

Stat 511 Outline

(Spring 2009 Revision of 2008 Version)

Steve Vardeman
Iowa State University

February 27, 2009

Abstract

This outline summarizes the main points of lectures based on Ken Koehler's class notes and other sources.

Contents

1	Linear Models	3
1.1	Ordinary Least Squares	3
1.2	Estimability and Testability	4
1.3	Means and Variances for Ordinary Least Squares (Under the Gauss-Markov Assumptions)	6
1.4	Generalized Least Squares	7
1.5	Reparameterizations and Restrictions	8
1.6	Normal Distribution Theory and Inference	8
1.7	Normal Theory “Maximum Likelihood” and Least Squares	10
1.8	Linear Models and Regression	11
1.9	Linear Models and Two-Way Factorial Analyses	14
2	Nonlinear Models	18
2.1	Ordinary Least Squares in the Nonlinear Model	18
2.2	Inference Methods Based on Large n Theory for the Distribution of MLE's	20
2.3	Inference Methods Based on Large n Theory for Likelihood Ratio Tests/Shape of the Likelihood Function	22
3	Mixed Models	24
3.1	Maximum Likelihood in Mixed Models	25
3.2	Estimation of an Estimable Vector of Parametric Functions $C\beta$	26
3.3	Best Linear Unbiased Prediction and Related Inference in the Mixed Model	26
3.4	Confidence Intervals and Tests for Variance Components	28

3.5	Linear Combinations of Mean Squares and the Cochran-Satterthwaite Approximation	30
3.6	Mixed Models and the Analysis of Balanced Two-Factor Nested Data	31
3.7	Mixed Models and the Analysis of Unreplicated Balanced One-Way Data in Complete Random Blocks	34
3.8	Mixed Models and the Analysis of Balanced Split-Plot/Repeated Measures Data	37
4	Bootstrap Methods	39
4.1	Bootstrapping in the iid (Single Sample) Case	39
4.2	Bootstrapping in Nonlinear Models	44
5	Generalized Linear Models	45
5.1	Inference for the Generalized Linear Model Based on Large Sample Theory for MLEs	47
5.2	Inference for the Generalized Linear Model Based on Large Sample Theory for Likelihood Ratio Tests/Shape of the Likelihood Function	48
5.3	Common GLMs	49
6	Smoothing Methods	49
7	Appendix	52
7.1	Some Useful Facts About Multivariate Distributions (in Particular Multivariate Normal Distributions)	52
7.2	The Multivariate Delta Method	53
7.3	Large Sample Inference Methods	54
7.3.1	Large n Theory for Maximum Likelihood Estimation	54
7.3.2	Large n Theory for Inference Based on the Shape of the Likelihood Function/Likelihood Ratio Testing	55

1 Linear Models

The basic linear model structure is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

for \mathbf{Y} an $n \times 1$ vector of observables, \mathbf{X} an $n \times k$ matrix of known constants, $\boldsymbol{\beta}$ a $k \times 1$ vector of (unknown) constants (parameters), and $\boldsymbol{\epsilon}$ an $n \times 1$ vector of unobservable random errors. Almost always one assumes that $\mathbf{E}\boldsymbol{\epsilon} = \mathbf{0}$. Often one also assumes that for an unknown constant (a parameter) $\sigma^2 > 0$, $\text{Var}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$ (these are the Gauss-Markov model assumptions) or somewhat more generally assumes that $\text{Var}\boldsymbol{\epsilon} = \eta^2\mathbf{V}$ (these are the Aitken model assumptions). These assumptions can be phrased as “the mean vector $\mathbf{E}\mathbf{Y}$ is in the column space of the matrix \mathbf{X} ($\mathbf{E}\mathbf{Y} \in C(\mathbf{X})$) and the variance-covariance matrix $\text{Var}\mathbf{Y}$ is known up to a multiplicative constant.”

1.1 Ordinary Least Squares

The ordinary least squares estimate for $\mathbf{E}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ is made by minimizing

$$(\mathbf{Y} - \widehat{\mathbf{Y}})' (\mathbf{Y} - \widehat{\mathbf{Y}})$$

over choices of $\widehat{\mathbf{Y}} \in C(\mathbf{X})$. $\widehat{\mathbf{Y}}$ is then “the (perpendicular) projection of \mathbf{Y} onto $C(\mathbf{X})$.” This is minimization of the squared distance between \mathbf{Y} and $\widehat{\mathbf{Y}}$ belonging to $C(\mathbf{X})$. Computation of this projection can be accomplished using a (unique) “projection matrix” $\mathbf{P}_{\mathbf{X}}$ as

$$\widehat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$$

There are various ways of constructing $\mathbf{P}_{\mathbf{X}}$. One is as

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-} \mathbf{X}'$$

for $(\mathbf{X}'\mathbf{X})^{-}$ any generalized inverse of $\mathbf{X}'\mathbf{X}$. As it turns out, $\mathbf{P}_{\mathbf{X}}$ is both symmetric and idempotent. It is sometimes called the “hat matrix” and written as \mathbf{H} rather than $\mathbf{P}_{\mathbf{X}}$. (It is used to compute the “ y hats.”)

