THE CLASSICAL LINEAR REGRESSION MODEL

Herman J. Bierens

Pennsylvania State University

September 1, 2002

1. Introduction

The classical linear regression model takes the form

$$y_j = \theta_1 x_{1,j} + ... + \theta_k x_{k,j} + u_j,\ j = 1,...,n, \tag{1}$$

where y_j is the dependent variable, the $x_{i,j}$'s are the independent (or explanatory) variables, the u_j's are unobservable error terms, n is the sample size, and the θ_i's are the model parameters. If the model contains an intercept, then one of the $x_{i,j}$'s is equal to 1, say $x_{1,j}$.

Denoting

$$x_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{k,j} \end{pmatrix},\ \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}, \tag{2}$$

we can write model **(1)** more compactly as

$$y_j = x_j^T \theta + u_j,\quad j = 1,...,n \tag{3}$$

For the time being we assume:

ASSUMPTION 1: *The $x_{i,j}$'s are non-stochastic.*

ASSUMPTION 2: *The u_j's are independent $N(0,\sigma^2)$ distributed.*

In particular Assumption 1 is very unrealistic for economic data, but we impose it for pedagogical reasons only. As will appear later on, we may without loss of generality assume that the $x_{i,j}$'s are

random (except for the one corresponding to the intercept). For example, we may replace Assumption 1 by the assumption that for each $t = 1,..,n$, u_t is independent of all the $x_{i,j}$'s.

Denoting:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} x_{1,1} & \cdots & x_{k,1} \\ x_{1,2} & \cdots & x_{k,2} \\ \vdots & \cdots & \vdots \\ x_{1,n} & \cdots & x_{k,n} \end{pmatrix}, \; u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \tag{4}$$

we can cast model **(1)** is vector/matrix form as:

$$y = X\theta + u, \quad u \sim N_n(0, \sigma^2 I_n), \tag{5}$$

hence,

$$y \sim N_n(X\theta, \sigma^2 I_n), \tag{6}$$

because by Assumption 1, $X\theta$ is non-random.

2. Least squares estimation

For an arbitrary vector $\theta_* \in \mathbb{R}^k$ we have:

$$\begin{aligned} E\left((y - X\theta_*)^T(y - X\theta_*)\right) &= E\left((u + X\theta - X\theta_*)^T(u + X\theta - X\theta_*)\right) \\ = E\left(u^T u - u^T X(\theta_* - \theta) - (\theta_* - \theta)^T X^T u + (\theta_* - \theta)^T X^T X(\theta_* - \theta)\right) \\ = n\sigma^2 + (\theta_* - \theta)^T X^T X(\theta_* - \theta), \end{aligned} \tag{7}$$

which is minimal for $\theta_* = \theta$. This solution is unique if:

ASSUMPTION 3: *The matrix $X^T X$ is nonsingular,*

because then X^TX is positive definite. This result motivates the least squares estimation of θ. *In general an estimator is a function of the data which serves as an approximation of a parameter or a parameter vector.* In particular, the least squares estimator $\hat{\theta}$ of θ is the solution of the minimization problem:

$$\begin{aligned}\min_{\hat{\theta}} (y - X\hat{\theta})^T(y - X\hat{\theta}) &= \min_{\hat{\theta}} \left(y^Ty - \hat{\theta}^TX^Ty - y^TX\hat{\theta} + \hat{\theta}^TX^TX\hat{\theta}\right) \\ &= \min_{\hat{\theta}} \left(y^Ty - 2\hat{\theta}^TX^Ty + \hat{\theta}^TX^TX\hat{\theta}\right).\end{aligned} \tag{8}$$

The first-order condition for the minimum involved is:

$$\frac{\partial\left(y^Ty - 2\hat{\theta}^TX^Ty + \hat{\theta}^TX^TX\hat{\theta}\right)}{\partial\hat{\theta}^T} = -2X^Ty + 2X^TX\hat{\theta} = 0, \tag{9}$$

hence:

$$\hat{\theta} = \left(X^TX\right)^{-1}X^Ty. \tag{10}$$

Remarks: In **(9)** we have used the notation:

$$\frac{\partial f(x)}{\partial x^T} = \begin{pmatrix} \partial f(x)/\partial x_1 \\ \vdots \\ \partial f(x)/\partial x_n \end{pmatrix}, \text{ where } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ and } f(x) \text{ is a function of } x. \tag{11}$$

In particular, if $f(x) = a + x^Tb + x^TCx$ where

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \; C = \begin{pmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & \cdots & \vdots \\ c_{n,1} & \cdots & c_{n,n} \end{pmatrix}, \text{ with } c_{i,j} = c_{j,i} \; (\text{thus } C = C^T), \tag{12}$$

then

$$\partial f(x)/\partial x_k = \frac{\partial\left(a + \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} x_i c_{i,j} x_j\right)}{\partial x_k}$$

$$= \sum_{i=1}^{n} b_i \frac{\partial x_i}{\partial x_k} + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial x_i c_{i,j} x_j}{\partial x_k} = b_k + 2c_{k,k}x_k + \sum_{\substack{i=1 \\ i\neq k}}^{n} x_i c_{i,k} + \sum_{\substack{j=1 \\ j\neq k}}^{n} c_{k,j} x_j \qquad \textbf{(13)}$$

$$= b_k + 2\sum_{j=1}^{n} c_{k,j} x_j, \quad k = 1,\ldots,n,$$

hence

$$\frac{\partial f(x)}{\partial x^T} = b + 2Cx. \qquad \textbf{(14)}$$

If C is not symmetric, we may without loss of generality replace C in the quadratic function $f(x)$ by the symmetric matrix $(C + C^T)/2$ (because $x^TCx = (x^TCx)^T = x^TC^Tx$), so that then

$$\frac{\partial f(x)}{\partial x^T} = b + Cx + C^T x. \qquad \textbf{(15)}$$

3. Properties of the least squares estimator for fixed sample size

Substituting **(5)** in **(10)** yields:

$$\hat{\theta} = (X^TX)^{-1}X^T(X\theta + u) = \theta + (X^TX)^{-1}X^Tu. \qquad \textbf{(16)}$$

