Stat 5102 Notes: Regression

Charles J. Geyer

April 27, 2007

In these notes we do not use the "upper case letter means random, lower case letter means nonrandom" convention. Lower case normal weight letters (like x and β) indicate scalars (real variables). Lowercase bold weight letters (like $\mathbf{x}$ and $\boldsymbol{\beta}$) indicate vectors. Upper case bold weight letters (like $\mathbf{X}$) indicate matrices.

1 The Model

The general linear model has the form

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i \tag{1.1}$$

where i indexes individuals and j indexes different predictor variables. Explicit use of (1.1) makes theory impossibly messy. We rewrite it as a vector equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1.2}$$

where $\mathbf{y}$ is a vector whose components are y_i, where $\mathbf{X}$ is a matrix whose components are x_{ij}, where $\boldsymbol{\beta}$ is a vector whose components are β_j, and where $\mathbf{e}$ is a vector whose components are e_i. Note that $\mathbf{y}$ and $\mathbf{e}$ have dimension n, but $\boldsymbol{\beta}$ has dimension p. The matrix $\mathbf{X}$ is called the *design matrix* or *model matrix* and has dimension $n \times p$.

As always in regression theory, we treat the predictor variables as nonrandom. So $\mathbf{X}$ is a nonrandom matrix, $\boldsymbol{\beta}$ is a nonrandom vector of unknown parameters. The only random quantities in (1.2) are $\mathbf{e}$ and $\mathbf{y}$.

As always in regression theory the errors e_i are independent and identically distributed mean zero normal. This is written as a vector equation

$$\mathbf{e} \sim \text{Normal}(0, \sigma^2 \mathbf{I}),$$

where σ^2 is another unknown parameter (the error variance) and $\mathbf{I}$ is the identity matrix. This implies

$$\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),$$

where

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{1.3}$$

2 Least Squares

If $\mathbf{z}$ is an n-dimensional vector with components z_i, then the *length* of $\mathbf{z}$ is given by

$$\|\mathbf{z}\| = \sqrt{\sum_{i=1}^{n} z_i^2}.$$

The sum is written more conveniently using vector notation

$$\|\mathbf{z}\| = \sqrt{\mathbf{z}'\mathbf{z}}.$$

The least squares estimator of the unknown parameter $\boldsymbol{\beta}$ is the value of $\boldsymbol{\beta}$ that minimizes the "residual sum of squares," which is the squared length of $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

$$Q_1(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Since this is a positive semi-definite quadratic form, the minimum occurs where the gradient vector (vector of first partial derivatives)

$$\nabla Q_1(\boldsymbol{\beta}) = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

is zero. This happens where

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{2.1}$$

These are a set of linear equations to be solved for $\boldsymbol{\beta}$. They are sometimes called the "normal equations."

There is a unique solution to the normal equations, if and only if $\mathbf{X}'\mathbf{X}$ is an invertible matrix, in which case it is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.2}$$

If the inverse does not exist, then (2.2) makes no sense, the solution to (2.1) is not unique, and the least squares estimator of $\boldsymbol{\beta}$ can be taken to be any solution to (2.1).

3 Regression as Projection

Abstractly, regression is orthogonal projection in n-dimensional space. The data $\mathbf{y}$ are a vector of dimension n, which we consider an element of the vector space $\mathbb{R}^n$. Let $\mathbf{x}_1$, …, $\mathbf{x}_p$ denote the columns of $\mathbf{X}$. Then

$$\mathbf{X}\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j \mathbf{x}_j$$

so each possible mean vector (1.3) is a linear combination of the columns of $\mathbf{X}$, and the set of all possible mean vectors is a vector space

$$V = \{\, \mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p \,\} \tag{3.1}$$

called the *column space* of $\mathbf{X}$, because it is the vector space spanned by the columns of $\mathbf{X}$.

We can rephrase the least squares problem as follows. The least squares estimator of the unknown parameter $\boldsymbol{\mu}$ given by (1.3) is the value $\boldsymbol{\mu}$ that minimizes the "residual sum of squares"

$$Q_2(\boldsymbol{\mu}) = \|\mathbf{y} - \boldsymbol{\mu}\|^2 = (\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ ranges over the column space V given by (3.1). Thus least squares finds the μ that is the point in V closest to $\mathbf{y}$. Such a point always exists and is unique, a fact from linear algebra that we will consider intuitively obvious and not prove.