The vector

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}$$

is the vector of residuals. As it turns out, the matrix $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$ is also a perpendicular projection matrix. It projects onto the subspace of \mathbf{R}^n consisting of all vectors perpendicular to the elements of $C(\mathbf{X})$. That is, $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$ projects onto

$$C(\mathbf{X})^{\perp} \equiv \{\mathbf{u} \in \mathbf{R}^n \mid \mathbf{u}'\mathbf{v} = 0 \ \forall \mathbf{v} \in C(\mathbf{X})\}$$

It is the case that $C(\mathbf{X})^{\perp} = C(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$ and

$$\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{P}_{\mathbf{X}}) = \text{dimension of } C(\mathbf{X}) = \text{trace}(\mathbf{P}_{\mathbf{X}})$$

and

$$\text{rank}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \text{dimension of } C(\mathbf{X})^\perp = \text{trace}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$$

and

$$n = \text{rank}(\mathbf{I}) = \text{rank}(\mathbf{P}_{\mathbf{X}}) + \text{rank}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$$

Further, there is the Pythagorean Theorem/ANOVA identity

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{P}_{\mathbf{X}}\mathbf{Y})'(\mathbf{P}_{\mathbf{X}}\mathbf{Y}) + ((\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y})'((\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}) = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \mathbf{e}'\mathbf{e}$$

When $\text{rank}(\mathbf{X}) = k$ (one has a “full rank” \mathbf{X}) every $\mathbf{w} \in C(\mathbf{X})$ has a unique representation as a linear combination of the columns of \mathbf{X} . In this case there is a unique \mathbf{b} that solves

$$\mathbf{X}\mathbf{b} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} = \widehat{\mathbf{Y}} = \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}} \quad (2)$$

We can call this solution of equation (2) the ordinary least squares estimate of $\boldsymbol{\beta}$. Notice that we then have

$$\mathbf{X}\mathbf{b}_{\text{OLS}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-} \mathbf{X}'\mathbf{Y}$$

so that

$$\mathbf{X}'\mathbf{X}\mathbf{b}_{\text{OLS}} = \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-} \mathbf{X}'\mathbf{Y}$$

These are the so called “normal equations.” $\mathbf{X}'\mathbf{X}$ is $k \times k$ with the same rank as \mathbf{X} (namely k) and is thus non-singular. So $(\mathbf{X}'\mathbf{X})^{-} = (\mathbf{X}'\mathbf{X})^{-1}$ and the normal equations can be solved to give

$$\mathbf{b}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

When \mathbf{X} is not of full rank, there are multiple \mathbf{b} 's that will solve equation (2) and multiple $\boldsymbol{\beta}$'s that could be used to represent $\mathbf{E}\mathbf{Y} \in C(\mathbf{X})$. There is thus no sensible “least squares estimate of $\boldsymbol{\beta}$.”

1.2 Estimability and Testability

In full rank cases, for any $\mathbf{c} \in \mathbf{R}^k$ it makes sense to define the least squares estimate of the linear combination of parameters $\mathbf{c}'\boldsymbol{\beta}$ as

$$\widehat{\mathbf{c}'\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{c}'\mathbf{b}_{\text{OLS}} = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3)$$

When \mathbf{X} is not of full rank, the above expression doesn't make sense. But even in such cases, for some $\mathbf{c} \in \mathbf{R}^k$ the linear combination of parameters $\mathbf{c}'\boldsymbol{\beta}$ can be unambiguous in the sense that every $\boldsymbol{\beta}$ that produces a given $\mathbf{E}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ produces the same value of $\mathbf{c}'\boldsymbol{\beta}$. As it turns out, the \mathbf{c} 's that have this property can be characterized in several ways.

Theorem 1 *The following 3 conditions on $\mathbf{c} \in \mathbf{R}^k$ are equivalent:*

- a) $\exists \mathbf{a} \in \mathbf{R}^n$ such that $\mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} \forall \boldsymbol{\beta} \in \mathbf{R}^k$
- b) $\mathbf{c} \in C(\mathbf{X}')$
- c) $\mathbf{X}\boldsymbol{\beta}_1 = \mathbf{X}\boldsymbol{\beta}_2$ implies that $\mathbf{c}'\boldsymbol{\beta}_1 = \mathbf{c}'\boldsymbol{\beta}_2$

Condition c) says that $\mathbf{c}'\boldsymbol{\beta}$ is unambiguous in the sense mentioned before the statement of the result, condition a) says that there is a linear combination of the entries of \mathbf{Y} that is unbiased for $\mathbf{c}'\boldsymbol{\beta}$, and condition b) says that \mathbf{c}' is a linear combinations of the rows of \mathbf{X} . When $\mathbf{c} \in \mathbf{R}^k$ satisfies the characterizations of Theorem 1 it makes sense to try and estimate $\mathbf{c}'\boldsymbol{\beta}$. Thus, one is led to the following definition.

