A FULL VERSION OF THE MATHEMATICAL APPENDIX (§10) TO

THE PHYLOGENETIC REGRESSION

BY A. GRAFEN

Animal Behaviour Research Group, Zoology Department, Oxford University

MATHEMATICAL APPENDIX

(a) Preliminary remarks.

The purpose of this appendix is to state results of importance to the phylogenetic regression. The mathematics done is all quite simple, but in order to express it economically it has been necessary to adopt a rather formal approach. A major notational problem is the formal treatment of an arbitrary phylogeny. In these preliminary remarks, the meaning and relevance of the four theorems is discussed. Throughout the appendix, formal remarks are made to explain the direction of the developing argument. In many cases the actual objects of interest are not mentioned in the mathematics at all. These objects are statistical tests. The first is the *standard regression*. The variables involved are a y-variable y, a set of x-variables X to be controlled for, and a set of x-variables Z to be tested for. The regression is defined by

$$\mathbf{E}(y) = \mathbf{1}_t\mu + X\beta + Z\gamma, \qquad (y-X\beta-Z\gamma) \propto \mathbf{N}(0,V),$$

where $\mathbf{1}_t$ is the constant term, V is defined by

$$V_{ij}(\rho) = (1 - h_{ij}^{\rho}),$$

and h_{ij} is the height in the initial working phylogeny at which the paths to species i and j diverge. “$\propto$" is used to mean that the variance-covariance matrix of the error is assumed only to be proportional to V, not necessarily equal to it. For the purpose of this appendix the path segment lengths are fixed, so that ρ is considered known.

The first theorem states that the standard regression is equivalent to the *long regression*, defined by

$$\mathbf{E}(Ly) = S\delta + LX\beta + LZ\gamma, \qquad (Ly-S\delta-LX\beta-LZ\gamma) \propto \mathbf{N}(0,C),$$

in which L, S and C are matrices defined formally later. The two regressions are shown to be equivalent in the sense that the residual sum of squares of the long regression, concentrated for β and δ, is the same function of y, Z and γ as the residual sum of squares of the standard regression concentrated for μ and β. This shows that the significance tests for γ=0, controlling for $\mathbf{1}_t$ and X in the case of the standard regression and for LX and S in the case of the long regression, will yield the same test statistic with the same distribution. Each datapoint in the long regression represents the deviation of a node's value from its parent node's value. The data in this form is suitable for defining the randomization test explained in §3(c). It is important that C is a diagonal matrix, so that this theorem allows the standard regression to be fitted by a package which cannot handle non-diagonal variance-covariance matrices. GLIM is such a package. The reason it is necessary to prove this first theorem is to show that the formulae for L and C are correct - their forms are far from obvious *a priori*. L represents the process of “hanging on the tree" described in §3(a).

The second and third theorems concern the *short regression*, defined by

$$\mathbf{E}(GC^{-1}Ly) = GC^{-1}LX\beta + GC^{-1}LZ\gamma, \qquad (GC^{-1}L(y-X\beta-Z\gamma)) \propto \mathbf{N}(0,I),$$

The distribution of $(GC^{-1}L(y-X\beta-Z\gamma))$ is understood as a distribution conditional on G, as G is a random matrix because it depends on the value of y. The second theorem states that the process of performing the long regression, defining the random linear contrasts GC^{-1} and forming the elements of the short regression does indeed result in the same, standard, statistical test as the short regression. This is shown by proving that conditional on G, the residual in the short regression after regression of $GC^{-1}Ly$ on $GC^{-1}LX$ has the same probability density whether the randomness arises through ε, the error in the standard regression, as transmuted by construction of G and the formation of the short regression; or whether the randomness is assumed to arise as a $\mathbf{N}(0,I)$ variable in the short regression itself. The first reason it is necessary to prove this theorem is to show that the formula for G is correct. The second reason is that G is a random matrix, as it depends on ε. In general, using contrasts G that depend on ε will violate the standard formulae for the variances and covariances of the contrasts, which rely on fixed G. As was seen in the simulations in §5, the short regression has high mean square error in its parameter estimate under the null hypothesis, and has biassed estimates under the alternative hypothesis. It is therefore not at all obvious that the short regression will be valid, but, as the theorem shows, it is.

The third theorem states that the short regression is equivalent to the *long regression with T*, defined by

$$\mathbf{E}(Ly) = S\delta + LX\beta + T\tau + LZ\gamma, \qquad (Ly\text{-}S\delta\text{-}LX\beta\text{-}T\theta\text{-}LZ\gamma) \propto \mathbf{N}(0,C),$$

T is a matrix representing a set of artificial variables added to the long regression to ensure that no matter what value Z may take, the residuals after regression on S, LX, T and LZ will remain proportional, within each radiation separately, to the residuals after regression on S and LX alone. T therefore depends on y, and like G is a random matrix. Equivalence means that the residual sum of squares for the long regression with T, concentrated for β, δ and θ, is the same function of y, Z and γ as the residual sum of squares of the short regression concentrated for β. The theorem is proved to show that the phylogenetic regression can be interpreted as conditioning within the standard regression on the patterns of the residual in each radiation, in the sense of “pattern" explained in §3(c).

The fourth theorem shows that the randomization test explained in §3(c) and defined formally below, is equivalent to the short regression in the sense that the null distribution of the test statistic of the randomization test is also an F-distribution with the required degrees of freedom.

As well as these four results, the mathematical development defines the matrices used to construct the long and short regressions, and so formally defines the phylogenetic regression.

The published version of this appendix contained only the definitions, theorems and remarks. This full version contains in addition all lemmas and proofs needed to prove the theorems.

(b) Mathematical development

A preliminary note on matrix notation. I shall define matrices as $\Delta\times\Theta$, where Δ and Θ are finite sets, rather than as $m\times n$, were m and n are integers. An $\Delta\times\Theta$ matrix A will have elements A_{ij}, where $i\in\Delta$ and $j\in\Theta$. Where a matrix is defined as $m\times n$ or $\Delta\times n$, the integers m and n should be understood as shorthand for the sets $\{1, 2 \ldots m\}$ and $\{1, 2 \ldots n\}$. The advantage of this notation is that if Δ' and Θ' are subsets of Δ and Θ, respectively, then a submatrix A' can be concisely defined as the $\Delta'\times\Theta'$ submatrix of A.

Lemma 1. If A is an $n_A\times n_A$ matrix of full rank, B is an $n_A\times n_B$ matrix of full rank, λ is a scalar, $n_B\leq n_A$, I represents the $n_B\times n_B$ identity matrix and $(I+\lambda B^TA^{-1}B)$ is of full rank, then

$$(A + \lambda BB^T)^{-1} = A^{-1} - \lambda A^{-1}B(I+\lambda B^TA^{-1}B)^{-1}B^TA^{-1}$$

Proof. By multiplication of the proposed inverses and collection of terms in $B(...)B^TA^{-1}$.

Lemma 2. If A is an $n_A\times n_A$ positive definite matrix, B and D are $n_A\times n_B$ and $n_A\times n_D$ matrices of full rank, $n_B+n_D=n_A$, $B^TA^{-1}D=0$, and I is the $n_A\times n_A$ identity matrix, then

$$I - D(D^TA^{-1}D)^{-1}D^TA^{-1} = B(B^TA^{-1}B)^{-1}B^TA^{-1}$$

Proof. Any $n_A \times 1$ vector x can be expressed uniquely as Bb+Dd, for some $n_B \times 1$ vector b and some $n_D \times 1$ vector d, because together the columns of B and D span the whole space. The LHS pre-multiplies x into Bb, and so does the RHS. They are therefore the same matrix.

Lemma 3. If A is an $n_A \times n_A$ positive definite matrix, and B is an $n_A \times n_B$ matrix of full rank, $n_B < n_A$, then there exists an $n_A \times n_D$ matrix D such that

i) D is of full rank

ii) $B^T A^{-1} D = 0$

iii) $n_B + n_D = n_A$

Proof. Let the inner product of two vectors a and b of $\mathbf{R}^{n_A}$ be defined by $a^T A^{-1} b$. Then D must be chosen so that its columns minimally span the subspace complementary to that spanned by the columns of B.

Lemma 4. If A is an $n_A \times n_A$ positive definite matrix, and B, D, E and F are full rank matrices of size $n_A \times n_B$, $n_A \times n_D$, $n_A \times n_E$, and $n_A \times n_F$, respectively, and rk(B|E)=rk(B)+rk(E), then if

i) $n_D + n_B = n_A$

ii) $D^T A^{-1} B = 0$, and

iii) $F^T A^{-1} (B|E) = 0$,

then there exists an $n_D \times n_F$ matrix H such that H is of full rank and F=DH.

Proof. The columns of F must lie in the subspace spanned by the columns of D. The ith column of H is the linear combination of the columns of D which equals the ith column of F, $1\leq i\leq n_D$. H is of full rank because $rk(F)\leq\min\{rk(D),rk(H)\}$, $rk(F)=n_F$, and $n_F\leq n_D$.