If the $\mathbf{X}'\mathbf{X}$ is invertible, which happens if and only if the columns of $\mathbf{X}$ are linearly independent vectors, then the solution is given by

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

It is traditional to write $\hat{\mathbf{y}}$ instead of $\hat{\boldsymbol{\mu}}$, and we shall follow tradition henceforth, writing

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}, \tag{3.2}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{3.3}$$

has been given the cutesy name *hat matrix* because it puts the "hat" on $\mathbf{y}$. The components of $\hat{\mathbf{y}}$ are called the *predicted values.*

If the columns of $\mathbf{X}$ are not linearly independent, then we can find a maximal linearly independent subset of the columns and form a new model matrix from them, and then (3.2) gives the solution of the least squares problem where (3.3) uses the new model matrix with linearly independent columns.

4 Sampling Distributions

4.1 Regression Coefficients

As a linear function of a normal random vector, $\hat{\boldsymbol{\beta}}$ is normal. It is only necessary to determine its mean vector and variance matrix to determine which normal distribution it has.

Recall the formulas for calculating the mean vector and variance matrix for a linear transformation $\mathbf{a} + \mathbf{B}\mathbf{y}$ of a random vector $\mathbf{y}$

$$E(\mathbf{a} + \mathbf{B}\mathbf{y}) = \mathbf{a} + \mathbf{B}E(\mathbf{y}) \tag{4.1a}$$

and

$$\operatorname{Var}(\mathbf{a} + \mathbf{B}\mathbf{y}) = \mathbf{B}\operatorname{Var}(\mathbf{y})\mathbf{B}' \tag{4.1b}$$

Applying these to the random vector (2.2) gives

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

which says that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ (when the assumptions of the model are true), and

$$\begin{aligned} \operatorname{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\operatorname{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Getting rid of the clutter of the derivations we have the following.

Theorem 4.1. *If* $\mathbf{X}'\mathbf{X}$ *is invertible, then* $\hat{\boldsymbol{\beta}}$ *is multivariate normal and*

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{4.2a}$$

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{4.2b}$$

4.2 Predicted Values and Residuals

It is easily checked that the hat matrix is symmetric (meaning $\mathbf{H} = \mathbf{H}'$) and idempotent (meaning $\mathbf{H}^2 = \mathbf{H}$). Another easily checked property of the hat matrix is $\mathbf{H}\mathbf{X} = \mathbf{X}$. This implies, for example,

$$\mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}. \tag{4.3}$$

The *residuals* are the estimated errors. The error vector is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

and the residual vector is

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \tag{4.4}$$

Theorem 4.2.

$$\hat{\mathbf{y}} \sim \operatorname{Normal}(\boldsymbol{\mu}, \sigma^2\mathbf{H}) \tag{4.5a}$$

$$\hat{\mathbf{e}} \sim \operatorname{Normal}\bigl(0, \sigma^2(\mathbf{I} - \mathbf{H})\bigr) \tag{4.5b}$$

Proof. We only do the latter. The proof of the other is similar. Since the residuals are linear functions of $\mathbf{y}$, they are multivariate normal. Their mean vector is

$$\begin{aligned} E(\hat{\mathbf{e}}) &= (\mathbf{I} - \mathbf{H})E(\mathbf{y}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} \\ &= 0 \end{aligned}$$

because $\mathbf{H}\mathbf{X} = \mathbf{X}$. Their variance matrix is

$$\begin{aligned} \operatorname{Var}(\hat{\mathbf{e}}) &= (\mathbf{I} - \mathbf{H}) \operatorname{Var}(\mathbf{y})(\mathbf{I} - \mathbf{H}) \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

because $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent. □

If h_{ii} denote the diagonal elements of that hat matrix, then this says

$$\operatorname{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii}).$$

So we see that the residuals are multivariate normal and do have mean zero, but they are not homoscedastic nor are they uncorrelated (because $\mathbf{I} - \mathbf{H}$ is not, in general, a diagonal matrix).

4.3 Independence of Residuals and Predicted Values

Theorem 4.3. *The residuals are independent of the predicted values and the regression coefficients.*

Proof. First observe that

$$\mathbf{X}'\hat{\mathbf{e}} = \mathbf{X}'(\mathbf{I} - \mathbf{H})\mathbf{y} = 0 \tag{4.6}$$

because $\mathbf{H}\mathbf{X} = \mathbf{X}$.