Definition 2 If $\mathbf{c} \in \mathbf{R}^k$ satisfies the characterizations of Theorem 1 the parametric function $\mathbf{c}'\boldsymbol{\beta}$ is said to be estimable.

If $\mathbf{c}'\boldsymbol{\beta}$ is estimable, $\exists \mathbf{a} \in \mathbf{R}^n$ such that

$$\mathbf{c}'\boldsymbol{\beta} = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{a}'\mathbf{E}\mathbf{Y} \quad \forall \boldsymbol{\beta} \in \mathbf{R}^k$$

and so it makes sense to invent an ordinary least squares estimate of $\mathbf{c}'\boldsymbol{\beta}$ as

$$\widehat{\mathbf{c}'\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{a}'\widehat{\mathbf{Y}} = \mathbf{a}'\mathbf{P}_X\mathbf{Y} = \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

which is a generalization of the full rank formula (3).

It is often of interest to simultaneously estimate several parametric functions

$$\mathbf{c}'_1\boldsymbol{\beta}, \mathbf{c}'_2\boldsymbol{\beta}, \dots, \mathbf{c}'_l\boldsymbol{\beta}$$

If each $\mathbf{c}'_i\boldsymbol{\beta}$ is estimable, it makes sense to assemble the matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}'_1 \\ \mathbf{c}'_2 \\ \vdots \\ \mathbf{c}'_l \end{pmatrix} \quad (4)$$

and talk about estimating the vector

$$\mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{c}'_1\boldsymbol{\beta} \\ \mathbf{c}'_2\boldsymbol{\beta} \\ \vdots \\ \mathbf{c}'_l\boldsymbol{\beta} \end{pmatrix}$$

An ordinary least squares estimator of $\mathbf{C}\boldsymbol{\beta}$ is then

$$\widehat{\mathbf{C}\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Related to the notion of the estimability of $\mathbf{C}\boldsymbol{\beta}$ is the concept of “testability” of hypotheses. Roughly speaking, several hypotheses like $H_0:\mathbf{c}'_i\boldsymbol{\beta} = \#$ are (simultaneously) testable if each $\mathbf{c}'_i\boldsymbol{\beta}$ can be estimated and the hypotheses are not internally inconsistent. To be more precise, suppose that (as above) \mathbf{C} is an $l \times k$ matrix of constants.

Definition 3 For a matrix \mathbf{C} of the form (4), the hypothesis $H_0:\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ is testable provided each $\mathbf{c}'_i\boldsymbol{\beta}$ is estimable and $\text{rank}(\mathbf{C}) = l$.

Cases of testing $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ are of particular interest. If such a hypothesis is testable, $\exists \mathbf{a}_i \in \mathbf{R}^n$ such that $\mathbf{c}'_i = \mathbf{a}'_i \mathbf{X}$ for each i , and thus with

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}'_1 \\ \mathbf{a}'_2 \\ \vdots \\ \mathbf{a}'_l \end{pmatrix}$$

one can write

$$\mathbf{C} = \mathbf{A}\mathbf{X}$$

Then the basic linear model says that $\mathbf{E}\mathbf{Y} \in C(\mathbf{X})$ while the hypothesis says that $\mathbf{E}\mathbf{Y} \in C(\mathbf{A}')^\perp$. $C(\mathbf{X}) \cap C(\mathbf{A}')^\perp$ is a subspace of $C(\mathbf{X})$ of dimension

$$\text{rank}(\mathbf{X}) - \text{rank}(\mathbf{A}') = \text{rank}(\mathbf{X}) - l$$

and the hypothesis can thus be thought of in terms of specifying that the mean vector is in a subspace of $C(\mathbf{X})$.

1.3 Means and Variances for Ordinary Least Squares (Under the Gauss-Markov Assumptions)

Elementary rules about how means and variances of linear combinations of random variables are computed can be applied to find means and variances for OLS estimators. **Under the Gauss-Markov Model** some of these are

$$\begin{aligned} \mathbf{E}\hat{\mathbf{Y}} &= \mathbf{X}\boldsymbol{\beta} \text{ and } \text{Var}\hat{\mathbf{Y}} = \sigma^2 \mathbf{P}_{\mathbf{X}} \\ \mathbf{E}\mathbf{e} &= \mathbf{0} \text{ and } \text{Vare} = \sigma^2 (\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \end{aligned}$$

Further, for l estimable functions $\mathbf{c}'_1\boldsymbol{\beta}, \mathbf{c}'_2\boldsymbol{\beta}, \dots, \mathbf{c}'_l\boldsymbol{\beta}$ and