Definition of Π, Π_h, Π_s, Π_t, Π_i, Π_{di}. These definitions are made with respect to the working phylogeny. Let Π be the set of all nodes, Π_t the set of species nodes, Π_h the set of higher (i.e. non-species) nodes and Π_s the set of all nodes except the root. Let Π_i, $i\in\Pi$, be the set of species nodes which are descendants of (or equal to) node i. Let Π_{di}, $i\in\Pi_h$, be the set of daughter nodes of node i. Associate each node with a distinct integer, to establish an arbitrary ordering over Π.

Definition of P, P_i. Let P denote the partition $\{\Pi_{di}\}_{\in\Pi_h}$ of Π_s, and let P_i, $i\in\Pi$, denote the partition $\{\Pi_j\}_{j\in\Pi_{di}}$ of Π_i.

Definition of n, n_t, n_s, n_h. Let n be the number of nodes in the working phylogeny, n_t be the number of species nodes, n_s be n-1, and n_h be the number of higher nodes. Note every species is either a species node or a higher node but not both, so that $n_h+n_t=n$. It follows that $n_s-n_h=n_t-1$.

Definition of ′. Let $i'\in\Pi_h$ denote the parent node of i, $i\in\Pi_s$.

Definition of κ_i, h_i. Let κ_i be arbitrary non-negative real numbers representing the length of the path segment between i and i′, $i\in\Pi_s$, with $\kappa_i>0$ if $i\in\Pi_t$. Let h_i be the summed length of the path segments between the root and i, $i\in\Pi$.

Definition of $\alpha(i,j)$. Let $\alpha(i,j)$ be the lowest common ancestor of i and j, $i,j\in\Pi$.

Remark. The h_i just defined are related to the h_{ij} used in the body of the paper by the relationship $h_{\alpha(i,j)} = 1 - h_{ij}^{\rho}$. The working phylogeny as used in the appendix is taken as having already undergone transformation by ρ.

Definition of Ω_t, Ω_s, Ω_{ti}, Ω_{si}. Let Ω_t be the set of column vectors with real elements indexed by Π_t, and let Ω_{ti}, $i\in\Pi$, be the subspace of Ω_t with only those elements indexed by Π_i. Let Ω_s be the set of column vectors with real elements indexed by Π_s, and let Ω_{si}, $i \in \Pi_h$, be the subspace of Ω_s with only those elements indexed by Π_{di}.

Remark. The definitions of the various Π's allows means at higher nodes to be dealt with in the same way as species values. Ω_t and Ω_s are the dataspaces of the standard and long regressions, respectively.

Definition of $\mathbf{1}_{ti}$, $\mathbf{I}_{ti}$, $\mathbf{1}_{si}$. Let $\mathbf{1}_{ti} \in \Omega_{ti}$, $i \in \Pi$, be the vector each of whose elements equals one. Let $\mathbf{I}_{ti}$ be the identity matrix over Ω_{ti}. Let $\mathbf{1}_{si} \in \Omega_{si}$ be the vector each of whose elements equals one.

Definition of U_i. Let U_i, $i \in \Pi$, be the $\Pi_i \times \Pi_t$ matrix defined by

$$(U_i)_{jk} = \begin{matrix} 1 & \quad j=k \\ 0 & \quad j \neq k \end{matrix}$$

Remark. U_i is a matrix which picks out from a vector $x \in \Omega_t$ those elements indexed by elements of Π_i. $(U_i x) \in \Omega_i$, and equals x over those elements held in common. U_i^T transforms a vector $x \in \Omega_i$ into a vector which is an element of Ω_t, equals x in those elements indexed in common, and equals zero elsewhere.

Lemma 5. If $\alpha(i,k)=i$, $\alpha(j,m)=j$ and $\alpha(i,j) \neq i,j$, then $\alpha(k,m)=\alpha(i,j)$.

Proof. This is obvious from the nature of a tree.

Definition of V. Let V be the $\Pi_t \times \Pi_t$ matrix defined by

$$V_{ij} = h_{\alpha(i,j)}, \text{ for } i,j \in \Pi_t$$

Extension of subscript notation for V. As an extension of the usual subscript notation, let V_{ij} also be defined when i and j are not necessarily species nodes, as the $\Pi_i \times \Pi_j$ submatrix of V. Further, let V_i denote V_{ii}.

Lemma 6. $V_{ij} = h_{\alpha(i,j)}\mathbf{1}_i\, \mathbf{1}_j^T$ $\alpha(i,j)\neq i,j$.

Proof. By definition of V, $V_{km} = h_{\alpha(k,m)}$, $k,m\in \Pi_t$. Under the conditions of the lemma, $\alpha(k,m)=\alpha(i,j)$ $\forall k,m$, by Lemma 5, proving the result.

Remark. The restrictions in the following lemma are needed, as V_i - $h_i\, \mathbf{1}_i\, \mathbf{1}_i^T$ equals zero for species nodes, and V_i - $h_{i'}\, \mathbf{1}_i\, \mathbf{1}_i^T$ is undefined for the root node.

Lemma 7. V_i, $i\in \Pi$, and V_i - $h_i\, \mathbf{1}_i\, \mathbf{1}_i^T$, $i\in \Pi_h$, and V_i - $h_{i'}\, \mathbf{1}_i\, \mathbf{1}_i^T$, $i\in \Pi_s$, are all of full rank and positive definite.

Proof. For this proof, let PD denote “positive definite". First it is shown that

If V_i - $h_i\, \mathbf{1}_i\, \mathbf{1}_i^T$ is PD then V_i - $h_{i'}\, \mathbf{1}_i\, \mathbf{1}_i^T$ is PD, $i\in \Pi_s$ (A1)

If the left hand matrix is A and the right hand is B, then we have

$$B = A + (h_i - h_{i'})\mathbf{1}_i\, \mathbf{1}_i^T$$

Now $(h_i - h_{i'})$ is non-negative because by definition of h it equals κ_i. Hence for any conformable vector x,

$$x^T Bx = x^T Ax + (h_i - h_{i'})(x^T\mathbf{1}_i)^2$$

The second term on the right is non-negative for arbitrary x. Hence if A is positive definite, then so is B, as required. Similar arguments, using the non-negativity of $h_{i'}$ or h_i instead of $(h_i-h_{i'})$, show that

If V_i - $h_{i'}\, \mathbf{1}_i\, \mathbf{1}_i^T$ is PD then V_i is PD, $i\in \Pi_s$ (A2a)

If V_i - $h_i\, \mathbf{1}_i\, \mathbf{1}_i^T$ is PD then V_i is PD, $i\in \Pi$ (A2b)

Next it is shown that

If V_j - $h_i\, \mathbf{1}_j\, \mathbf{1}_j^T$ is PD for all $j\in \Pi_{di}$, then V_i - $h_i\, \mathbf{1}_i\, \mathbf{1}_i^T$ is PD, $i\in \Pi_h$ (A3)

This follows because, by Lemma 6, considered as a $\Pi_{di}\times\Pi_{di}$ matrix according to the partition P_i of Π_i, $V_i - h_i \mathbf{1}_i \mathbf{1}_i^T$ is diagonal and its jth element is $V_j - h_i \mathbf{1}_j \mathbf{1}_j^T$. A block diagonal matrix is positive definite if all of its diagonal blocks are positive definite.

The three results (A1), (A2) and (A3) now allow the lemma to be proved. (A1) and (A3) show that the property that $V_i - h_{i'} \mathbf{1}_i \mathbf{1}_i^T$ is PD is inherited from daughters to parents in the sense that if all the daughters of node i possess it then so does node i, provided it is defined for that node. But for a species node, $i\in\Pi_t$, $V_i - h_{i'} \mathbf{1}_i \mathbf{1}_i^T$ is a 1×1 matrix whose element equals κ_i, which is by definition strictly positive for $i\in\Pi_t$. Hence the property that $V_i - h_{i'} \mathbf{1}_i \mathbf{1}_i^T$ is PD is possessed by all species nodes and so is inherited by all $i\in\Pi_s$. Now by (A3), $V_i - h_i \mathbf{1}_i \mathbf{1}_i^T$ is PD for $i\in\Pi_h$. The only case remaining in the statement of the lemma is V_i. We have now shown for every node either $V_i - h_{i'} \mathbf{1}_i \mathbf{1}_i^T$ is PD, or $V_i - h_i \mathbf{1}_i \mathbf{1}_i^T$ is PD; and by (A2a) and (A2b) this is sufficient to show that V_i is PD, $i\in\Pi$. Positive definiteness has been established for all the cases in the statement of the lemma.

Finally, it is sufficient to note that a positive definite matrix must be of full rank. This completes the proof.

Definition of σ_i^2. Let $\sigma_i^2 = (\mathbf{1}_i^T V_i^{-1} \mathbf{1}_i)^{-1}$, $i\in\Pi$. Lemma 7 shows that both the inverses exist.

Remark. σ_i^2 is the sampling variance of the mean of all the species below node i, of a variable whose variance-covariance matrix is V.