Since u is multivariate normally distributed and X is assumed to be nonstochastic, it follows that $\hat{\theta}$ is k-variate normally distributed with expectation

$$E(\hat{\theta}) = \theta + E((X^TXt)^{-1}X^Tu) = \theta + (X^TX)^{-1}X^TE(u) = \theta \qquad \textbf{(17)}$$

(hence $\hat{\theta}$ is an *unbiased* estimator) and variance matrix

$$E\left((\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\right) = E\left((X^TX)^{-1}X^Tuu^TX(X^TX)^{-1}\right) = \left(X^TX\right)^{-1}X^T\left(E(uu^T)\right)X\left(X^TX\right)^{-1}$$
$$= \sigma^2\left(X^TX\right)^{-1}. \tag{18}$$

Moreover, the least squares estimator is the *best linear unbiased estimator* (BLUE), in the sense that for all estimators of the form $\hat{\theta}_* = Cy$, where C is a $k \times n$ matrix such that $E(\hat{\theta}_*) = \theta$, we have that $Var(\hat{\theta}_*) = E[(\hat{\theta}_* - \theta)(\hat{\theta}_* - \theta)^T] = \sigma^2(X^TX)^{-1} + D$, where D is a positive semi-define matrix.

The proof of this proposition is quite easy. First, observe that the unbiasedness condition implies that $CX = I_k$, hence $\hat{\theta}_* = C(X\theta + u) = \theta + Cu$, and thus $Var(\hat{\theta}_*) = \sigma^TCC^T$. Now

$$D = \sigma^2[CC^T - (X^TX)^{-1}] = \sigma^2[CC^T - CX(X^TX)^{-1}X^TC^T]$$
$$= \sigma^2C[I_k - X(X^TX)^{-1}X^T]C^T = \sigma^2CMC^T, \tag{19}$$

say, where the second equality follows from the unbiasedness condition $CX = I_k$. The matrix

$$M = I_n - X\left(X^TX\right)^{-1}X^T \tag{20}$$

is idempotent:

$$M^2 = \left(I_n - X(X^TX)^{-1}X^T\right)\left(I_n - X(X^TX)^{-1}X^T\right) = I_n - 2X\left(X^TX\right)^{-1}X^T$$
$$+ X\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}X^T = I_n - X\left(X^TX\right)^{-1}X^T = M, \tag{21}$$

hence its eigenvalues are either 1 or 0. Since all the eigenvalues are non-negative, M is positive semi-definite, and so is CMC^T. Thus we have:

THEOREM 1: *Under Assumptions 1-3,* $\hat{\theta} - \theta \sim N_k\left(0, \sigma^2(X^TX)^{-1}\right)$, *and* $\hat{\theta}$ *is BLUE.*

The latter result is known as the Gauss-Markov theorem.

4. Estimation of the error variance

Since by **(1)**,

$$\min_{\theta_*} \frac{1}{n} E\left((y - X\theta_*)^T(y - X\theta_*)\right) = \sigma^2, \tag{22}$$

it seems at first sight a good idea (but not at second sight as will appear) to estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\theta})^T(y - X\hat{\theta}) = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - x_j^T\hat{\theta}\right)^2 = \frac{1}{n}\sum_{j=1}^{n}\hat{u}_j^2, \tag{23}$$

where $\hat{u}_j = y_j - x_j^T\hat{\theta}$ is called the residual of y_j (i.e., the part of y_j that is left over after accounting for the effect of x_j). The quadratic form involved is called the Residual Sum of Squares (RSS):

$$RSS = (y - X\hat{\theta})^T(y - X\hat{\theta}) = \sum_{j=1}^{n}\hat{u}_j^2. \tag{24}$$

This is sometimes also referred to as the Sum of Squared Residuals (SSR).

Note that the RSS can be computed without computing each of the $\hat{u}_j$'s separately, as follows:

$$RSS = y^Ty - y^TX\hat{\theta} - \hat{\theta}^TX^Ty + \hat{\theta}^TX^TX\hat{\theta} = y^Ty - \hat{\theta}^TX^Ty = y^Ty - \hat{\theta}^TX^TX\hat{\theta}. \tag{25}$$

(See Exercise 1) Substituting **(5)** and **(16)** in **(24)** yield:

$$\begin{aligned} RSS &= (u + X\theta - X\hat{\theta})^T(u + X\theta - X\hat{\theta}) = \left(u - X(\hat{\theta}-\theta)\right)^T\left(u - X(\hat{\theta}-\theta)\right) \\ &= \left(u - X(X^TX)^{-1}X^Tu\right)^T\left(u - (X^TX)^{-1}X^Tu\right) = u^TM^2u, \end{aligned} \tag{26}$$

where M is defined by **(20)**. As shown in **(21)**, M is idempotent, hence

$$\begin{aligned} \text{rank}(M) &= \text{trace}(M) = \text{trace}\left(I_n - X(X^TX)^{-1}X^T\right) = \text{trace}(I_n) - \text{trace}\left(X(X^TX)^{-1}X^T\right) \\ &= \text{trace}(I_n) - \text{trace}\left((X^TX)^{-1}X^TX\right) = \text{trace}(I_n) - \text{trace}(I_k) = n - k. \end{aligned} \tag{27}$$

Thus:

$$RSS = u^TMu \tag{28}$$

and consequently, by one of the results for the multivariate normal distribution, hereafter indicated as "an *MND* result" (which one?),

THEOREM 2: *Under Assumptions 1-3,* $RSS/\sigma^2 \sim \chi^2_{n-k}$.