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}'_1 \\ \mathbf{c}'_2 \\ \vdots \\ \mathbf{c}'_l \end{pmatrix}$$

the estimator $\widehat{\mathbf{C}}\boldsymbol{\beta}_{\text{OLS}} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$ has mean and covariance matrix

$$\mathbf{E}\widehat{\mathbf{C}}\boldsymbol{\beta}_{\text{OLS}} = \mathbf{C}\boldsymbol{\beta} \text{ and } \text{Var}\widehat{\mathbf{C}}\boldsymbol{\beta}_{\text{OLS}} = \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}'$$

Notice that in the case that \mathbf{X} is full rank and $\boldsymbol{\beta} = \mathbf{I}\boldsymbol{\beta}$ is estimable, the above says that

$$\mathbf{E}\mathbf{b}_{\text{OLS}} = \boldsymbol{\beta} \text{ and } \text{Var}\mathbf{b}_{\text{OLS}} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

It is possible to use Theorem 5.2A of Rencher about the mean of a quadratic form to also argue that

$$\mathbf{E}\mathbf{e}'\mathbf{e} = \mathbf{E}(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \sigma^2 (n - \text{rank}(\mathbf{X}))$$

This fact suggests the ratio

$$MSE \equiv \frac{\mathbf{e}'\mathbf{e}}{n - \text{rank}(\mathbf{X})}$$

as an obvious estimate of σ^2 .

In the Gauss-Markov model, ordinary least squares estimation has some optimality properties. Foremost there is the guarantee provided by the Gauss-Markov Theorem. This says that under the linear model assumptions with $\text{Var}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$, for estimable $\mathbf{c}'\boldsymbol{\beta}$ the ordinary least squares estimator $\widehat{\mathbf{c}'\boldsymbol{\beta}}_{\text{OLS}}$ is the Best (in the sense of minimizing variance) Linear (in the entries of \mathbf{Y}) Unbiased (having mean $\mathbf{c}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$) Estimator of $\mathbf{c}'\boldsymbol{\beta}$.

1.4 Generalized Least Squares

For \mathbf{V} positive definite, suppose that $\text{Var}\boldsymbol{\epsilon} = \eta^2\mathbf{V}$. There exists a symmetric positive definite square root matrix for \mathbf{V}^{-1} , call it $\mathbf{V}^{-\frac{1}{2}}$. Then

$$\mathbf{U} = \mathbf{V}^{-\frac{1}{2}}\mathbf{Y}$$

satisfies the Gauss-Markov model assumptions with model matrix

$$\mathbf{W} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}$$

It then makes sense to do ordinary least squares estimation of $\mathbf{EU} \in C(\mathbf{W})$ (with $\mathbf{P}_{\mathbf{W}}\mathbf{U} = \widehat{\mathbf{U}} = \widehat{\mathbf{W}}\boldsymbol{\beta}$). Note that for $\mathbf{c} \in C(\mathbf{W}')$ the BLUE of the parametric function $\mathbf{c}'\boldsymbol{\beta}$ is

$$\mathbf{c}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{U} = \mathbf{c}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-\frac{1}{2}}\mathbf{Y}$$

(any linear function of the elements of \mathbf{U} is a linear function of elements of \mathbf{Y} and vice versa). So, for example, in full rank cases, the best linear unbiased estimator of the $\boldsymbol{\beta}$ vector is

$$\mathbf{b}_{\text{OLS}(\mathbf{U})} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{U} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} \quad (5)$$

Notice that ordinary least squares estimation of \mathbf{EU} is minimization of

$$\left(\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} - \widehat{\mathbf{U}}\right)' \left(\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} - \widehat{\mathbf{U}}\right) = \left(\mathbf{Y} - \mathbf{V}^{\frac{1}{2}}\widehat{\mathbf{U}}\right)' \mathbf{V}^{-1} \left(\mathbf{Y} - \mathbf{V}^{\frac{1}{2}}\widehat{\mathbf{U}}\right)$$

over choices of $\widehat{\mathbf{U}} \in C(\mathbf{W}) = C\left(\mathbf{V}^{-\frac{1}{2}}\mathbf{X}\right)$. This is minimization of

$$\left(\mathbf{Y} - \widehat{\mathbf{Y}}^*\right)' \mathbf{V}^{-1} \left(\mathbf{Y} - \widehat{\mathbf{Y}}^*\right)$$

over choices of $\widehat{\mathbf{Y}}^* \in C\left(\mathbf{V}^{\frac{1}{2}}\mathbf{W}\right) = C(\mathbf{X})$. This is so-called “generalized least squares” estimation in the original variables \mathbf{Y} .