Lemma 8.

$$\mathbf{1}_i^T (V_i - h_i\mathbf{1}_i\mathbf{1}_i^T)^{-1} \mathbf{1}_i = \sum_{j\in\Pi_{di}} (\sigma_j^2 - h_i)^{-1}, \ i\in\Pi_h$$

$$\sigma_i^2 = \left(\sum_{j \in \Pi_{di}} (\sigma_j^2 - h_i)^{-1} \right)^{-1} + h_i \,, \ i \in \Pi_h$$

Proof. The inverses employed throughout this proof are shown to exist by Lemma 7. According to Lemma 6, $V_i - h_i \mathbf{1}_i \mathbf{1}_i^T$, $i \in \Pi_h$, is diagonal when considered as a $\Pi_{di} \times \Pi_{di}$ matrix according to the partition P_i. Hence

$$\mathbf{1}_i^T (V_i - h_i \mathbf{1}_i \mathbf{1}_i^T)^{-1} \mathbf{1}_i = \sum_{j \in \Pi_{di}} \mathbf{1}_j^T (V_j - h_i \mathbf{1}_j \mathbf{1}_j^T)^{-1} \mathbf{1}_j$$

But $V_j - h_i \mathbf{1}_j \mathbf{1}_j^T$ can be inverted by Lemma 1, pre- and post-multiplied by $\mathbf{1}_j$, and rearranged to give

$$\mathbf{1}_j^T (V_j - h_i \mathbf{1}_j \mathbf{1}_j^T)^{-1} \mathbf{1}_j = \frac{\mathbf{1}_j^T V_j^{-1} \mathbf{1}_j}{1 - h_i \mathbf{1}_j^T V_j^{-1} \mathbf{1}_j} = \frac{1}{\sigma_j^2 - h_i}$$

and so

$$\mathbf{1}_i^T (V_i - h_i \mathbf{1}_i \mathbf{1}_i^T)^{-1} \mathbf{1}_i = \sum_{j \in \Pi_{di}} (\sigma_j^2 - h_i)^{-1}$$

establishing the first part of the lemma. V_i in the form $(V_i - h_i \mathbf{1}_i \mathbf{1}_i^T) + h_i \mathbf{1}_i \mathbf{1}_i^T$ can be inverted by Lemma 1, pre- and post-multiplied by $\mathbf{1}_i$, and then rearranged to give

$$(\mathbf{1}_i^T V_i^{-1} \mathbf{1}_i)^{-1} = (\mathbf{1}_i^T (V_i - h_i \mathbf{1}_i \mathbf{1}_i^T)^{-1} \mathbf{1}_i)^{-1} + h_i$$

Together, these last two equations establish the second part of the lemma, completing the proof.

Definition of f_i. Let $f_i \in \Omega_t$, $i \in \Pi$, be defined by $f_i = U_i^T V_i^{-1} \mathbf{1}_i (\mathbf{1}_i^T V_i^{-1} \mathbf{1}_i)^{-1}$.

<u>Lemma</u> <u>9</u>. $$f_i^T Vf_j = \begin{cases} h_{\alpha(i,j)} & \alpha(i,j)\neq i,j. \\ \sigma_i^2 & \alpha(i,j)=i. \end{cases}$$

<u>Proof</u>. After expanding the f_i and f_j, it is necessary to notice that $U_i VU_j^T = V_{ij}$, and that when $\alpha(i,j)=i$, $U_i^T U_i U_j^T = U_j^T$ and $U_j U_i^T \mathbf{1}_i = \mathbf{1}_j$. With the definition of σ_i^2 and Lemma 6 on the form of V_{ij}, the results then follow immediately by direct computation.

<u>Definition</u> <u>of</u> L, L_i, W <u>and</u> K. Let L be a $\Pi_s\times\Pi_t$ matrix whose ith row is denoted by L_i, and defined by

$$L_i = [f_i^T - f_{i'}^T]$$

and let W be a $\Pi_s\times\Pi_s$ matrix defined by $W=LVL^T$. Let K be a $\Pi_t\times\Pi_s$ matrix defined by

$$K_{ij} = \begin{cases} 1 & \alpha(i,j)=j \\ 0 & \text{o.w.} \end{cases}$$

<u>Remark</u>. L is the matrix of linear contrasts which transforms the variables of the standard regression into the corresponding variables of the long regression. f_i is a vector which maps (by taking the inner product) a vector of species values into the mean value for species below node i. So f_i-$f_{i'}$ produces the deviation of the mean of the species below node i from the mean of the species below the parent of node i. If the variance covariance matrix of a random vector x is V, then that of Lx is W. K is a matrix with a row for every species, and a column for every node except the root. An element equal to 1 indicates that the column-node is an ancestor of (or is equal to) the row-node.

<u>Notational</u> <u>convention</u> <u>of</u> <u>bracketed</u> <u>subscripts</u>. Any array dimension indexed by Π_s can also be considered to be indexed by Π_h, according to the partition P of Π_s. It is

convenient to be able to use both forms of indexing explicitly. Accordingly unbracketed subscripts will refer in the usual way to indexing by Π_s, while bracketed subscripts will refer to the partitional indexing. Thus W_{ij} is a single element of the matrix W, defined for $i,j\in\Pi_s$. $W_{(i)j}$ is a $\Pi_{di}\times 1$ vector defined for $i\in\Pi_h$, $j\in\Pi_s$. $W_{(ij)}$ is the $\Pi_{di}\times\Pi_{dj}$ submatrix of W, defined for $i,j\in\Pi_h$.

<u>Lemma</u> <u>10</u>.

$$W_{ij} = \begin{cases} 0 & \text{if } i'\neq j' \\ -(\sigma^2_{i'} - h_{i'}) & \text{if } i'=j',\ i\neq j \\ \sigma^2_i - \sigma^2_{i'} & \text{if } i'=j',\ i=j \end{cases}$$

Equivalently,

$$W_{(ij)} = \begin{cases} 0 & i\neq j \\ \mathrm{diag}_{k\in\Pi_{di}}(\sigma^2_k - h_i) - (\sigma^2_i - h_i)\mathbf{1}_{si}\mathbf{1}^T_{si} & i=j \end{cases}$$

<u>Proof</u>. It is convenient to assume without loss of generality that if i′ and j′ are ancestor and descendant, then it is i′ that is the ancestor. Formally, if $\alpha(i',j')=j'$, then $i'=j'$. The proof considers in turn five distinct and mutually exhaustive cases. By definition, $W_{ij}=L_i VL^T_j$. Expanding $L_i VL^T_j$ using Lemma 9 yields:

Case 1, $\alpha(i',j')\neq i'$: $L_i VL^T_j = h_{\alpha(i,j)} - h_{\alpha(i,j)} - h_{\alpha(i,j)} + h_{\alpha(i,j)} = 0$.

Case 2, $\alpha(i',j')=i'$, $i'\neq j'$, $\alpha(i,j')=i$: $L_i VL^T_j = \sigma^2_i - \sigma^2_{i'} - \sigma^2_i + \sigma^2_{i'} = 0$.

Case 3, $\alpha(i',j')=i'$, $i'\neq j'$, $\alpha(i,j')\neq i$: $L_i VL^T_j = h_{\alpha(i,j')} - \sigma^2_{i'} - h_{\alpha(i,j')} + \sigma^2_{i'} = 0$.

Case 4, $\alpha(i',j')=i'$, $i'=j'$, $i\neq j$: $L_i VL^T_j = h_{i'} - \sigma^2_{i'} - \sigma^2_{i'} + \sigma^2_{i'} = h_{i'} - \sigma^2_{i'}$.

Case 5, $\alpha(i',j')=i'$, $i'=j'$, $i=j$: $L_i VL^T_j = \sigma^2_i - \sigma^2_{i'} - \sigma^2_{i'} + \sigma^2_{i'} = \sigma^2_i - \sigma^2_{i'}$.

Cases 1, 2 and 3 all have $i'\neq j'$, and so show the first part of the proposition. Case 4 and Case 5 demonstrate the second and third parts respectively.

Definition of C. Let C be a $\Pi_s\times\Pi_s$ matrix defined by $C=\mathrm{diag}_{i\in\Pi_s}(\sigma_i^2 - h_{i'})$.

Definition of |. If A and B are two matrices with the same number of rows, then let A|B denote the matrix formed by juxtaposing the columns of A and B.

Definition of M^t , N^t , M^s , N^s . If A is a $\Pi_t\times n_A$ matrix of full rank, $n_A\leq n_t$, then let $M^t_A = A(A^TV^{-1}A)^{-1}A^TV^{-1}$, and let $N^t_A = \mathbf{I}_t - M^t_A$. If A is a $\Pi_s\times n_A$ matrix of full rank, $n_A\leq n_s$, then let $M^s_A = A(A^TC^{-1}A)^{-1}A^TC^{-1}$, and let $N^s_A = \mathbf{I}_s - M^s_A$. In each case, $rk(M_A)=rk(A)$. In each case, if A is a null matrix, then let $M_A=0$, and $N_A=I$.