This result implies that

$$E[\hat{\sigma}^2] = (n-k)\sigma^2/n, \tag{29}$$

hence $\hat{\sigma}^2$ is a *biased* estimator. However, the following correction yields an *unbiased* estimator:

$$s^2 = \frac{1}{n-k}(y - X\hat{\theta})^T(y-X\hat{\theta}). \tag{30}$$

Next we show that s^2 and $\hat{\theta}$ are independent, by showing that $(X^TX)^{-1}X^Tu$ and u^TMu are independent. A necessary and sufficient condition for the latter is that

$$(X^TX)^{-1}X^TM = O. \tag{31}$$

Condition **(31)** follows from:

$$\begin{aligned}(X^TX)^{-1}X^T(I - X(X^TX)^{-1}X^T) &= (X^TX)^{-1}X^T - (X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= (X^TX)^{-1}X^T - (X^TX)^{-1}X^T = O.\end{aligned} \tag{32}$$

Summarizing, we have shown that:

THEOREM 3: *Under Assumptions 1-3,* $E(s^2) = \sigma^2$, $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$, *and* s^2 *and* $\hat{\theta}$ *are independent*.

5. The t-test.

Suppose we want to test the null hypothesis that the *i*-th component θ_i of θ equals zero:

$$H_0: \theta_i = 0, \tag{33}$$

which amounts to the hypothesis that the corresponding variable $x_{i,j}$ can be deleted from model **(1)**, against the alternative hypothesis

$$H_1: \theta_i \neq 0. \tag{34}$$

The general procedure for testing statistical hypothesis is to construct a function of the data, called test statistic, that has under the null hypothesis a particular distribution, and under the alternative hypothesis a distribution that differs from the one under the null hypothesis.

In order to construct a test statistic for the null hypothesis **(33)**, we first isolate θ_i and its least squares estimator $\hat{\theta}_i$ from θ and $\hat{\theta}$, respectively, by taking the linear transformations

$$\theta_i = e_i^T\theta, \ \hat{\theta}_i = e_i^T\hat{\theta} \tag{35}$$

where e_i is the i-th column of the $k \times k$ unit matrix I_k. Thus, e_i is a k-vector of zeros, except for the i-the component, which equals 1. Then it follows from Theorem 1, together with an MND result (*Exercise*: Which one?), that

$$\hat{\theta}_i - \theta_i = e_i^T(\hat{\theta} - \theta) \sim N\left(0, \sigma^2 e_i^T(X^TX)^{-1}e_i\right), \tag{36}$$

hence

$$\frac{\hat{\theta}_i - \theta_i}{\sigma\sqrt{e_i^T(X^TX)^{-1}e_i}} \sim N(0,1). \tag{37}$$

Note that $e_i^T(X^TX)^{-1}e_i$ is just the i-th diagonal element of $(X^TX)^{-1}$. Using the definition of the t-distribution, and the result in Theorem 3, it is now easy to verify that

THEOREM 4: *Under Assumptions 1-3,*

$$\frac{\hat{\theta}_i - \theta_i}{s\sqrt{e_i^T (X^T X)^{-1} e_i}} \sim t_{n-k}. \tag{38}$$

Proof: Exercise.

The denominator in the left-hand side expression involved is called the *standard error* (*se*) of $\hat{\theta}_i$:

$$se(\hat{\theta}_i) = s\sqrt{e_i^T (X^T X t)^{-1} e_i}. \tag{39}$$

Now under the null hypothesis **(33)** we have

$$\hat{t}_i = \frac{\hat{\theta}_i}{s\sqrt{e_i^T (X^T X)^{-1} e_i}} \sim t_{n-k}. \tag{40}$$

The statistic $\hat{t}_i$ is called the *t-value* of $\hat{\theta}_i$, and will be our test statistic. Most econometric software packages report either the standard error, or the t-value, or both, for each least squares estimate $\hat{\theta}_i$.

Before we turn to the actual testing procedure, we need to pay attention to what happens with the t-value if the null hypothesis is false, i.e., if **(34)** is true. Then

$$\hat{t}_i = \frac{\hat{\theta}_i - \theta_i}{s\sqrt{e_i^T (X^T X)^{-1} e_i}} + \frac{\theta_i}{s\sqrt{e_i^T (X^T X)^{-1} e_i}}. \tag{41}$$

The first term in this expression is *t* distributed with $n - k$ degrees of freedom, which converges in distribution to the $N(0,1)$ distribution if n increases to infinity:

$$\frac{\hat{\theta}_i - \theta_i}{s\sqrt{e_i^T (X^T X)^{-1} e_i}} \to N(0,1) \text{ in distr. if } n \to \infty. \tag{42}$$

Moreover,

$$\operatorname{plim}_{n\to\infty} s^2 = \sigma^2, \tag{43}$$

hence

$$\operatorname{plim}_{n\to\infty} s = \sigma. \tag{44}$$

Next, assume that

ASSUMPTION 4: $\lim_{n\to\infty}(1/n)\sum_{j=1}^{n} x_j x_j^T = Q$, *where Q is a finite positive definite matrix.*

Then $\lim_{n\to\infty} X^T X/n = Q$, hence $\lim_{n\to\infty} n(X^T X)^{-1} = Q^{-1}$ and thus

$$\lim_{n\to\infty} \sqrt{n}\sqrt{e_i^T (X^T X)^{-1} e_i} = \sqrt{e_i^T Q^{-1} e_i} > 0. \tag{45}$$

It follows now from **(40)**, **(42)**, **(44)** and **(45)** that

THEOREM 5: *Under Assumptions 1-3 and the null hypothesis* **(33)**, $\hat{t}_i \sim t_{n-k}$, *whereas under Assumptions 1-4 and the alternative hypothesis* **(34)**, $\operatorname{plim}_{n\to\infty} \hat{t}_t / \sqrt{n} = \theta_i / \sqrt{\sigma^2 e_i^T Q^{-1} e_i}$.

