text{or } (x-\tilde{x})'v \leq 0.$$

But v is an arbitrary point in C, thus $x - \tilde{x} \in C^\circ$. Consequently, we have shown that

$$x = \tilde{x} + (x-\tilde{x}), \quad (x-\tilde{x})'\tilde{x} = 0, \quad x - \tilde{x} \in C^\circ, \tilde{x} \in C.$$

By part i this shows that

$$x = P(x|C) + P(x|C^\circ), \quad \text{and} \quad P(x|C)'P(x|C^\circ) = 0. \tag{1.181}$$

q.e.d.

For more detailed discussion of the issues raised in connection with the specialization of Propositions 4, and 5 to convex sets, see Hiriart-Urruti and Lemarechal (1993), p. 116 ff.

The results above may be particularized to the problems we consider in this chapter. Thus, let

$$y = X\beta + u, \quad \text{with the possible constraint} \quad R\beta = 0, \tag{1.182}$$

in the usual regression context, with i.i.d. errors. Here we are dealing with the linear spaces

$$\begin{aligned} \mathcal{M}_1 &= \{y : y = X\beta, \beta \in R^m\}, \\ \mathcal{M}_2 &= \{y : y = X\beta, \beta \in C\}, \quad \text{such that } C = \{\beta : R\beta = 0\}. \end{aligned} \tag{1.183}$$

It is evident that $\mathcal{M}_2 \subseteq \mathcal{M}_1$ so that, by vi of Proposition 3,

$$\inf_{\beta \in C} \|y - X\beta\| = \inf_{\beta \in C} \|y_1 - X\beta\|, \quad y_1 = P(y|\mathcal{M}_1), \tag{1.184}$$

in the sense that the minimizer of the left member is **precisely** the minimizer obtained by minimizing the right member.

An entirely similar result will hold if the restricted space is defined by

$$\mathcal{M}_2 = \{y : y = X\beta, \beta \in C\}, \quad \text{such that } C = \{\beta : R\beta \geq 0\}, \tag{1.185}$$

since in this case C is a closed (convex) cone over R^+ contained in R^n, so that again $\mathcal{M}_2 \subseteq \mathcal{M}_1$, and $\mathcal{M}_2$ is a **linear space**, i.e. if $y, z \in \mathcal{M}_2$ then $c_1 y + c_2 z \in \mathcal{M}_2$ as well, for $c_i \geq 0, i = 1, 2$.

REFERENCES

Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Andrews, D.W. K. (1999), "Estimation When a Parameter is on a Boundary, *Econometrica*, vol 67, pp. 1341-1383.

Chernoff, H. (1959) "On the Distriibution of the Likelihood Ratio, *The Annals of Mathematical Statistics*, vol. 25, pp. 573-578.

Dhrymes, P. J. (1978) *Introductory Econometrics*, New York: Springer-Verlag.

Dhrymes, P. J. (1989) *Topics in Advanced Econometrics: Probability Foundations*, New York: Springer-Verlag.

Dhrymes, P. J. (1994) *Topics in Advanced Econometrics, vol. II: Linear and Nonlinear Simultaneous Equations*, New York: Springer-Verlag.

Dhrymes, P. J. (1998) *Time Series, Unit Roots and Cointegration*, New York: Academic Press.

Dhrymes, P. J. (2000) *Mathematics for Econometrics*, third edition, New York: Springer-Verlag.

Gourieroux C., A. Holly and A. Montforte (1982), Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on Regression Parameters, *Econometrica*, vol. 50, pp. 63-80

Hiriart-Urruti, J-B. and C. Lemarechal (1993), *Convex Analysis and Minimization Algorithms, vols. I nd II*, Berlin: Springer-Verlag.

Rockafellar, R.T. (1970), *Convex Analysis*, Princeton: Princeton University Press.

Potscher, B.M. and I.R. Prucha (1989), 'A Uniform Law of Large Numbers for Dependent and Heterogeneous Data Processes, *Econometrica*, vol. 57, pp. 675-683.

Shapiro, A. (1985) Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures under Inequality Constraints, *Biometrika* vol. 72 pp. 133-144.

Shapiro, A. (1987) A Conjecture Related to Chi-bar-squared Distributions, *The American Mathematical Monthly*, vol. 94, pp. 46-48.

Shapiro, A. (1988) Toward a Unified Theory if Inequality Constrained Testing in Multivariate Analysis, *International Statistical Review*, vol. 56, pp.49-62.

Shapiro, A. (1989) Asymptotic Properties of Statistical Estimators in Stochastic Programming, *The Annals of Statistics*, vol. 17, pp. 841-858

Shapiro, A. (2000) On the Asymptotics of Constrained Local M-Estimators, *The Annals of Statistics*, vol. 28, pp. 948-960

Silvapulle, M. J. (1994) On Tests Against One-Sided Hypotheses in Some Generalized Linear Models, *Biometrics*, vol. 50, pp. 853-858.

Silvapulle, M. J. and P. K. Sen (2005), *Constraint Statistical Inference*, New York: Wiley-Interscience.

Stoer, J. and C. Witzgall (1970) *Convexity and Optimization in Finite Dimensions, I*, New York: Springer-Verlag.