Remark. The M's and N's are orthogonal projection matrices in the Ω space indicated by their superscript. M projects onto the columns of the subscripted matrix, while N projects onto the space orthogonal to them. Orthogonality in Ω_t is taken with respect to V^{-1}, and in Ω_s is taken with respect to C^{-1}. The principal properties of projection matrices, which will be used without comment, are that $M_{A|B}A=A$, and $N_{A|B}A=0$; that $M_AM_A=M_A$, $N_AN_A=N_A$ and $M_AN_A=0$; that if the columns of A and B span the same subspace, then $M_A=M_B$ and $N_A=N_B$; and that if the columns of A are orthogonal to the columns of B then $M_AB=0$ and $N_AB=B$.

Lemma 11. $KL=N^t_{\mathbf{1}_t}$

Proof. The ith row of KL is

$$([f_i^T - f_{i'}^T] + [f_{i'}^T - f_{i''}^T] + [f_{i''}^T - f_{i'''}^T] + \ldots + - f_r^T]),$$

where r represents the root node. This equals $[f_i^T - f_r^T]$. f_i^T , $i\in\Pi_t$, by definition contains a 1 in position i, and zeroes elsewhere. f_r^T is constant for all rows, and as $U_r=\mathbf{I}_t$, f_r^T equals $(\mathbf{1}_i^T V_i^{-1} \mathbf{1}_i)^{-1}\mathbf{1}_i^T V_i^{-1}$. Hence

$$KL = \mathbf{I}_t - \mathbf{1}_t(\mathbf{1}_t^T V^{-1}\mathbf{1}_t)^{-1}\mathbf{1}_t^T V^{-1},$$

which equals $N^t_{\mathbf{1}_t}$ as required.

Lemma 12. $L\mathbf{1}_t=0$

Proof. The ith row of L is $[f_i^T - f_{i'}^T]$, but calculation from the definition of f_i using $U_i\mathbf{1}_t=\mathbf{1}_i$ shows that $f_i^T\mathbf{1}_t=1$, for all $i\in\Pi$. Hence each element of $L\mathbf{1}_t$ equals 1-1=0, as required.

Definition of S. Let S be a $\Pi_s\times\Pi_h$ matrix defined by

$$S_{ij} = \begin{array}{ll} 1 & i\in\Pi_{dj} \\ 0 & i\notin\Pi_{dj} \end{array}$$

Equivalently, when considered as a $\Pi_h\times\Pi_h$ matrix according to the partition P of Π_s, S is diagonal with $S_{(i)i}=\mathbf{1}_{si}$, $i\in\Pi_h$.

Lemma 13. $W = C - S(S^TC^{-1}S)^{-1}S^T$.

Proof. C and S can be considered as $\Pi_h\times\Pi_h$ matrices according to the partition P of Π_s. The off-diagonal elements of both C and S all equal zero by definition. Hence it suffices to prove that

$$W_{(ii)} = C_{(ii)} - \mathbf{1}_{si}(\mathbf{1}_{si}^T C_{(ii)}^{-1}\mathbf{1}_{si})^{-1}\mathbf{1}_{si}^T, \quad i\in\Pi_h.$$

Consider first the off-diagonal elements. The off-diagonal elements of the left hand side are all equal to $h_i - \sigma_i^2$. The off-diagonal elements of the right hand side all equal

$$-(\mathbf{1}_{si}^T C_{(ii)}^{-1}\mathbf{1}_{si})^{-1} = -\left(\sum_{j\in\Pi_{ti}}(\sigma_j^2 - h_{j'})^{-1}\right)^{-1}$$

so the off-diagonal elements of the two sides are equal according to Lemma 8.

Consider next the diagonal elements. The diagonal element of the left hand side indexed by j, $j \in \Pi_{di}$, is $\sigma_j^2 - \sigma_i^2$ by Lemma 10. The corresponding diagonal element of the right hand side is C_{jj} plus the off-diagonal element. But as these are equal in the right and left hand sides, we can write

$$c_j = \sigma_i^2 - h_{i'} - (h_i - \sigma_i^2) = \sigma_j^2 - \sigma_i^2$$

which proves the result.

<u>Lemma</u> <u>14</u>. $S^T C^{-1} L = 0$

<u>Proof</u>. Let S^T be partitioned into a row of submatrices s_i, $i \in \Pi_h$, in which s_i contains (as columns) the rows of S indexed by Π_{di}. Let L be partitioned into a column of submatrices m_i, $i \in \Pi_h$, in which m_i contains the rows of L indexed by Π_{di}. We can now write

$$S^T C^{-1} L = \sum_{i \in \Pi_h} s_i C^{-1}_{(ii)} m_i$$

We proceed by showing that each element of the sum equals zero. The element indexed by i looks like this:

$$\begin{pmatrix} 00000\ldots \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ 00000\ldots \\ 11111\ldots \\ 00000\ldots \\ \ldots\ldots\ldots \end{pmatrix} \begin{pmatrix} C^{-1}_{j_1 j_1}(f^T_{j_1} - f^T_i) \\ C^{-1}_{j_2 j_2}(f^T_{j_2} - f^T_i) \\ C^{-1}_{j_3 j_3}(f^T_{j_3} - f^T_i) \\ C^{-1}_{j_4 j_4}(f^T_{j_4} - f^T_i) \\ \ldots\ldots\ldots \end{pmatrix}$$

where j_1, j_2 etc represent the elements of Π_{di}. The rows of the matrix product corresponding to the zero rows in the left hand factor will be zero. This leaves only the row indexed by i, which is the row of ones. The product of this row with the right hand factor is

$$\sum_{j\in\Pi_{di}} C_{jj}^{-1}(f_j-f_i)^T\ , \qquad \text{(A4)}$$

and the lemma will be proved if this can be shown to equal zero for all $i\in\Pi_h$. Expanding the f_i from its definition, and using the definition of C, we obtain that (A4) equals

$$\sum_{j\in\Pi_{di}} \frac{1}{\sigma_j^2-h_i}(\sigma_j^2\mathbf{1}_j^T V_j^{-1}U_j - \sigma_i^2\mathbf{1}_i^T V_i^{-1}U_i)$$

By Lemma 8,

$$\sum_{j\in\Pi_{di}} \frac{1}{\sigma_j^2-h_i} = \frac{1}{\sigma_i^2-h_i},$$

so the lemma will be proved if we can prove the equality

$$\sum_{j\in\Pi_{di}} \frac{\sigma_j^2}{\sigma_j^2-h_i}\mathbf{1}_j^T V_j^{-1}U_j = \frac{\sigma_i^2}{\sigma_i^2-h_i}\mathbf{1}_i^T V_i^{-1} U_i \qquad \text{(A5)}$$

which we now proceed to do.

V_i in the form $(V_i - h_i\,\mathbf{1}_i\,\mathbf{1}_i^T) + h_i\,\mathbf{1}_i\,\mathbf{1}_i^T$ can be inverted by Lemma 1, pre-multiplied by $\mathbf{1}_i$, and re-arranged to yield

$$\mathbf{1}_i^T V_i^{-1} = \frac{1}{1 + h_i\mathbf{1}_i^T(V_i - h_i\mathbf{1}_i\mathbf{1}_i^T)^{-1}\mathbf{1}_i}\ \mathbf{1}_i^T(V_i - h_i\mathbf{1}_i\mathbf{1}_i^T)^{-1} \qquad \text{(A6)}$$

The scalar factor on the RHS of (A6) equals, by the first part of Lemma 8,

$$\frac{1}{1 + h_i \sum_{j\in\Pi_{di}} (\sigma_j^2 - h_i)^{-1}}$$

which using the second part of Lemma 8 becomes

$$\frac{\sigma_i^2 - h_i}{\sigma_i^2}$$

Using

$$(V_i \ - h_i \mathbf{1}_i \mathbf{1}_i^T) = \mathrm{diag}_{j\in\Pi_{di}}(V_j\text{-}h_i\mathbf{1}_j\mathbf{1}_j^T)$$

from Lemma 6, the matrix inverse on the RHS of (A6) can be expressed using Lemma 1 as

$$\mathrm{diag}_{j\in\Pi_{di}}(V_j^{-1} + \frac{h_i}{1 - h_i\mathbf{1}_j^T V_j^{-1} \mathbf{1}_j} V_j^{-1}\mathbf{1}_j\mathbf{1}_j^T V_j^{-1})$$

Hence the RHS as a whole equals

$$\frac{\sigma_i^2 - h_i}{\sigma_i^2} \ \mathrm{row}_{j\in\Pi_{di}}(\frac{\sigma_i^2}{\sigma_j^2 - h_i} \mathbf{1}_j^T V_j^{-1})$$

We can therefore conclude from (A6) that

$$\frac{\sigma_i^2}{\sigma_i^2 - h_i} \ \mathbf{1}_i^T \ V_i^{-1} \ = \mathrm{row}_{j\in\Pi_{di}}(\frac{\sigma_i^2}{\sigma_j^2 - h_i} \mathbf{1}_j^T V_j^{-1})$$

This equality differs only notationally from the equality (A5) we had to prove, and so completes the proof.