Proof: Exercise.

Thus, under the alternative **(34)** we have for an arbitrary large positive number K,

$$\lim_{n\to\infty} P(\hat{t}_i > K) = 1 \text{ if } \theta_i > 0, \ \lim_{n\to\infty} P(\hat{t}_i < -K) = 1 \text{ if } \theta_i < 0. \tag{46}$$

This result suggests a decision rule where we accept the null hypothesis (33) if for an a priori chosen constant $K > 0$, $|\hat{t}_i| \leq K$, and we reject the null hypothesis (33) in favor of the alternative hypothesis (34) is $|\hat{t}_i| > K$. Of course it is possible that the correct null hypothesis is rejected, with probability

$\alpha = P(|t_{n-k}| > K)$. This probability α is called the *Type I* error, or the *significance level*, and can be controlled by choosing K such that $P(|t_{n-k}| > K) = \alpha$ for a given value of α, using the table of the t distribution. Traditional values for the significance level are $\alpha = 0.05$ (5% significance level) and $\alpha = 0.1$ (10% significance level).

The *power* function of the test is:

$$\phi_n(\theta_i) = P(|\hat{t}_i| > K), \tag{47}$$

where K, called *critical value*, is chosen such that $P(|t_{n-k}| > K) = \alpha$ for a given significance level α. The value of the power function under the alternative hypothesis **(34)** is called the *power* of the test, and 1 minus the power is called the *Type II* error, which is the probability of not rejecting the null hypothesis if the alternative is true.

The t-test discussed so far is a two-sided test, because under the alternative both $\theta_i > 0$ and $\theta_i < 0$ are possible. If the latter is not possible, and the choice is between the null hypothesis $\theta_i = 0$ against the alternative $\theta_i > 0$, we should conduct a one-sided test: Choose for given significance level α the critical value K such that $P(t_{n-k} > K) = \alpha$. Then accept the null hypothesis $\theta_i = 0$ if $\hat{t}_i \leq K$ and reject the null in favor of the alternative $\theta_i > 0$ if $\hat{t}_i > K$. For the case of the alternative $\theta_i < 0$ we accept the null if $\hat{t}_i \geq -K$ and we reject the null in favor of the alternative if $\hat{t}_i < -K$.

6. The F-test

We now consider testing of a null hypothesis of the form

$$H_0: \quad R\theta = q, \tag{48}$$

where R is a given $r \times k$ matrix with rank r, and q is $r \times 1$ vector of given constants. The alternative we are going to consider is the alternative that this null hypothesis is false.

It follows from Theorem 1, together with an MND result (*Exercise*: Which one?), that

$$R(\hat{\theta} - \theta) \sim N_r\left(0, \sigma^2 R(X^T X)^{-1} R^T\right), \tag{49}$$

hence it follows from an MND result (which one?) that

$$\frac{(\hat{\theta} - \theta)^T R^T \left(R(X^T X)^{-1} R^T\right)^{-1} R(\hat{\theta} - \theta)}{\sigma^2} \sim \chi^2_r. \tag{50}$$

Combining this result with the results of Theorem 3, if follows from the definition of the F distribution that

$$\frac{(\hat{\theta} - \theta)^T R^T \left(R(X^T X)^{-1} R^T\right)^{-1} R(\hat{\theta} - \theta)/r}{s^2} \sim F_{r,n-k}. \tag{51}$$

Thus under the null hypothesis **(48)** we have:

$$\hat{F} = \frac{(R\hat{\theta} - q)^T \left(R(X^T X)^{-1} R^T\right)^{-1} (R\hat{\theta} - q)/r}{s^2} \sim F_{r,n-k}, \tag{52}$$

which is the test statistic of the F-test under review.

If the null hypothesis **(48)** is false, we have

$$\begin{aligned} \hat{F} = & \frac{(\hat{\theta} - \theta)^T R^T \left(R(X^T X)^{-1} R^T\right)^{-1} R(\hat{\theta} - \theta)/r}{s^2} \\ & + 2\frac{(R\theta - q)^T \left(R(X^T X)^{-1} R^T\right)^{-1} R(\hat{\theta} - \theta)/r}{s^2} \\ & + \frac{(R\theta - q)^T \left(R(X^T X)^{-1} R^T\right)^{-1} (R\theta - q)/r}{s^2}, \end{aligned} \tag{53}$$

hence:

THEOREM 6: *Under Assumptions 1-3 and the null hypothesis* **(48)**, $\hat{F} \sim F_{r,n-k}$. *If the null hypothesis* **(48)** *is false then under Assumptions 1-4,*

$$\operatorname*{plim}_{n\to\infty} \hat{F}/n = \frac{(R\theta - q)^T \left(RQ^{-1}R^T\right)^{-1}(R\theta - q)/r}{\sigma^2} > 0. \qquad \textbf{(54)}$$

Proof: Exercise.

This result suggests the following one-sided test: Given a significance level α, look up in the table of the F distribution the critical value K for which $P(F_{r,n-k} > K) = \alpha$. Then accept the null hypothesis **(48)** if $\hat{F} \leq K$, and reject it if $\hat{F} > K$.