1.5 Reparameterizations and Restrictions

When two superficially different linear models

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and} \quad \mathbf{Y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

(with the same assumptions on $\boldsymbol{\epsilon}$) have $C(\mathbf{X}) = C(\mathbf{W})$, they are really fundamentally the same. $\widehat{\mathbf{Y}}$ and $\mathbf{Y} - \widehat{\mathbf{Y}}$ are the same in the two models. Further,

$$\begin{aligned} \mathbf{c}'\boldsymbol{\beta} \text{ is estimable} &\iff \exists \mathbf{a} \in \mathbf{R}^n \text{ with } \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \\ &\iff \exists \mathbf{a} \in \mathbf{R}^n \text{ with } \mathbf{a}'\mathbf{E}\mathbf{Y} = \mathbf{c}'\boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \end{aligned}$$

That is, $\mathbf{c}'\boldsymbol{\beta}$ is estimable exactly when it is a linear combination of the entries of $\mathbf{E}\mathbf{Y}$. Though this is expressed differently in the two models, it must be the case that the two equivalent models produce the same set of estimable functions, i.e. that

$$\{\mathbf{c}'\boldsymbol{\beta} | \mathbf{c}'\boldsymbol{\beta} \text{ is estimable}\} = \{\mathbf{d}'\boldsymbol{\gamma} | \mathbf{d}'\boldsymbol{\gamma} \text{ is estimable}\}$$

The correspondence between estimable functions in the two model formulations is as follows. Since every column of \mathbf{W} is a linear combination of the columns of \mathbf{X} , there must be a matrix \mathbf{F} such that

$$\mathbf{W} = \mathbf{X}\mathbf{F}$$

Then for an estimable $\mathbf{c}'\boldsymbol{\beta}$, $\exists \mathbf{a} \in \mathbf{R}^n$ such that $\mathbf{c}' = \mathbf{a}'\mathbf{X}$. But then

$$\mathbf{c}'\boldsymbol{\beta} = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{a}'\mathbf{W}\boldsymbol{\gamma} = \mathbf{a}'\mathbf{X}\mathbf{F}\boldsymbol{\gamma} = (\mathbf{c}'\mathbf{F})\boldsymbol{\gamma}$$

So, for example, in the Gauss-Markov model, if $\mathbf{c} \in C(\mathbf{X}')$, the BLUE of $\mathbf{c}'\boldsymbol{\beta} = (\mathbf{c}'\mathbf{F})\boldsymbol{\gamma}$ is $\widehat{\mathbf{c}'\boldsymbol{\beta}}_{\text{OLS}}$.

What then is there to choose between two fundamentally equivalent linear models? There are two issues. Computational/formula simplicity pushes one in the direction of using full rank versions. Sometimes, scientific interpretability of parameters pushes one in the opposite direction. It must be understood that the set of inferences one can make can ONLY depend on the column space of the model matrix, NOT on how that column space is represented.

1.6 Normal Distribution Theory and Inference

If one adds to the basic Gauss-Markov (or Aitken) linear model assumptions an assumption that $\boldsymbol{\epsilon}$ (and therefore \mathbf{Y}) is multivariate normal, inference formulas (for making confidence intervals, tests and predictions) follow. These are based primarily on two basic results.

Theorem 4 (Koehler 4.7, panel 309. See also Rencher Theorem 5.5.A.) Suppose that \mathbf{A} is $n \times n$ and symmetric with $\text{rank}(\mathbf{A}) = k$, $\mathbf{Y} \sim \text{MVN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma}$ positive definite. If $\mathbf{A}\boldsymbol{\Sigma}$ is idempotent, then

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi_k^2(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$$

(So, if in addition $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, then $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi_k^2$.)

Theorem 5 (Theorem 1.3.7 of Christensen) Suppose that $\mathbf{Y} \sim MVN(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ and $\mathbf{B}\mathbf{A} = \mathbf{0}$.

- a) If \mathbf{A} is symmetric, $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ and $\mathbf{B}\mathbf{Y}$ are independent, and
- b) if both \mathbf{A} and \mathbf{B} are symmetric, then $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ and $\mathbf{Y}'\mathbf{B}\mathbf{Y}$ are independent.

(Part b) is Koehler's 4.8. Part a) is a weaker form of Corollary 1 to Rencher's Theorem 5.6.A and part b) is a weaker form of Corollary 1 to Rencher's Theorem 5.6.B.) Here are some implications of these theorems.

Example 6 In the normal Gauss-Markov model

$$\frac{1}{\sigma^2} (\mathbf{Y} - \hat{\mathbf{Y}})' (\mathbf{Y} - \hat{\mathbf{Y}}) = \frac{SSE}{\sigma^2} \sim \chi_{n-\text{rank}(\mathbf{X})}^2$$

This leads, for example, to $1 - \alpha$ level confidence limits for σ^2

$$\left(\frac{SSE}{\text{upper } \frac{\alpha}{2} \text{ point of } \chi_{n-\text{rank}(\mathbf{X})}^2}, \frac{SSE}{\text{lower } \frac{\alpha}{2} \text{ point of } \chi_{n-\text{rank}(\mathbf{X})}^2} \right)$$

Example 7 (Estimation and testing for an estimable function) In the normal Gauss-Markov model, if $\mathbf{c}'\boldsymbol{\beta}$ is estimable,

$$\frac{\widehat{\mathbf{c}'\boldsymbol{\beta}}_{OLS} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{MSE} \sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-\text{rank}(\mathbf{X})}$$