Lemma 15. If A is a $\Pi_t\times n_A$ matrix of full rank, B is a $\Pi_t\times n_B$ matrix of full rank, and A and B are linearly independent, then

$$N^t_{A|B} = N^t_{(N^t_A B)} N^t_A$$

Proof. The lemma states that residualizing on (A|B) is equivalent to residualizing on A, and then on B residualized on A. Let D be a matrix of full rank such that (D|A|B) spans Ω_t and such that $D^TV^{-1}(A|B)=0$. The existence of D is guaranteed by Lemma 3. Then there is a unique decomposition of a vector $x\in\Pi_t$ into $x=Dd+(A|B)\begin{pmatrix}a\\b\end{pmatrix}$ The LHS of the statement of the lemma pre-multiplies x into Dd. $N^t_A\, x=Dd+(0|N^t_A\, B)\begin{pmatrix}a\\b\end{pmatrix}$, and the second term is annihilated on pre-multiplication by $N^t_{(N^t_A B)}$ because $N^t_{(N^t_A B)}$ $N^t_A\, B=0$. To prove the lemma, it remains to show that $N^t_{(N^t_A B)}\, D=D$. By properties of projection matrices, this will be true if $N^t_A\, B$ and D are orthogonal. But

$$B^TN^{t\;T}_A V^{-1}D = B^T(\mathbf{I}_t - V^{-1}A(A^TV^{-1}A)^{-1}A^T)V^{-1}D,$$

and as $B^TV^{-1}D=0$ and $A^TV^{-1}D=0$, the whole expression equals zero as required. Hence D and $N^t_A\, B$ are orthogonal and the lemma is proved.

Lemma 16. If A is a $\Pi_t\times n_A$ matrix of full rank, and if $\mathbf{1}_t$ and A are linearly independent, then

$$N^{t\;\;T}_{\mathbf{1}_t|A} V^{-1}N^t_{\mathbf{1}_t|A} \quad = \quad L^T\, N^{s\;\;T}_{LA|S}\, C^{-1}N^s_{LA|S}\, L$$

Proof. Expanding the projection matrices, the RHS becomes

$$L^TC^{-1}L - L^TC^{-1}(LA|S)\begin{pmatrix} A^TL^TC^{-1}LA & A^TL^TC^{-1}S \\ S^TC^{-1}LA & S^TC^{-1}S \end{pmatrix}^{-1}(LA|S)^{\,T}C^{-1}L$$

By Lemma 14, $S^TC^{-1}L=0$, and so this simplifies to

$$L^TC^{-1}L - (L^TC^{-1}LA|0)\begin{pmatrix} A^TL^TC^{-1}LA & 0 \\ 0 & S^TC^{-1}S \end{pmatrix}^{-1}(L^TC^{-1}LA|0)^{\,T}$$

and so to

$$L^TC^{-1}L - L^TC^{-1}LA(A^TL^TC^{-1}LA)^{-1}A^TL^TC^{-1}L$$

By Lemma 15, we have $N^t_{\mathbf{1}_t|A} = N^t_{(N^t_{\mathbf{1}_t}A)}\; N^t_{\mathbf{1}_t} = (\mathbf{I}_t - M^t_{(N^t_{\mathbf{1}_t}A)})N^t_{\mathbf{1}_t}$, which is defined because the condition of the theorem that $\mathbf{1}_t$ and A are linearly independent ensures

that $N^{t}_{\mathbf{1}_t}$ A is of full rank. We can use this to expand the LHS as follows
$N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t} - N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t}A(A^{T}N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t}A)^{-1}A^{T}N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t}$
To show that the RHS and LHS are equal, it will therefore suffice to show that $L^{T}C^{-1}L=N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t}$. Lemma 13 implies that $WC^{-1}W=W$, and W by definition equals LVL^{T}. Hence

$$LVL^{T}C^{-1}LVL^{T} = LVL^{T}$$

Pre-multiplying by K and postmultiplying by K^{T}, and applying Lemma 11 we obtain

$$N^{t}_{\mathbf{1}_t}VL^{T}C^{-1}LVN^{t}_{\mathbf{1}_t}{}^{T} = N^{t}_{\mathbf{1}_t}VN^{t}_{\mathbf{1}_t}{}^{T}$$

Substituting $(\mathbf{I}_t-M^{t}_{\mathbf{1}_t})$ for $N^{t}_{\mathbf{1}_t}$ in the LHS, and using $L\mathbf{1}_t=0$ from Lemma 12 yields

$$VL^{T}C^{-1}LV = N^{t}_{\mathbf{1}_t}VN^{t}_{\mathbf{1}_t}{}^{T}$$

Calculation shows that $V^{-1}N^{t}_{\mathbf{1}_t}V=N^{t}_{\mathbf{1}_t}{}^{T}$, so pre- and post-multiplication by V^{-1} gives

$$L^{T}C^{-1}L = N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t}$$

as required. This completes the proof.

Remark. The following theorem says that the long regression, which has C as its variance-covariance matrix, is equivalent to the standard regression. C is a diagonal matrix. This allows GLIM, for example, to handle the standard regression, even though it does not allow covariances among the errors. The LHS of the statement of the theorem is the residual sum of squares in the standard regression, concentrated for the parameter vectors for $\mathbf{1}_t$ and X, as a function of y, Z and γ. The RHS is the residual sum of squares in the long regression, concentrated for the parameter vectors of S and LX, as a function of y, Z and γ. Before the theorem we formally define the data of the analysis. Note that the null hypothesis is implicit in the definition of y.

Definition of ε, y, μ, X, β, Z, n_X, n_Z. Let ε be a random variable over Ω_t, distributed as $\mathbf{N}(0,V)$. Let μ be a scalar, X a $\Pi_t\times n_X$ matrix of full rank which is linearly independent of $\mathbf{1}_t$, β an $n_X\times1$ vector, and Z a $\Pi_t\times n_Z$ matrix. Let y be a random variable defined by $y=(\mathbf{1}_t|X)\begin{pmatrix}\mu\\ \hline \beta\end{pmatrix}+\varepsilon$.

Theorem 1. If γ is an $n_Z\times1$ vector, then

$$(y-Z\gamma)^T {N^{t}_{\mathbf{1}_t|X}}^T V^{-1}N^{t}_{\mathbf{1}_t|X}(y-Z\gamma) = (Ly-LZ\gamma)^T {N^{s}_{LX|S}}^T C^{-1}N^{s}_{LX|S}(Ly-LZ\gamma)$$

Proof. $N^{s}_{LX|S}$ is well defined only if (LX|S) is of full rank, and this is established first. S is of full rank by construction, and by Lemma 14 $S^TC^{-1}L=0$ so to show (LX|S) of full rank it remains to show that LX is of full rank. Suppose not, then there exists a vector a≠0 such that LXa=0 and so by Lemma 11 we have $KLXa=N^{t}_{\mathbf{1}_t} Xa=0$. As X is of full rank, if a≠0 then Xa≠0. But only multiples of $\mathbf{1}_t$ are annihilated by $N^{t}_{\mathbf{1}_t}$, hence Xa is a multiple of $\mathbf{1}_t$. However, this contradicts the definition of X, which states that $\mathbf{1}_t$ and X are linearly independent.

The statement of the theorem will be true if

$${N^{t}_{\mathbf{1}_t|X}}^T V^{-1}N^{t}_{\mathbf{1}_t|X} \quad = \quad L^T {N^{s}_{LX|S}}^T C^{-1}N^{s}_{LX|S}\, L$$

but this is the statement of Lemma 16, with X in the place of A. This completes the proof.

Definition of e, Π_g, n_g, Ω_g, $\mathbf{I}_g$, τ, λ. Let e be a random variable over Ω_s defined by

$$e = N^{s}_{S|LX}\, L\varepsilon$$

Let $\Pi_g=\{i|i\in\Pi_h, e_{(i)}\neq0\}$. Let n_g be the number of elements of Π_g. Let Ω_g be the set of column vectors with real elements indexed by Π_g. Let $\mathbf{I}_g$ be the identity matrix over Ω_g. Let $j_i=\min\{j|j\in\Pi_{di}, \tau_j\neq0\}$, $i\in\Pi_g$. Let τ be a random variable over Ω_s defined by

$$\tau_{(i)} \propto e_{(i)}\,, \quad \tau_{(i)}^{T}\, C_{(ii)}^{-1}\, \tau_{(i)} = 1, \quad \tau_{j_i} > 0 \qquad\qquad i \in \Pi_g$$

$$\tau_{(i)} = 0 \qquad\qquad i \notin \Pi_g$$

For $i \in \Pi_g$, the conditions define in turn the relative values of the elements of $\tau_{(i)}$, the magnitude of $\tau_{(i)}$ and the sign of $\tau_{(i)}$. Let λ be a random variable over Ω_g defined by $e_{(i)}=\lambda_i\tau_{(i)}$, $i \in \Pi_g$.

Remark. It is formally possible that $\Pi_g=\{\}$, if all of the variability in y has been explained by X. In what follows I shall tacitly assume that this is not the case. In practical terms, this situation will be obvious because of a zero sum of squares in the standard regression, and in theoretical terms it has no particular interest. There is no possibility of discovering if Z explains variability in y from such a dataset.