In practice the F-test is conducted differently, as follows: Implement the null hypothesis **(48)**, and re-estimate the regression model **(5)** with the parameter vector θ reparametrized such that the condition $R\theta = q$ holds. This can be done by augmenting the matrix R with a $(k-r) \times k$ matrix R_* such that

$$\begin{pmatrix} R_* \\ R \end{pmatrix} \theta = \begin{pmatrix} \beta \\ q \end{pmatrix}, \text{ where } \begin{pmatrix} R_* \\ R \end{pmatrix} \text{ is nonsingular}. \qquad \textbf{(55)}$$

Then

$$\theta = \begin{pmatrix} R_* \\ R \end{pmatrix}^{-1} \begin{pmatrix} \beta \\ q \end{pmatrix} = R_1\beta + R_2 q, \qquad \textbf{(56)}$$

say. Substituting **(56)** in model **(3)** yields the restricted model

$$y_j - x_j^T R_2 q = x_j^T R_1 \beta + u_j, \quad j = 1, \ldots, n. \qquad \textbf{(57)}$$

Now estimate the parameter vector β by least squares, and compute the Residual Sum of Squares RSS_0 involved. Then

THEOREM 7: *The test statistic* $\hat{F}$ *defined by* **(52)** *is equal to*

$$\hat{F} = \frac{(RSS_0 - RSS)/r}{RSS/(n-k)}. \tag{58}$$

Proof: Note that

$$RSS_0 = \min_{R\hat{\theta}_0 = q} (y - X\hat{\theta}_0)^T (y - X\hat{\theta}_0). \tag{59}$$

The Lagrange function for this minimization problem is:

$$\mathcal{L} = y^T y - 2\hat{\theta}_0^T X^T y + \hat{\theta}_0^T X^T X \hat{\theta}_0 - 2(R\hat{\theta}_0 - q)^T \lambda. \tag{60}$$

The first-order conditions for the minimum are:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{\theta}_0^T} &= -2X^T y + 2X^T X\hat{\theta}_0 - 2R^T\lambda = 0 \Rightarrow \hat{\theta}_0 = \left(X^T X\right)^{-1} X^T y + \left(X^T X\right)^{-1} R^T \lambda, \\ \frac{\partial \mathcal{L}}{\partial \lambda^T} &= -2(R\hat{\theta}_0 - q) = 0 \Rightarrow R\hat{\theta}_0 = q. \end{aligned} \tag{61}$$

Thus,

$$\hat{\theta}_0 = \hat{\theta} + \left(X^T X\right)^{-1} R^T \lambda, \tag{62}$$

and

$$q = R\hat{\theta}_0 = R\hat{\theta} + R\left(X^T X\right)^{-1} R^T \lambda, \tag{63}$$

hence

$$\lambda = -\left(R(X^T X)^{-1} R^T\right)^{-1} (R\hat{\theta} - q). \tag{64}$$

Substituting **(64)** in **(62)** yields:

$$\hat{\theta}_0 = \hat{\theta} - \left(X^T X\right)^{-1} R^T \left(R(X^T X)^{-1} R^T\right)^{-1} (R\hat{\theta} - q). \tag{65}$$

Therefore

$$\begin{aligned} RSS_0 &= \left(y - X\hat{\theta} - X(X^TX)^{-1}R^T\left(R(X^TX)^{-1}R^T\right)^{-1}(R\hat{\theta} - q)\right)^T \\ &\times \left(y - X\hat{\theta} - X(X^TX)^{-1}R^T\left(R(X^TX)^{-1}R^T\right)^{-1}(R\hat{\theta} - q)\right) \\ &= (y - X\hat{\theta})^T(y - X\hat{\theta}) - 2(y - X\hat{\theta})^T\left(X(X^TX)^{-1}R^T(R(X^TX)^{-1}R^T)^{-1}(R\hat{\theta} - q)\right) \\ &+ \left(X(X^TX)^{-1}R^T(R(X^TX)^{-1}R^T)^{-1}(R\hat{\theta} - q)\right)^T\left(X(X^TX)^{-1}R^T(R(X^TX)^{-1}R^T)^{-1}(R\hat{\theta} - q)\right) \end{aligned} \quad \textbf{(66)}$$

Since

$$(y - X\hat{\theta})^TX = 0 \quad \textbf{(67)}$$

(why?), this expression simplifies to:

$$RSS_0 = RSS + (R\hat{\theta} - q)^T\left(R(X^TX)^{-1}R^T\right)^{-1}(R\hat{\theta} - q). \quad \textbf{(68)}$$

Using the fact that $s^2 = RSS/(n-k)$, the theorem follows. Q.E.D.

Some econometric software packages automatically report the F-statistic. This statistic is the test statistic of the F-test that all the slope coefficients are zero, in a linear regression model with an intercept. Thus let the model be as in **(1)**, with $x_{1,j} = 1$. Then the null hypothesis involved is that

$$H_0: \theta_2 = \theta_3 = ... = \theta_k = 0. \quad \textbf{(69)}$$

This null hypothesis is equivalent to the hypothesis that the model can be simplified to

$$y_j = \theta_1 + u_j. \quad \textbf{(70)}$$

As is easy to verify, the least squares estimator $\hat{\theta}_1$ of the parameter θ_1 in model **(70)** is just the sample mean of the y_j's:

$$\hat{\theta}_1 = \bar{y} = (1/n)\sum_{j=1}^{n} y_j \quad \textbf{(71)}$$

and the RSS of model **(70)** is also called the Total Sum of Squares (TSS):

$$TSS = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2. \qquad \textbf{(72)}$$

Thus, the F-test involved is:

$$\hat{F} = \frac{(TSS - RSS)/(k-1)}{RRR/(n-k)}, \qquad \textbf{(73)}$$

which under the null hypothesis **(69)** has an $F_{k\text{-}1,n\text{-}k}$ distribution.