This implies that $H_0: \mathbf{c}'\boldsymbol{\beta} = \#$ can be tested using the statistic

$$T = \frac{\widehat{\mathbf{c}'\boldsymbol{\beta}}_{OLS} - \#}{\sqrt{MSE} \sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}}$$

and a $t_{n-\text{rank}(\mathbf{X})}$ reference distribution. Further, if t is the upper $\frac{\alpha}{2}$ point of the $t_{n-\text{rank}(\mathbf{X})}$ distribution, $1 - \alpha$ level two-sided confidence limits for $\mathbf{c}'\boldsymbol{\beta}$ are

$$\widehat{\mathbf{c}'\boldsymbol{\beta}}_{OLS} \pm t \sqrt{MSE} \sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}$$

Example 8 (Prediction) In the normal Gauss-Markov model, suppose that $\mathbf{c}'\boldsymbol{\beta}$ is estimable and $y^* \sim N(\mathbf{c}'\boldsymbol{\beta}, \gamma\sigma^2)$ independent of \mathbf{Y} is to be observed. (We assume that γ is known.) Then

$$\frac{\widehat{\mathbf{c}'\boldsymbol{\beta}}_{OLS} - y^*}{\sqrt{MSE} \sqrt{\gamma + \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-\text{rank}(\mathbf{X})}$$

This means that if t is the upper $\frac{\alpha}{2}$ point of the $t_{n-\text{rank}(\mathbf{X})}$ distribution, $1 - \alpha$ level two-sided prediction limits for y^* are

$$\widehat{\mathbf{c}'\boldsymbol{\beta}}_{OLS} \pm t \sqrt{MSE} \sqrt{\gamma + \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}$$

Example 9 (Testing) In the normal Gauss-Markov model, suppose that the hypothesis $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ is testable. Then with

$$SS_{H_0} = \left(\widehat{\mathbf{C}\boldsymbol{\beta}}_{OLS} - \mathbf{d} \right)' \left(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' \right)^{-1} \left(\widehat{\mathbf{C}\boldsymbol{\beta}}_{OLS} - \mathbf{d} \right)$$

it's easy to see that

$$\frac{SS_{H_0}}{\sigma^2} \sim \chi_l^2(\delta^2)$$

for

$$\delta^2 = \frac{1}{\sigma^2} (\mathbf{C}\boldsymbol{\beta} - \mathbf{d})' \left(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' \right)^{-1} (\mathbf{C}\boldsymbol{\beta} - \mathbf{d})$$

This in turn implies that

$$F = \frac{SS_{H_0}/l}{MSE} \sim F_{l, n-\text{rank}(\mathbf{X})}(\delta^2)$$

So with f the upper α point of the $F_{l, n-\text{rank}(\mathbf{X})}$ distribution, an α level test of $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ can be made by rejecting if $\frac{SS_{H_0}/l}{MSE} > f$. The power of this test is

$$\text{power}(\delta^2) = P[\text{an } F_{l, n-\text{rank}(\mathbf{X})}(\delta^2) \text{ random variable} > f]$$

Or taking a significance testing point of view, a p -value for testing this hypothesis is

$$P \left[\text{an } F_{l, n-\text{rank}(\mathbf{X})} \text{ random variable exceeds the observed value of } \frac{SS_{H_0}/l}{MSE} \right]$$

1.7 Normal Theory “Maximum Likelihood” and Least Squares

One way to justify/motivate the use of least squares in the linear model is through appeal to the statistical principle of “maximum likelihood.” That is, in the normal Gauss-Markov model, the joint pdf for the n observations is

$$\begin{aligned} f(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2) &= (2\pi)^{-\frac{n}{2}} |\det(\sigma^2\mathbf{I})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right) \\ &= (2\pi)^{-\frac{n}{2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right) \end{aligned}$$

Clearly, for fixed σ this is maximized as a function of $\mathbf{X}\boldsymbol{\beta} \in C(\mathbf{X})$ by minimizing

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

i.e. with $\widehat{\mathbf{Y}} = \widehat{\mathbf{X}\boldsymbol{\beta}}_{OLS}$. Then consider

$$\ln f(\mathbf{Y}|\widehat{\mathbf{Y}}, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}SSE$$

Setting the derivative with respect to σ^2 equal to 0 and solving shows this function of σ^2 to be maximized when $\sigma^2 = \frac{SSE}{n}$. That is, the (joint) maximum

likelihood estimator of the parameter vector $(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$ is $(\widehat{\mathbf{Y}}, \frac{SSE}{n})$. It is worth noting that

$$\frac{SSE}{n} = \left(\frac{n - \text{rank}(\mathbf{X})}{n} \right) MSE$$

so that

$$E \frac{SSE}{n} = \left(\frac{n - \text{rank}(\mathbf{X})}{n} \right) \sigma^2 < \sigma^2$$

and the MLE of σ^2 is biased low.