Definition of M^g , N^g . If A is a $\Pi_g \times n_A$ matrix of full rank, $n_A \leq n_g$, then let M_A^g = $A(A^TA)^{-1}A^T$, and let $N_A^g = \mathbf{I}_g - M_A^g$. Note that $rk(M_A)=rk(A)$. In the case that A is a null matrix, let $M_A^g=0$, and $N_A^g=\mathbf{I}_g$. M^g and N^g are orthogonal projection matrices over Ω_g, and orthogonality is taken with respect to I_g.

Definition of G. Let G be a $\Pi_g \times \Pi_s$ matrix defined by

$$G_{i(j)} = \begin{cases} \tau_{(i)}^{T} & i=j \\ 0 & i \neq j \end{cases}$$

Remark. e is the residual in the long regression after regression of y on X. Π_g is the set of higher nodes at which these residuals are not identically zero. The circumstances in which some of the residuals are identically zero is discussed in §3(e). Usually, $\Pi_g=\Pi_h$. Ω_g is the dataspace of the short regression. τ is a vector containing the “pattern" of the residuals, and λ contains the “magnitudes" in the sense of §3(c). G is a matrix which in combination with C will form the linear contrasts GC^{-1} which transform the long regression into the short regression. G^TGC^{-1} is a projection matrix, as the following lemma shows. The short regression can therefore be seen as a projection of the long regression onto the columns of G^T. The ith column of G^T has zero everywhere except in the radiation of node i, and there it is proportional to $e_{(i)}$, the residuals in the long regression of y on X. This projection ensures that all the residuals after regression on $GC^{-1}LZ$ must lie in the same space, and so must be proportional, within each radiation, to $e_{(i)}$.

<u>Lemma</u> <u>17</u>. $GC^{-1}G^T = \mathbf{I}_g$ and $GC^{-1}S = 0$

<u>Proof</u>. The first part is obvious in view of the diagonality of C and the definition of G. C and S are diagonal matrices according to the partition P of Π_s. From the definition of G,

$$\begin{array}{ll} G_{i(j)} \; = \; 0 & \qquad\qquad i\neq j \\ (GC^{-1}S)_{ii} \propto e^{T}_{(i)}\, C^{-1}S_{(i)i} & \qquad\qquad i\in\Pi_g \end{array}$$

Hence if $e^TC^{-1}S=0$, then so does $GC^{-1}S$. But from the definition of e, $e^TC^{-1}S$ equals $(Ly)^TN^{s}_{S|LX}{}^{T}C^{-1}S = (Ly)^TC^{-1}N^{s}_{S|LX}\,S$. However, $N^{s}_{S|LX}\,S = 0$ and so the second part of the lemma is proved.

<u>Definition</u> <u>of</u> X^g. Let X^g be a $\Pi_g\times rk(GC^{-1}LX)$ matrix defined such that the columns of X^g span the same subspace as those of $GC^{-1}LX$. X^g may be a null matrix.

<u>Remark</u>. This definition is needed in case $GC^{-1}LX$ is not of full rank even though LX is. See §3(e). $X^g=GC^{-1}LX$ will satisfy the definition when $GC^{-1}LX$ is of full rank.

<u>Lemma</u> <u>18</u>. If A is a $\Pi_t\times n_A$ matrix of full rank, and $\mathbf{1}_t$ and A are linearly independent, then $LN^{t}_{\mathbf{1}_t|A} = N^{s}_{S|LA}\,L$.

<u>Proof</u>. The lemma will be true if

$$LM^{t}_{\mathbf{1}_t|A} = M^{s}_{S|LA}\,L$$

which we now prove. ($\mathbf{1}_t$|A) and ($\mathbf{1}_t|N^{t}_{\mathbf{1}_t}$ A) span the same subspace of Ω_t, so it follows that $M^{t}_{\mathbf{1}_t|A} = M^{t}_{\mathbf{1}_t|N^{t}_{\mathbf{1}_t}A}$. The LHS therefore equals

$$L(\mathbf{1}_t|N^{t}_{\mathbf{1}_t}\ A)\begin{pmatrix}\mathbf{1}_t^{T}V^{-1}\mathbf{1}_t & 0\\ 0 & A^{T}N^{t}_{\mathbf{1}_t}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t}A\end{pmatrix}^{-1}(\mathbf{1}_t|N^{t}_{\mathbf{1}_t}A)\ {}^{T}V^{-1}$$

which using L$\mathbf{1}_t$=0 from Lemma 12, its consequence that $LN^{t}_{\mathbf{1}_t}$ =L, and the block diagonality of the inverse of the partitioned matrix, reduces to

$$LA(A^{T}N^{t}_{\mathbf{1}_t}\ {}^{T}V^{-1}N^{t}_{\mathbf{1}_t}\ A)^{-1}(N^{t}_{\mathbf{1}_t}\ A)^{T}V^{-1} \qquad \text{(A7)}$$

Because (S|LA) and (S|N^{s}_{S} LA) span the same subspace of Ω_s, $M^{s}_{S|LA} = M^{s}_{S|N^{s}_{S}LA}$, and so the RHS equals

$$(S|N^{s}_{S}\ LA)\begin{pmatrix}S^{T}C^{-1}S & 0\\ 0 & A^{T}L^{T}N^{s}_{S}{}^{T}C^{-1}N^{s}_{S}LA\end{pmatrix}^{-1}(S|N^{s}_{S}LA)\ {}^{T}C^{-1}L$$

Using $S^{T}C^{-1}L$=0 from Lemma 14 , its consequence that N^{s}_{S} L=L, and the block diagonality of the inverse of the partitioned matrix, this equals

$$LA(A^{T}L^{T}C^{-1}LA)^{-1}(LA)^{T}C^{-1}L \qquad \text{(A8)}$$

It will now be shown piecewise that (A7) equals (A8) thus proving the lemma. Both formulae begin with LA. The matrix inverses that follow are equal because $L^{T}C^{-1}L=N^{t}_{\mathbf{1}_t}\ {}^{T}V^{-1}N^{t}_{\mathbf{1}_t}$ by Lemma 16, and the remaining portions are equal for the same reason and because $N^{t}_{\mathbf{1}_t}\ {}^{T}V^{-1}N^{t}_{\mathbf{1}_t} = N^{t}_{\mathbf{1}_t}\ {}^{T}V^{-1}$. This completes the proof.

<u>Lemma</u> <u>19</u>. Conditional on τ, λ ~ **N**(0,N^{g}_{Xg}).

<u>Proof</u>. First it is established that the support of the probability distribution of e is the subspace of Ω_s orthogonal to (S|LX), and that there the density is proportional to

$$\exp(-\tfrac{1}{2}\ e^{T}C^{-1}e)$$

Let $\eta = N^{t}_{\mathbf{1}_t|X}\,\varepsilon$, and $\varphi = M^{t}_{\mathbf{1}_t|X}\,\varepsilon$. As the cross-product $M^{t}_{\mathbf{1}_t|X}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t|X}$ equals zero, the density of ε is proportional to

$$\exp(-\tfrac{1}{2}\,\varepsilon^{T}V^{-1}\varepsilon) \;=\; \exp(-\tfrac{1}{2}\,\varepsilon^{T}N^{t}_{\mathbf{1}_t|X}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t|X}\,\varepsilon - \tfrac{1}{2}\,\varepsilon^{T}M^{t}_{\mathbf{1}_t|X}{}^{T}V^{-1}M^{t}_{\mathbf{1}_t|X}\,\varepsilon)$$

which in turn yields

$$= \exp(-\tfrac{1}{2}\,\eta^{T}V^{-1}\eta - \tfrac{1}{2}\,\varphi^{T}V^{-1}\varphi)$$

Hence the density of η is zero on the subspace orthogonal to $(\mathbf{1}_t|X)$, where it is proportional to

$$\exp(-\tfrac{1}{2}\,\eta^{T}V^{-1}\eta)\;\int \exp(-\tfrac{1}{2}\,\varphi^{T}V^{-1}\varphi)\,d\varphi$$

and so it is also proportional simply to

$$\exp(-\tfrac{1}{2}\,\eta^{T}V^{-1}\eta)$$

$e=L\eta$, so the density of e will be as claimed if

$$\eta^{T}V^{-1}\eta = e^{T}C^{-1}e \quad \text{where } N^{t}_{\mathbf{1}_t|X}\,\eta=\eta \text{ and } e=L\eta$$

which we now prove. Over the relevant subspace, the LHS is equal to

$$\eta^{T}N^{t}_{\mathbf{1}_t|X}{}^{T}V^{-1}N^{t}_{\mathbf{1}_t|X}\,\eta$$

and the RHS equals

$$\eta^{T}N^{t}_{\mathbf{1}_t|X}{}^{T}L^{T}C^{-1}LN^{t}_{\mathbf{1}_t|X}\,\eta$$

By Lemma 18, $LN^{t}_{\mathbf{1}_t|X} = N^{s}_{S|LX}\,L$, so the RHS equals

$$\eta^{T}L^{T}N^{s}_{S|LX}\,C^{-1}N^{s}_{S|LX}\,L\eta$$

But Lemma 16 shows that

$$N^{t}_{\mathbf{1}_t|X}{}^{T} V^{-1} N^{t}_{\mathbf{1}_t|X} = L^T N^{s}_{S|LX} C^{-1} N^{s}_{S|LX} L$$

and so the two sides are equal, completing the proof of the distribution of e.