Since $E(\hat{\mathbf{e}}) = 0$ by (4.5b), we can write the covariance of $\hat{\mathbf{e}}$ and $\hat{\mathbf{y}}$ as

$$E\{\hat{\mathbf{e}}\hat{\mathbf{y}}'\} = E\{(\mathbf{I} - \mathbf{H})\mathbf{e}(\boldsymbol{\mu} + \mathbf{e})'\mathbf{H}\} = \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{H} = 0$$

because $\mathbf{H}$ is idempotent. Since $\hat{\mathbf{e}}$ and $\hat{\mathbf{y}}$ are jointly multivariate normal, uncorrelated implies independent.

For the regression coefficients we use a trick. We can actually write

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{y}}$$

because

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{y}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

then the result about $\hat{\boldsymbol{\beta}}$ follows from the result about $\hat{\mathbf{y}}$. □

It is also interesting that certain empirical correlations are zero. This happens when there is an "intercept" in the regression model.

Theorem 4.4. *If the vector with all components equal to one is in the column space of the design matrix, then the empirical covariance of* $\hat{\mathbf{e}}$ *with each predictor variable* $\mathbf{x}_j$ *and with* $\hat{\mathbf{y}}$ *is zero.*

Proof. Let $\mathbf{u}$ denote the vector all components equal to one. Note that (4.6) can be rewritten as p equations $\mathbf{x}_j'\hat{\mathbf{e}} = 0$, $j = 1, \ldots, p$. Hence, since $\mathbf{u}$ is a linear combination of the $\mathbf{x}_j$, we also have $\mathbf{u}'\hat{\mathbf{e}} = 0$. This says the empirical expectation of $\hat{\mathbf{e}}$

$$\frac{1}{n} \sum_{i=1}^{n} \hat{e}_i = \frac{\mathbf{u}'\hat{\mathbf{e}}}{n}$$

is zero. Thus the empirical covariance with $\mathbf{x}_j$

$$\frac{1}{n} \sum_{i=1}^{n} \hat{e}_i x_j = \frac{\mathbf{x}_j'\hat{\mathbf{e}}}{n}$$

is also zero. Since $\hat{\mathbf{y}}$ is a linear combination of the $\mathbf{x}_j$ its empirical correlation with $\hat{\mathbf{e}}$ is also zero. □

Thus a plot of residuals versus fitted values or residuals versus any predictor variable should look a plot of independent normals versus whichever.

4.4 Distribution of the Residual Sum of Squares

Any symmetric matrix $\mathbf{A}$ has a *spectral decomposition*

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' \tag{4.7}$$

where $\mathbf{D}$ is diagonal and $\mathbf{O}$ is orthogonal, which means $\mathbf{O}^{-1} = \mathbf{O}'$. We take this as a fact from linear algebra, which we will not prove. The diagonal elements of $\mathbf{D}$ are called the *eigenvalues* of $\mathbf{A}$. The columns of $\mathbf{O}$ are called the *eigenvectors* of $\mathbf{A}$.

If we consider the case where A is symmetric and idempotent we have

$$\begin{aligned} \mathbf{A}^2 &= \mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}' \\ &= \mathbf{O}\mathbf{D}^2\mathbf{O}' \end{aligned}$$

Thus, $\mathbf{O}$ being invertible, $\mathbf{A}^2 = \mathbf{A}$ implies $\mathbf{D}^2 = \mathbf{D}$. Since $\mathbf{D}$ is diagonal, so is $\mathbf{D}^2$. The eigenvalues of a diagonal matrix are its diagonal components. Hence the eigenvalues $\mathbf{D}^2$ are the squares of the eigenvalues of $\mathbf{D}$. Zero and one being the only numbers that are their own squares, all of the eigenvalues of $\mathbf{D}$ are zero or one. These are also the eigenvalues of $\mathbf{A}$.

The spectral decomposition (4.7) can also be written

$$\mathbf{A} = \sum_{i=1}^{n} d_i \mathbf{o}_i \mathbf{o}_i', \tag{4.8}$$

where d_i are the eigenvalues of $\mathbf{A}$ and $\mathbf{o}_i$ are the corresponding eigenvectors. The ordering of the eigenvalues is arbitrary; suppose we order them in decreasing order. Suppose $\mathbf{A}$ is idempotent and p of its eigenvalues are one, in which case we say it has *rank* p. Then its spectral decomposition becomes

$$\mathbf{A} = \sum_{i=1}^{p} \mathbf{o}_i \mathbf{o}_i'$$

and

$$\mathbf{A}\mathbf{y} = \sum_{i=1}^{p} \mathbf{o}_i (\mathbf{o}_i' \mathbf{y}). \tag{4.9}$$

The term in parentheses in (4.9) being scalar, this says any vector in the column space of $\mathbf{A}$ is a linear combination of $\mathbf{o}_1$, $\ldots$, $\mathbf{o}_p$, which, being orthogonal, are linearly independent. Thus we have proved that the dimension of the column space of an idempotent matrix is equal to its rank.