1.8 Linear Models and Regression

A most important application of the general framework of the linear model is to (multiple linear) regression analysis, usually represented in the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_r x_{ri} + \epsilon_i$$

for “predictor variables” x_1, x_2, \dots, x_r . This can, of course, be written in the usual linear model format (1) for $k = r + 1$ and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{r1} \\ 1 & x_{12} & x_{22} & \cdots & x_{r2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{rn} \end{pmatrix} = (\mathbf{1} | \mathbf{x}_1 | \mathbf{x}_2 | \cdots | \mathbf{x}_r)$$

Unless n is small or one is very unlucky, in regression contexts \mathbf{X} is of full rank (i.e. is of rank $r + 1$). A few specifics of what has gone before that are of particular interest in the regression context are as follows.

As always $\widehat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}} \mathbf{Y}$ and in the regression context it is especially common to call $\mathbf{P}_{\mathbf{X}} = \mathbf{H}$ the hat matrix. It is $n \times n$ and its diagonal entries h_{ii} are sometimes used as indices of “influence” or “leverage” of a particular case on the regression fit. It is the case that each $h_{ii} \geq 0$ and

$$\sum h_{ii} = \text{trace}(\mathbf{H}) = \text{rank}(\mathbf{H}) = r + 1$$

so that the “hats” h_{ii} average to $\frac{r+1}{n}$. In light of this, a case with $h_{ii} > \frac{2(r+1)}{n}$ is sometimes flagged as an “influential” case. Further

$$\text{Var} \widehat{\mathbf{Y}} = \sigma^2 \mathbf{P}_{\mathbf{X}} = \sigma^2 \mathbf{H}$$

so that $\text{Var} \widehat{y}_i = h_{ii} \sigma^2$ and an estimated standard deviation of \widehat{y}_i is $\sqrt{h_{ii}} \sqrt{MSE}$. (This is useful, for example, in making confidence intervals for $E y_i$.)

Also as always, $\mathbf{e} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \mathbf{Y}$ and $\text{Vare} = \sigma^2 (\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \sigma^2 (\mathbf{I} - \mathbf{H})$. So $\text{Vare}_i = (1 - h_{ii}) \sigma^2$ and it is typical to compute and plot standardized versions of the residuals

$$e_i^* = \frac{e_i}{\sqrt{MSE} \sqrt{1 - h_{ii}}}$$

The general testing of hypothesis framework discussed in Section 1.6 has a particular important specialization in regression contexts. That is, it is common in regression contexts (for $p < r$) to test

$$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_r = 0 \quad (6)$$

and in first methods courses this is done using the “full model/reduced model” paradigm. With

$$\mathbf{X}_i = (\mathbf{1} | \mathbf{x}_1 | \mathbf{x}_2 | \cdots | \mathbf{x}_i)$$

this is the hypothesis

$$H_0 : E\mathbf{Y} \in C(\mathbf{X}_p)$$

It is also possible to write this hypothesis in the standard form $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ using the matrix

$$\mathbf{C}_{(r-p) \times (r+1)} = \left(\begin{array}{c|c} \mathbf{0}_{(r-p) \times (p+1)} & \mathbf{I}_{(r-p) \times (r-p)} \end{array} \right)$$

So from Section 1.6 the hypothesis can be tested using an F test with numerator sum of squares

$$SS_{H_0} = (\mathbf{C}\mathbf{b}_{OLS})' \left(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}' \right)^{-1} (\mathbf{C}\mathbf{b}_{OLS})$$

What is interesting and perhaps not initially obvious is that

$$SS_{H_0} = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_p})\mathbf{Y} \quad (7)$$

and that this kind of sum of squares *is* the elementary $SSR_{\text{full}} - SSR_{\text{reduced}}$. (A proof of the equivalence (7) is on a handout posted on the course web page.)

Further, the sum of squares in display (7) can be made part of any number of interesting partitions of the (uncorrected) overall sum of squares $\mathbf{Y}'\mathbf{Y}$. For example, it is clear that

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'(\mathbf{P}_1 + (\mathbf{P}_{\mathbf{X}_p} - \mathbf{P}_1) + (\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_p}) + (\mathbf{I} - \mathbf{P}_{\mathbf{X}}))\mathbf{Y} \quad (8)$$

so that

$$\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}_1\mathbf{Y} = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_p} - \mathbf{P}_1)\mathbf{Y} + \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_p})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}$$

In elementary regression analysis notation

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}_1\mathbf{Y} &= SSTot \text{ (corrected)} \\ \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_p} - \mathbf{P}_1)\mathbf{Y} &= SSR_{\text{reduced}} \\ \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_p})\mathbf{Y} &= SSR_{\text{full}} - SSR_{\text{reduced}} \\ \mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} &= SSE_{\text{full}} \end{aligned}$$

(and then of course $\mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_1)\mathbf{Y} = SSR_{\text{full}}$). These four sums of squares are often arranged in an ANOVA table for testing the hypothesis (6).