We now proceed to find the distribution of λ. From the definition of τ and λ, we have

$$e^{T}_{(i)} C^{-1}_{(ii)} e_{(i)} = \lambda^{2}_{i} \tau^{T}_{(i)} C^{-1}_{(ii)} \tau_{(i)} = \lambda^{2}_{i} \quad i \in \Pi_g$$

hence

$$e^T C^{-1} e = \lambda^T \lambda, \text{ and so } \exp(-\tfrac{1}{2} e^T C^{-1} e) = \exp(-\tfrac{1}{2} \lambda^T \lambda)$$

To derive the distribution of λ, it remains to establish over what subset of Ω_g the density is non-zero, and what the Jacobian of the transformation is. The subset of Ω_g is those values of λ corresponding to an e belonging to the subspace of Ω_s defined by $M^{s}_{S|LX}$ e=0. As $e_{(i)}=\lambda_i\tau_{(i)}$ this condition becomes

$$(S|LX)^T C^{-1} \operatorname{col}_{i\in\Pi_g} \{\lambda_i \tau_{(i)}\} = 0$$

But $S^T C^{-1} G^T=0$ by Lemma 17, so $(S^T C^{-1})^{T}_{(i)} \tau_{(i)}=0$ and therefore S may be dropped. Writing $(LX)_{(i)}$ for the submatrix of LX containing only those rows indexed by Π_{di}, this condition is equivalent to

$$\sum_{i\in\Pi_h} \lambda_i (LX)^{T}_{(i)} C^{-1}_{(ii)} \tau_{(i)} = 0, \text{ or } (GC^{-1}LX)^T \lambda = 0, \text{ hence to } (X^g)^T \lambda = 0$$

The subset of Ω_g over which the density of λ is defined conditional on τ is therefore a subspace defined by $(X^g)^T\lambda=0$. Conditional on τ, $e_{(i)}=\lambda_{(i)}\tau_{(i)}$ implies that $de \propto d\lambda$, so there is no non-constant term in the Jacobian. We have now established that the density of λ is proportional to

$$\exp(-\tfrac{1}{2} \lambda^T \lambda)$$

where $M^g_{Xg}\,\lambda=0$. A random variable with distribution $\mathbf{N}(0,N^g_{Xg})$ has density proportional to $\exp(-\frac{1}{2}\,\xi^T\xi)$ where $M^g_{Xg}\,\xi=0$, and zero elsewhere. But this is the same density as λ, and so λ too is distributed as $\mathbf{N}(0,N^g_{Xg})$. This completes the proof.

Remark. The following theorem shows that the short regression is analytically valid in the case where the working phylogeny is the true phylogeny, and ρ is known and taken as fixed at its true value. It does this by giving the probability density of $N^g_{Xg}\,GC^{-1}Ly$, the residual vector after regression of $GC^{-1}Ly$ on $GC^{-1}LX$, based on the whole process of computing the long regression, conditioning on τ, and using the random linear contrasts GC^{-1} to form the short regression. The theorem shows that conditional on τ, the probability density is the same as it would have been if $GC^{-1}LX$ were taken as fixed, and the residual's density calculated on the basis of an error ψ distributed as $\mathbf{N}(0,\mathbf{I}_g)$ in the regression $y^g=GC^{-1}LX\beta+\psi$. This equivalence of the residual density in the two cases establishes the exactness of the short regression for testing for the addition of $GC^{-1}LZ$. Note that conditional on τ, $GC^{-1}LZ$ is fixed and not random.

Theorem 2. Conditional on τ

$$N^g_{Xg}\,GC^{-1}Ly \sim \mathbf{N}(0,N^g_{Xg})$$

Proof. We show that $N^g_{Xg}\,\lambda=\lambda$, and then the theorem follows immediately from Lemma 19. We begin with the identity

$$Ly = L(\,(\mathbf{1}_t|X)\begin{pmatrix}\mu\\ \hline \beta\end{pmatrix} + \varepsilon\,) = M^s_{S|LX}\,L(\,(\mathbf{1}_t|X)\begin{pmatrix}\mu\\ \hline \beta\end{pmatrix} + \varepsilon\,) + N^s_{S|LX}\,L(\,(\mathbf{1}_t|X)\begin{pmatrix}\mu\\ \hline \beta\end{pmatrix} + \varepsilon\,)$$

$L\mathbf{1}_t=0$ from Lemma 12. We also know that $N^s_{S|LX}\,L\varepsilon=e$ by definition, and $N^s_{S|LX}$ annihilates LX while $M^s_{S|LX}$ preserves it. Hence these expressions also equal

$$M^s_{S|LX}\,L(X\beta+\varepsilon) + e$$

Pre-multiplying by GC^{-1} yields

$$GC^{-1}Ly = GC^{-1}M^{s}_{S|LX}\ L(X\beta+\varepsilon) + GC^{-1}e$$

$N^{g}_{X^g}\ GC^{-1}(S|LX)=0$ because $GC^{-1}S=0$ by Lemma 17, and because the columns of X^g and the columns of $GC^{-1}LX$ span the same subspace of Ω_g. Expansion of the projection matrix therefore shows that $N^{g}_{X^g}$ annihilates the first term of the RHS. Further, calculation shows that $GC^{-1}e=\lambda$. Hence pre-multiplying by $N^{g}_{X^g}$ gives

$$N^{g}_{X^g}\ GC^{-1}Ly = N^{g}_{X^g}\ \lambda$$

But $N^{g}_{X^g}\ \lambda=\lambda$ because by Lemma 19 the density of λ is zero except over the subspace where this is true. This completes the proof.

Definition of T, n_T. Let $n_T=n_s-n_h-n_g$. Let T be a $\Pi_s\times n_T$ matrix of full rank which satisfies $T^TC^{-1}(S|G^T)=0$, and $rk(T|S|G^T)=n_s$. If $n_s-n_h-n_g=0$, then T will be a null matrix.

Remark. T will be null only when the working phylogeny is binary and as a consequence the phylogenetic regression and the standard regression are the same.

Remark. The following theorem shows that the long regression with T is equivalent to the short regression. The LHS is the sum of squares of the long regression, concentrated for the parameter vectors associated with LX, S and T. The RHS is the residual sum of squares of the short regression, concentrated for the parameter vectors associated with $GC^{-1}LX$. The introduction of A allows for possible collinearity between T and LX, or in other words that $rk(X^g)<rk(LX)$.

Theorem 3. Let A be a $\Pi_s\times rk(S|LX|T)$ matrix of full rank whose columns span the same subspace of Ω_s as the columns of (S|LX|T). Then conditional on τ,

$$(Ly-LZ\gamma)^T\ N^{s\ T}_{A}\ C^{-1}N^{s}_{A}(Ly-LZ\gamma)$$

$$= (GC^{-1}Ly-GC^{-1}LZ\gamma)^{T}\, N_{X^g}^{g\,T}\, \mathbf{I}_g^{-1}\, N_{X^g}^{g}(GC^{-1}Ly-GC^{-1}LZ\gamma)$$

<u>Proof</u>. The statement of the theorem will be true if

$$N_A^{s\,T}\, C^{-1}N_A^{s} = C^{-1}G^{T}\, N_{X^g}^{g\,T}\, \mathbf{I}_g^{-1}\, N_{X^g}^{g}\, GC^{-1}$$

which by definition of the projection matrices simplifies to

$$C^{-1}N_A^{s} = C^{-1}G^{T}\, N_{X^g}^{g}\, GC^{-1}$$

which we now prove. Let n_R=n_s-rk(S|T|LX)=n_s-rk(A). By Lemma 3, there exists a $\Pi_s \times n_R$ matrix R of full rank, which satisfies rk(A|R)=rk(S|T|LX|R)=n_s, and $A^TC^{-1}R=0$ so that $(S|T|LX)^TC^{-1}R=0$. As $rk(S|T|G^T)=n_s$, and $G^TC^{-1}(S|T)=0$ from Lemma 17 and construction of T, by Lemma 4 there exists a $\Pi_g \times n_R$ matrix Q of full rank such that $R=G^TQ$.