In linear algebra, the rank of a general (not necessarily symmetric) matrix is defined to be the dimension of its column space. For a symmetric matrix, this is the same as the number of nonzero eigenvalues.

Theorem 4.5. *The model matrix and the hat matrix have the same rank.*

Proof. Since any possible mean vector can be written as $\mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$ or as $\mathbf{H}\mathbf{y}$ for some $\mathbf{y}$, it follows that $\mathbf{X}$ and $\mathbf{H}$ have the same column space. □

Since the eigenvalues of $\mathbf{I}$ are all equal to one, the eigenvalues of $\mathbf{I} - \mathbf{H}$ are one minus the corresponding eigenvalues of $\mathbf{H}$. Hence $\operatorname{rank}(\mathbf{I} - \mathbf{H}) = n - \operatorname{rank}(\mathbf{H})$, where n is the sample size.

Note that because of $\mathbf{H}\mathbf{X} = \mathbf{X}$ we have $(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = 0$. Thus we can rewrite (4.4) as

$$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{e}$$

and the residual sum of squares as

$$\text{RSS} = \hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{e}'(\mathbf{I} - \mathbf{H})\mathbf{e} \tag{4.10}$$

again using the idempotency of $\mathbf{I} - \mathbf{H}$.

Theorem 4.6. *If* $\mathbf{A}$ *is any nonrandom symmetric idempotent matrix and* $\mathbf{z}$ *is a multivariate standard normal random vector, then* $\mathbf{z}'\mathbf{A}\mathbf{z}$ *has a chi-square distribution with* $\operatorname{rank} \mathbf{A}$ *degrees of freedom.*

Proof. Write $\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}'$. By Theorem 7.3.2 in DeGroot and Schervish $\mathbf{w} = \mathbf{O}'\mathbf{z}$ is also multivariate standard normal. Then $\mathbf{z}'\mathbf{A}\mathbf{z} = \mathbf{w}'\mathbf{D}\mathbf{w}$. If the first p eigenvalues of $\mathbf{A}$ are equal to one and the rest zero, then

$$\mathbf{w}'\mathbf{D}\mathbf{w} = \sum_{i=1}^{p} w_i^2$$

is chi-square with p degrees of freedom. □

Corollary 4.7. *With RSS given by* (4.10) *RSS*$/\sigma^2$ *has a chi-square distribution with* $n - p$ *degrees of freedom, where* p *is the rank of the model matrix.*

Proof. The vector $\mathbf{z} = \mathbf{e}/\sigma$ is multivariate standard normal,

$$\frac{\text{RSS}}{\sigma^2} = \mathbf{z}'(\mathbf{I} - \mathbf{H})\mathbf{z},$$

$\mathbf{I} - \mathbf{H}$ is symmetric and idempotent and has rank $n - p$. Hence the corollary follows from the theorem. □

5 Model Comparison

Consider two regression models with model matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ and hat matrices $\mathbf{H}_1$ and $\mathbf{H}_2$. We say that these models are *nested* if the column space of one is a subspace of the column space of the other. We refer to them as the big model and the little model.

Lemma 5.1. *If models are nested, then* $\mathbf{H}_1\mathbf{H}_2 = \mathbf{H}_2\mathbf{H}_1 = \mathbf{H}_1$, *where 1 indicates the little model.*

Proof. Every vector of the form $\mathbf{H}_1\mathbf{y}_1$ is equal to $\mathbf{H}_2\mathbf{y}_2$ for some vector $\mathbf{y}_2$. Hence

$$\mathbf{H}_2\mathbf{H}_1\mathbf{y}_1 = \mathbf{H}_2^2\mathbf{y}_2 = \mathbf{H}_2\mathbf{y}_2 = \mathbf{H}_1\mathbf{y}_1,$$

which says $\mathbf{H}_2\mathbf{H}_1 = \mathbf{H}_1$.