7. The R-square

The R^2 statistic compares the RSS of the model **(1)** with intercept (thus $x_{1,j} = 1$) with the RSS (= TSS) of model **(70)**:

$$R^2 = 1 - \frac{RSS}{TSS}. \qquad \textbf{(74)}$$

It is easy to verify that $0 \leq R^2 \leq 1$, where the value $R^2 = 0$ corresponds to RSS = TSS, hence the explanatory variables $x_{i,j}$, $i = 2,...,k$, in model **(1)** do not contribute anything at all to the explanation of y_j. The case $R^2 = 1$ corresponds to RSS = 0, hence Var(u_j) = 0. The model then fits without error. The R^2 is related to the F statistic **(73)**:

$$\hat{F} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}. \qquad \textbf{(75)}$$

8 Relaxing the assumption of non-stochastic regressors

As said before, Assumption 1 is too restrictive for economic data. However, we may replace Assumptions 1-4 by the following assumptions:

ASSUMPTION 1$^{(*)}$: *Conditionally on the random variables in the matrix X, the error vector u in model* **(5)** *satisfies* $u \sim N_n(0, I_n)$.

ASSUMPTION 2$^{(*)}$: $P(\det(X^T X) > 0) = 1$.

ASSUMPTION 3$^{(*)}$: $\operatorname{plim}_{n\to\infty} X^T X/n = Q$, *where Q is a positive definite matrix*.

Then:

THEOREM 8: *With Assumptions 1-3 replaced by Assumptions 1*$^{(*)}$ *and 2*$^{(*)}$, *and Assumption 4 by Assumption 3*$^{(*)}$, *Theorem 1 now holds conditionally on the random variables in the matrix X, and Theorems 2-7 hold unconditionally.*

Proof: Exercise.

Hints: It is easy to verify that all the previous results go through conditionally on X. In order to prove that s^2 and $\hat{\theta}$ are still independent (cf. Theorem 3), observe that the joint density of s^2 and $\hat{\theta}$ conditional on X is now the product of the conditional density of s^2 and the conditional density of $\hat{\theta}$, given X. But the conditional density of s^2 does not depend on X, because $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$ conditionally on X and therefore also unconditionally (why?). Integrating X out in the joint conditional density of s^2 and $\hat{\theta}$ then yields the product of the unconditional density of s^2 and the unconditional density of $\hat{\theta}$, which proves the result involved. The rest of the conclusion of Theorem 8 can be proved by a similar argument.

A more general version of the above argument is stated in the following lemma:

LEMMA 1: *Let x, y and z be random vector or variables such that y and z are*

conditionally independent, relative to x, i.e., the joint conditional distribution function of y and z, given x, is the product of the conditional distribution function of y and the conditional distribution function of z, given x. If z and x are independent then y and z are independent.

Proof: For convenience assume that the joint distribution of x, y and z is continuous, with marginal densities $f_x(x)$, $f_y(y)$ and $f_z(z)$, and conditional densities $f_{yz}(y,z \mid x)$, $f_y(y \mid x)$ and $f_z(z \mid x)$. Since z and x are independent we have $f_z(z \mid x) = f_z(z)$. Now

$$f_{yz}(y,z) = \int f_{yz}(y,z|x)f_x(x)dx = \int f_y(y|x)f_z(z|x)f_x(x)dx = \int f_y(y|x)f_x(x)dx\, f_z(z)$$
$$= f_y(y)f_z(z). \qquad \textbf{(76)}$$

This result, however, also holds without the assumption that x, y and z are continuously distributed.

9. Large sample theory without the normality assumption

If our sample size n is large, we may even get rid of the normality assumption on the errors u_j in model **(1)**. Assume that the source of the data $\{(y_j,x_j), j = 1,...,n\}$ is a random sample

ASSUMPTION 1$^{(**)}$: *The random vectors* $(y_j, x_j^T)^T$, $j = 1, ..., n, ...,$ (*or the sub-vectors of random variables if one of the components of* x_j *equals* 1) *are i.i.d. Moreover,* $E(y_j|x_j) = x_j^T\theta$, *hence* $E(u_j|x_j) = 0$. *Furthermore,* $Var(u_j|x_j) = \sigma^2 < \infty$, *and* $E(x_j x_j^T) = Q$, *where Q is the same positive definite matrix as in Assumption 3*$^{(*)}$.

Then:

THEOREM 9: *Under Assumptions 1*$^{(**)}$ *and 2*$^{(*)}$ *we have*:

$$E(\hat{\theta}) = \theta, \quad E(s^2) = \sigma^2. \qquad \textbf{(77)}$$

Conditional on X, the least squares estimator $\hat{\theta}$ *is BLUE. Moreover, under the additional*

Assumption 3(*) we have:

$$\operatorname{plim}_{n\to\infty} \hat{\theta} = \theta, \quad \operatorname{plim}_{n\to\infty} s^2 = \sigma^2, \tag{78}$$

$$\sqrt{n}(\hat{\theta} - \theta) \to N_k(0, \sigma^2 Q^{-1}) \text{ in distr. as } n \to \infty, \tag{79}$$

the t-value $\hat{t}_i$ *of the i-th component of* $\hat{\theta}$ *satisfies*

$$\begin{aligned} &(a) \text{ if } \theta_i = 0 \text{ then } \hat{t}_i \to N(0,1) \text{ in distr as } n \to \infty, \\ &(b) \text{ if } \theta_i \neq 0 \text{ then } \operatorname{plim}_{n\to\infty} \hat{t}_i/\sqrt{n} = \theta_i/\sqrt{\sigma^2 e_i^T Q^{-1} e_i} \neq 0, \end{aligned} \tag{80}$$

and the F-statistic $\hat{F}$ *for testing the null hypothesis* $R\theta = q$, *where R is a* $r \times k$ *matrix with rank r, satisfies*

$$\begin{aligned} &(a) \text{ if } R\theta = q \text{ then } r\hat{F} \to \chi_r^2 \text{ in distr as } n \to \infty, \\ &(b) \text{ if } R\theta \neq q \text{ then } \operatorname{plim}_{n\to\infty} r\hat{F}/n = (R\theta - q)^T (RQ^{-1}R^T)^{-1}(R\theta - q)/\sigma^2 > 0. \end{aligned} \tag{81}$$