By Lemma 2 the LHS equals $C^{-1}M_R^{s}$, which can in turn be written as

$$C^{-1}R(R^TC^{-1}R)^{-1}R^TC^{-1}$$

Substituting G^TQ for R, and as $GC^{-1}G^T=\mathbf{I}_g$ by Lemma 17, this equals

$$C^{-1}G^TQ(Q^TGC^{-1}G^TQ)^{-1}Q^TGC^{-1}$$

$$= C^{-1}G^TQ(Q^TQ)^{-1}Q^TGC^{-1}$$

$$= C^{-1}G^TM_Q^{g}\, GC^{-1}$$

To prove the theorem, it therefore remains to show that

$$C^{-1}G^TM_Q^{g}\, GC^{-1} = C^{-1}G^{T}\, N_{X^g}^{g}\, GC^{-1}$$

By Lemma 2, $M_Q^{g} = N_{X^g}^{g}$ and so this equality will hold, only provided $rk(X^g|Q)=n_g$, and $Q^TX^g=0$. $Q^T(GC^{-1}LX)=R^TC^{-1}LX$, which equals zero by the definition of R, and because the columns of X^g and those of $GC^{-1}LX$ span the same subspace of Ω_g,

$Q^TX^g=0$ too. Now GC^{-1} is a $\Pi_g\times\Pi_s$ matrix of full rank, and by construction of R, $rk(S|T|LX|R)=n_s$. Hence

$$rk(GC^{-1}(S|T|LX|R))=rk(GC^{-1}S|GC^{-1}T|GC^{-1}LX|GC^{-1}R)=n_g$$

As $GC^{-1}S=0$ by Lemma 17 and $GC^{-1}T=0$ by construction of T, and $GC^{-1}R=GC^{-1}G^TQ=Q$ by Lemma 17, it follows that

$$rk(GC^{-1}LX|Q)=n_g,$$

and so $rk(X^g|Q)=n_g$ too, as required. This completes the proof of the theorem.

<u>Definition</u> <u>of</u> F. If a is a $\Pi_g\times1$ vector, B is an $\Pi_g\times n_B$ matrix of full rank, $n_B<n_g$, D is a $\Pi_g\times n_D$ matrix of full rank, $n_B+n_D<n_g$, $rk(B|D)=rk(B)+rk(D)$, and $a^TN^g_{B|D}\,a\neq0$, then let F(a,B,D) equal

$$\frac{a^TM^g_{N_BD}a}{a^TN^g_{B|D}a} = \frac{a^TM^g_{N_BD}a}{a^TN^g_B\,a - a^TM^g_{N_BD}a}$$

<u>Definition</u> <u>of</u> Z^c. Let Z^c be a matrix of full rank formed by deleting columns from Z in such a way that $(X^g|GC^{-1}LZ^c)$ is of full rank and that $rk(X^g|GC^{-1}LZ^c)=rk(X^g|GC^{-1}LZ)$. Z^c may be a null matrix.

<u>Definition</u> <u>of</u> m_X, m_Z. Let $m_X=rk(X^g)$, and $m_Z=rk(Z^c)$.

<u>Definition</u> <u>of</u> Γ. Let Γ be defined almost surely as a $\Pi_g\times m_Z$ random matrix whose elements are independently distributed in normal distributions, with zero mean, and

$$\mathrm{Var}(\Gamma_{ij}) = \frac{e^T_{(i)}C^{-1}_{(i)}e_{(i)}}{(e^T_{(i)}C^{-1}_{(i)}(LZ^c)_{(i)j})^2}$$

This definition fails when the denominator is zero for any i,j.

<u>Definition</u> <u>of</u> Ψ. Let Ψ be a $\Pi_s\times m_Z$ random matrix defined almost surely by

$\Psi_{(i)j}$ =	$\Gamma_{ij}(LZ^c)_{(i)j}$	$i\in\Pi_g$, $j=1..m_Z$
	0	$i\in\Pi_h-\Pi_g$, $j=1..m_Z$

Remark. Ψ is the random alternative to Z of the randomization test described in §3(c).

Definition of Φ_1, Φ_2, Φ. Let Φ_1 be a random $\Pi_g\times n_Z$ matrix defined by $\Phi_1=GC^{-1}\Psi$, and let Φ_2 be a random $\Pi_g\times(n_g-m_X-m_Z)$ matrix whose elements have normal distributions with zero mean and unit variance, independent of each other and of Γ. Let $\Phi=(\Phi_1|\Phi_2)$.

Lemma 20. Φ_{ij} have independent normal distributions with zero mean and unit variance.

Proof. Independence and zero mean follow for Φ_1 because Φ_{ij} equals a constant times Γ_{ij}. Unit variance follows because that constant equals the inverse of the square root of the variance of Γ_{ij}. Φ_2 has the required distribution by definition.

Remark. The following theorem shows that the randomization test described in §3(c) is the same test as the short regression. Formally, the randomization test is to find the p-value for the null hypothesis that $\gamma=0$ by finding

$$\Pr\{F(GC^{-1}Ly,X^g,GC^{-1}\Psi)>F(GC^{-1}Ly,X^g,GC^{-1}LZ^c)\},$$

in which the substitution of Ψ for LZ^c is the only difference between the two sides of the inequality. If this probability is low, it implies that randomly selected explanatory variables would rarely explain as much of the remaining variation in Ly as LZ^c does. The short regression's test statistic would also have an F-distribution with m_Z and $n_g–m_X–m_Z$ degrees of freedom.

Theorem 4. Conditional on y, $F(GC^{-1}Ly,X^g,GC^{-1}\Psi)$ has an F-distribution with m_Z and $n_g–m_X–m_Z$ degrees of freedom

Proof. For brevity, let $a=GC^{-1}Ly$, $B=X^g$. By the definition of F and the relationship between the F-distribution and the Beta distribution, and noting that $GC^{-1}\Psi=\Phi_1$, it suffices to show that

$$q = \frac{a^T M^g_{N_B\Phi_1} a}{a^T N^g_B a}$$

has a Beta distribution with m_Z and $n_g-m_X-m_Z$ degrees of freedom. For a definition of the Beta distribution, and its relationship to the F distribution, see Abramowitz & Stegun (1965, 26.5.1, 26.5.2 and 26.5.28). We apply Gram-Schmidt orthogonalization (see for example Kingman and Taylor 1966, page 201) to the columns of $N^g_B(\Phi_1|\Phi_2)$ in order, yielding a $\Pi_g\times(n_g-m_X-m_Z)$ random matrix Λ conformally partitioned into $(\Lambda_1|\Lambda_2)$. There is still ambiguity in the definition of Λ, up to multiplication of each column by ±1. We complete the definition of Λ by requiring that the inner product of each pair of corresponding columns from $N^g_B(\Phi_1|\Phi_2)$ and Λ should be positive. Gram-Schmidt orthogonalization ensures that the columns of Λ are mutually orthogonal, and are of unit length. Further, that the first k columns of Λ, k=1,2, .. rk(Λ) span the same subspace as the first k columns of $N^g_B(\Phi_1|\Phi_2)$. By virtue of its construction Λ therefore satisfies

$$\Lambda^T\Lambda = I,\ \Lambda^T B = 0,$$

Because of symmetry between the columns of Φ, and the spherical symmetry of the normal distribution which Φ has by virtue of Lemma 20, the distribution of Λ satisfies

$$\mathrm{freq}(\Lambda) \propto \begin{array}{ll} 1 & \Lambda^T\Lambda=I,\ \Lambda^T B=0 \\ 0 & \text{o.w.} \end{array}$$

By construction, Λ_1 spans the same subspace as $N^g_B\,\Phi_1$, so there exists a random $m_Z\times m_Z$ matrix ς defined by $\Lambda_1 = N^g_B\,\Phi_1\,\varsigma$. Now

$$M^{g}_{N^{g}_{B}\Phi_1} = M^{g}_{\Lambda_1 \varsigma^{-1}} = \Lambda_1\, \varsigma^{-1}((\varsigma^{-1})^{T}\Lambda_1^{T}\, \Lambda_1\, \varsigma^{-1})^{-1}(\varsigma^{-1})^{T}\Lambda_1^{T} = \Lambda_1\, \Lambda_1^{T}$$

and so

$$q = \frac{a^{T}\Lambda_1\Lambda_1^{T}a}{a^{T}N^{g}_{B}a}$$

Let d be the random $(n_g\text{-}m_X)\times1$ vector defined by $d=\Lambda^{T}a$, and let $d_1=\Lambda_1^{T}$ a. The distribution of d in $\mathbf{R}^{n_g\text{-}m_X}$ is spherically symmetrical with fixed length $a^{T}N^{g}_{B}$ a. The volume element of d for which d_1 is a constant is the surface of an $(n_g–m_X–m_Z)$ sphere of radius $\sqrt{a^{T}N^{g}_{B}a - d^{T}d}$, and the volume element of vectors d_1 such that $d_1^{T}\, d_1$ equals a constant is the surface of an m_Z sphere of radius $\sqrt{d_1^{T}d_1}$. Hence the density of $d_1^{T}\, d_1$, which equals $a^{T}\Lambda_1\Lambda_1{}^{T}a$, is proportional to

$$\left(\sqrt{d_1^{T}d_1}\right)^{m_Z}\left(\sqrt{a^{T}N^{g}_{B}a\text{-}d_1^{T}d_1}\right)^{n_g\text{-}m_X\text{-}m_Z}$$

and so the density of q is proportional to

$$q^{m_Z/2}(1\text{-}q)^{(n_g\text{-}m_X\text{-}m_Z)/2}$$

which is the Beta distribution with m_Z and $n_g–m_X–m_Z$ degrees of freedom (Abramowitz & Stegun 1965, 26.5.1 and 26.5.2), as required.

References

Abramowitz, M. & Stegun, I.A. (ed.s) 1965 *Handbook of mathematical functions*. New York: Dover.

Kingman, J.F.C. & Taylor, S.J. 1966 *Introduction to measure and probability*. Cambridge: Cambridge University Press.