The eigenvectors corresponding to eigenvalue one span the column space of $\mathbf{H}_k$. The rest of the eigenvectors span the column space of $\mathbf{I} - \mathbf{H}_k$. Hence every vector of the form $(\mathbf{I} - \mathbf{H}_2)\mathbf{y}_2$ is equal to $(\mathbf{I} - \mathbf{H}_1)\mathbf{y}_1$ for some vector $\mathbf{y}_1$, and

$$\mathbf{H}_1(\mathbf{I} - \mathbf{H}_2)\mathbf{y}_2 = \mathbf{H}_1(\mathbf{I} - \mathbf{H}_1)\mathbf{y}_1 = 0,$$

which says $\mathbf{H}_1\mathbf{H}_2 = \mathbf{H}_1$. □

Theorem 5.2. *If models are nested and both correct, then* $\hat{\mathbf{e}}_2$ *and* $\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1$*, where 1 indicates the little model, are independent normal random vectors. Moreover,* $\|\hat{\mathbf{e}}\|^2/\sigma^2$ *and* $\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2/\sigma^2$ *are independent chi-square random variables with* $n - p_2$ *and* $p_2 - p_1$ *degrees of freedom, respectively, where* p_k *is the rank of the model matrix of model* k*.*

Proof. That $\hat{\mathbf{e}}_2$ and $\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1$ are normal random vectors follows from them being linear functions of $\mathbf{y}$. As before $E(\hat{\mathbf{e}}_2) = 0$ so the covariance of these vectors is

$$\begin{aligned} E\{\hat{\mathbf{e}}_2(\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1)'\} &= E\{(\mathbf{I} - \mathbf{H}_2)\mathbf{e}(\boldsymbol{\mu} + \mathbf{e})(\mathbf{H}_2 - \mathbf{H}_1)\} \\ &= \sigma^2(\mathbf{I} - \mathbf{H}_2)(\mathbf{H}_2 - \mathbf{H}_1) \end{aligned}$$

which equals zero because $(\mathbf{I} - \mathbf{H}_2)(\mathbf{H}_2 - \mathbf{H}_1) = 0$ by the lemma.

We already know the distribution of $\|\hat{\mathbf{e}}_2\|^2/\sigma^2$ from Corollary 4.7, and

$$\begin{aligned} \frac{\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2}{\sigma^2} &= \frac{(\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1)'(\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1)}{\sigma^2} \\ &= \frac{\mathbf{y}'(\mathbf{H}_2 - \mathbf{H}_1)^2\mathbf{y}}{\sigma^2} \end{aligned}$$

By Lemma 5.1

$$(\mathbf{H}_2 - \mathbf{H}_1)^2 = \mathbf{H}_2 - \mathbf{H}_1$$

so $\mathbf{H}_2 - \mathbf{H}_1$ is symmetric and idempotent. Also $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$ with $\boldsymbol{\mu}$ given by (1.3), and $\mathbf{H}_k\boldsymbol{\mu} = \boldsymbol{\mu}$ for both models because both models are correct. Thus

$$\frac{\mathbf{y}'(\mathbf{H}_2 - \mathbf{H}_1)^2\mathbf{y}}{\sigma^2} = \frac{\mathbf{e}'(\mathbf{H}_2 - \mathbf{H}_1)\mathbf{e}}{\sigma^2}$$

and Theorem 4.6 implies that this has a chi-square distribution on $p_2 - p_1$ degrees of freedom, because $\operatorname{rank}(\mathbf{H}_2 - \mathbf{H}_1)$ is the dimension of the column space of $\mathbf{H}_2 - \mathbf{H}_1$, which because the column space of $\mathbf{H}_1$ is contained in the column space of $\mathbf{H}_2$ is the difference of dimensions. □

Thus

$$F = \frac{\|\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1\|^2/(p_2 - p_1)}{\|\hat{\mathbf{e}}_2\|^2/(n - p_2)} \tag{5.1}$$

has an F distribution with $p_2 - p_1$ numinator degrees of freedom and $n - p_2$ denominator degrees of freedom when both models are correct. When only the big model is correct, then the denominator is still chi-square with $n - p_2$ degrees of freedom but the numerator is now approximately $\boldsymbol{\mu}(\mathbf{H}_2 - \mathbf{H}_1)^2\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the true unknown mean of $\mathbf{y}$. Now this is is not zero when the little model is incorrect. Thus (5.1) is a sensible test statistic for testing

H_0: the little model is correct

H_1: the little model is incorrect, but the big model is correct