Proof: Assumption $1^{(**)}$ implies that $E(u|X) = 0$ and $E(uu^T|X) = \sigma^T I_n$, which in their turn imply the unbiasedness of the two estimators (*Proof*: Exercise). Moreover, Assumption $1^{(**)}$ and the law of large numbers imply that $\operatorname{plim}_{n\to\infty} (1/n)X^T u = 0$. Together with Assumptions $2^{(*)}$ and $3^{(*)}$ this result implies **(78)**. Furthermore, Assumption $1^{(**)}$ and the central limit theorem imply that $(1/\sqrt{n})X^T u \to N_k(0, \sigma^2 Q)$ in distribution (*Proof*: Exercise). Together with Assumption $3^{(*)}$ this result implies **(79)**. The rest of Theorem 9 is easy to prove.

10. Tests of structural change: The Chow tests

The classical linear regression model **(1)** assumes that the parameters are the same for all observations. In order to test this crucial hypothesis, the sample is split in say *m* subsamples of sizes

$n_1,...,n_m$, respectively, with $n = \sum_{i=1}^{m} n_i$ and $n_i > k$, where the parameters are allowed to be different across subsamples:

$$\begin{aligned} y_j &= \theta_{1,1}x_{1,j} + ... + \theta_{1,k}x_{k,j} + u_j, \qquad j = 1,...,n_1, \\ y_j &= \theta_{2,1}x_{1,j} + ... + \theta_{2,k}x_{k,j} + u_j, \quad j = n_1 + 1,...,n_1+n_2 \\ &..... \\ y_j &= \theta_{m,1}x_{1,j} + ... + \theta_{m,k}x_{k,j} + u_j, \quad j = \sum_{i=1}^{m-1} n_i + 1,...,n. \end{aligned} \tag{82}$$

For each subsample i we can write the model in matrix form as

$$y^{(i)} = X^{(i)}\theta^{(i)} + u^{(i)}, \; i = 1,..,m, \tag{83}$$

and stacking all these equations yields the *unrestricted model*:

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} X^{(1)} & O & .. & O \\ O & X^{(2)} & .. & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & .. & X^{(m)} \end{pmatrix} \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \\ \vdots \\ \theta^{(m)} \end{pmatrix} + \begin{pmatrix} u^{(1)} \\ u^{(2)} \\ \vdots \\ u^{(m)} \end{pmatrix}. \tag{84}$$

Note that the total number of parameters is $k \times m$. It is not hard to verifythat the least squares estimators $\hat{\theta}^{(i)}$ of the parameter vectors $\theta^{(i)}$ are:

$$\hat{\theta}^{(i)} = \left(X^{(i)\,T}X^{(i)}\right)^{-1}X^{(i)\,T}y^{(i)} \tag{85}$$

and that the RSS_u of model **(84)** is just the sum of the residual sums of squares RSS_i of the regressions **(83)**:

$$RSS_u = \sum_{i=1}^{m} RSS_i \tag{86}$$

The null hypothesis that model **(3)** is correct corresponds to the set of hypotheses

$$H_0: \begin{array}{c} \theta^{(1)} = \theta^{(2)} \\ \theta^{(2)} = \theta^{(3)} \\ \vdots \\ \theta^{(m-1)} = \theta^{(m)} \end{array}, \qquad \textbf{(87)}$$

hence the total number of restrictions is $(m-1)\times k$. Therefore:

THEOREM 10: (*Chow test*) *The F test of the hypothesis* **(87)** *is*:

$$\hat{F} = \frac{(RSS - \sum_{i=1}^{m} RSS_i)/((m-1)k)}{(\sum_{i=1}^{m} RSS_i)/(n-km)} \qquad \textbf{(88)}$$

where RSS is the residual sum of squares of the restricted model **(3)**. *Under the null hypothesis involved this F statistic is* $F_{(m-1)k,n-km}$ *distributed*.

If one of the subsample sizes n_i is less or equal to the number k of parameters, it is possible to fit the corresponding regression **(83)** without any residuals, so that $RSS_i = 0$ (why?). This case, with $m = 2$, may occur for example in a situation where the model is used for prediction and we want to test whether the predicted dependent variable y_n is governed by the same model as for the observations $j \leq n$-1. In that case the degrees of freedom of the F test will be different, as we will demonstrate now for the case $m = 2$. Thus, we want to test the validity of the null model **(1)** against the alternative model

$$\begin{aligned} y_j &= \theta_{1,1}x_{1,j} + ... + \theta_{1,k}x_{k,j} + u_j, \quad j = 1,...,n_1, \\ y_j &= \theta_{2,1}x_{1,j} + ... + \theta_{2,k}x_{k,j} + u_j, \quad j = n_1 + 1,...,n, \end{aligned} \qquad \textbf{(89)}$$

where

$$n_2 = n - n_1 \leq k. \qquad \textbf{(90)}$$

Since in this case $RSS_2 = 0$, the F test now compares the *RSS* of the restricted model **(1)** for the full sample of size n, with the RSS_1 of the regression for the larger subsample of size n_1:

THEOREM 11: (*Chow predictive test*) *The F test of the null hypothesis that model* **(1)** *applies to all observations against the alternative that model* **(89)** *with* **(90)** *applies, takes the form*

$$\hat{F} = \frac{(RSS - RSS_1)/n_2}{RRS_1/(n_1 - k)}, \tag{91}$$

which under the null hypothesis has an $F_{n_2, n_1 - k}$ *distribution.*

Proof: Adopting the matrix notation **(83)** of the models for the two subsamples, we can write

$$\begin{aligned} RSS &= u^T\left(I - X(X^TX)^{-1}X^T\right)u \\ &= u^T\left(\begin{pmatrix} I_{n_1} & O \\ O & I_{n_2} \end{pmatrix} - \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}(X^TX)^{-1}(X^{(1)T}, X^{(2)T})\right)u \\ &= u^T\left(\begin{pmatrix} I_{n_1} & O \\ O & I_{n_2} \end{pmatrix} - \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix}\right)u = u^TMu, \end{aligned} \tag{92}$$

say, and

$$\begin{aligned} RSS_1 &= u^{(1)T}\left(I_{n_1} - X^{(1)}(X^{(1)T}X^{(1)})^{-1}X^{(1)T}\right)u^{(1)} \\ &= u^T\left(\begin{pmatrix} I_{n_1} & O \\ O & I_{n_2} \end{pmatrix} - \begin{pmatrix} X^{(1)}(X^{(1)T}X^{(1)})^{-1}X^{(1)T} & O \\ O & I_{n_2} \end{pmatrix}\right)u = u^TM_1u, \end{aligned} \tag{93}$$

say, hence

$$RSS - RSS_1$$

$$= u^T \begin{pmatrix} X^{(1)}(X^{(1)T}X^{(1)})^{-1}X^{(1)T} & O \\ O & I_{n_2} \end{pmatrix} u - u^T \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix} u \quad \textbf{(94)}$$

$$= u^T M_2 u ,$$

say. Clearly, the matrix M_1 is idempotent, with rank(M_1) = trace(M_1) = n_1 - k. Also M_2 is idempotent, because

$$\begin{pmatrix} X^{(1)}(X^{(1)T}X^{(1)})^{-1}X^{(1)T} & O \\ O & I_{n_2} \end{pmatrix} \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix}$$
$$= \begin{pmatrix} X^{(1)}(X^TX)^{-1}X^{(1)T} & X^{(1)}(X^TX)^{-1}X^{(2)T} \\ X^{(2)}(X^TX)^{-1}X^{(1)T} & X^{(2)}(X^TX)^{-1}X^{(2)T} \end{pmatrix} \quad \textbf{(95)}$$

hence $M_1M = MM_1 = M_1$ and thus $(M - M_1)^2 = M - M_1$. The rank of M_2 is: rank(M_2) = trace(M_2) = trace(M) - trace(M_1) = (n - k) - (n_1 - k) = n - n_1 = n_2. Using an MND result (which one?), it now follows that under the null hypothesis

$$(RSS - RSS_1)/\sigma^2 \sim \chi^2_{n_2}, \quad RSS_1/\sigma^2 \sim \chi^2_{n_1-k}. \quad \textbf{(96)}$$

Finally it remains to show that the two random variables involved are independent. This follows from the fact that $M_1M_2 = M_1(M - M_1) = M_1M - M_1M_1 = M_1 - M_1 = O$, and an MND result (which one?). Q.E.D.

11. Tests of partial structural change

If some but not all of the parameters are allowed to be different across subsamples, the F tests in section 10 no longer apply. We then have to merge the different models by using dummy variables. To illustrate this, consider the bivariate regressions

$$y_j = \theta_{1,1} + \theta_{2,1}x_j + u_j, \quad j = 1,\ldots,n_1$$
$$y_j = \theta_{1,2} + \theta_{2,2}x_j + u_j, \quad j = n_1+1,\ldots,n \qquad \textbf{(97)}$$

Consider first the case with maintained hypothesis

$$Case\ 1: \quad \theta_{2,1} = \theta_{2,2} = \theta_2$$

(A maintained hypothesis is a hypothesis that is assumed to hold under both the null and alternative hypothesis). Rewriting $\theta_{1,1} = \theta_1$, $\theta_{1,2} = \theta_1 + \theta_3$, and creating the dummy variable

$$d_j = 0 \ for \ j = 1,\ldots,n_1, \quad d_j = 1 \ for \ j = n_1+1,\ldots,n, \qquad \textbf{(98)}$$

we can now merge the two models involved into a single reparametrized model:

$$y_j = \theta_1 + \theta_2 x_j + \theta_3 d_j + u_j, \quad j = 1,\ldots,n, \qquad \textbf{(99)}$$

and the null hypothesis that the two models are the same is now equivalent to the hypothesis $\theta_3 = 0$, which can de tested by the t-test.

The case with maintained hypothesis

$$Case\ 2: \quad \theta_{1,1} = \theta_{1,2} = \theta_1$$

is similar. Rewriting $\theta_{2,1} = \theta_2$, $\theta_{2,2} = \theta_2 + \theta_3$, we can merge the two models involved into a single model:

$$y_j = \theta_1 + \theta_2 x_j + \theta_3 (d_j x_j) + u_j, \quad j = 1,\ldots,n \qquad \textbf{(10}$$

and the null hypothesis that the two models are the same is again equivalent to the hypothesis $\theta_3 = 0$. The extension of this approach to multivariate models is straightforward and left to the reader.

Exercises:

1. Prove **(25)**.
2. Which MND[1] result is used in the proof of Theorem 2?
3. Prove **(29)**.
4. Which MND results are used to show that s^2 and $\hat{\theta}$ are independent (c.f. Theorem 3)?
5. Which MND results are used in **(36)** and **(37)**?
6. Give the details of the proof of Theorem 4. In particular, indicate which MND results are used.
7. Prove **(42)**.
8. Prove **(43)**.
9. Give the details of the proof of Theorem 5. In particular, indicate which MND results are used.
10. Which MND results are used in **(49)** and **(50)**?
11. Give the details of the proof of Theorem 6. In particular, indicate which MND results are used.
12. Why is **(67)** true?
13. Give the details of the proof of Theorem 8.
14. Give the details of the proof of Theorem 9.
15. Which MND results are used in **(96)**?

[1] Recall that MND stands for Multivariate Normal Distribution.