It is common in regression analysis to use “reduction in sums of squares” notation and write

$$\begin{aligned} R(\beta_0) &= \mathbf{Y}'\mathbf{P}_1\mathbf{Y} \\ R(\beta_1, \dots, \beta_p|\beta_0) &= \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_p} - \mathbf{P}_1)\mathbf{Y} \\ R(\beta_{p+1}, \dots, \beta_r|\beta_0, \beta_1, \dots, \beta_p) &= \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_p})\mathbf{Y} \end{aligned}$$

so that in this notation, identity (8) becomes

$$\mathbf{Y}'\mathbf{Y} = R(\beta_0) + R(\beta_1, \dots, \beta_p|\beta_0) + R(\beta_{p+1}, \dots, \beta_r|\beta_0, \beta_1, \dots, \beta_p) + SSE$$

And in fact, even more elaborate breakdowns of the overall sum of squares are possible. For example,

$$\begin{aligned} R(\beta_0) &= \mathbf{Y}'\mathbf{P}_1\mathbf{Y} \\ R(\beta_1|\beta_0) &= \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_1)\mathbf{Y} \\ R(\beta_2|\beta_0, \beta_1) &= \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_2} - \mathbf{P}_{\mathbf{X}_1})\mathbf{Y} \\ &\vdots \\ R(\beta_r|\beta_0, \beta_1, \dots, \beta_{r-1}) &= \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_{r-1}})\mathbf{Y} \end{aligned}$$

represents a “Type I” or “Sequential” sum of squares breakdown of $\mathbf{Y}'\mathbf{Y} - SSE$. (Note that these sums of squares are appropriate numerator sums of squares for testing significance of individual β 's in models that include terms only up to the one in question.)

The enterprise of trying to assign a sum of squares to a predictor variable strikes Vardeman as of little real interest, but is nevertheless a common one. Rather than think of

$$R(\beta_i|\beta_0, \beta_1, \dots, \beta_{i-1}) = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_i} - \mathbf{P}_{\mathbf{X}_{i-1}})\mathbf{Y}$$

(which depends on the usually essentially arbitrary ordering of the columns of \mathbf{X}) as “due to x_i ” it is possible to invent other assignments of sums of squares. One is the “SAS Type II Sum of Squares” assignment. With

$$\mathbf{X}_{\bar{i}} = (\mathbf{1}|\mathbf{x}_1|\dots|\mathbf{x}_{i-1}|\mathbf{x}_{i+1}|\dots|\mathbf{x}_r)$$

(the original model matrix with the column \mathbf{x}_i deleted), These are

$$R(\beta_i|\beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_r) = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_{\bar{i}}})\mathbf{Y}$$

(appropriate numerator sums of squares for testing significance of individual β 's in the full model).

It is worth noting that Theorem B.47 of Christensen guarantees that any matrix like $\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_i}$ is a perpendicular projection matrix and that Theorem B.48 implies that $C(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_i}) = C(\mathbf{X}) \cap C(\mathbf{X}_i)^\perp$ (the orthogonal complement of $C(\mathbf{X}_i)$ with respect to $C(\mathbf{X})$ defined on page 395 of Christensen). Further, it is easy enough to argue using Christensen's Theorem 1.3.7b (Koehler's 4.7) that any set of sequential sums of squares has pair-wise independent elements. And one can apply Cochran's Theorem (Koehler's 4.9 on panel 333) to conclude that the whole set (including SSE) are mutually independent.

1.9 Linear Models and Two-Way Factorial Analyses

As a second application/specialization of the general linear model framework we consider the two-way factorial context, That is, for I levels of Factor A and J levels of Factor B, we consider situations where one can make observations under every combination of a level of A and a level of B. For $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$

y_{ijk} = the k th observation at level i of A and level j of B

The most straightforward way to think about such a situation is in terms of the “cell means model”

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad (9)$$

In this full rank model, provided every “within cell” sample size n_{ij} is positive, each μ_{ij} is estimable and thus so too is any linear combination of these means.

It is standard to think of the IJ means μ_{ij} as laid out in a two-way table as

μ_{11}	μ_{12}	\cdots	μ_{1J}
μ_{21}	μ_{22}	\cdots	μ_{2J}
\vdots	\vdots	\ddots	\vdots
μ_{I1}	μ_{I2}	\cdots	μ_{IJ}

Particular interesting parametric functions ($\mathbf{c}'\boldsymbol{\beta}$'s) in model (9) are built on row, column and grand average means

$$\mu_{i.} = \frac{1}{J} \sum_{j=1}^J \mu_{ij} \quad \text{and} \quad \mu_{.j} = \frac{1}{I} \sum_{i=1}^I \mu_{ij} \quad \text{and} \quad \mu_{..} = \frac{1}{IJ} \sum_{i,j} \mu_{ij}$$

Each of these is a linear combination of the $I \times J$ means μ_{ij} and is thus estimable. So are the linear combinations of them

$$\alpha_i = \mu_{i.} - \mu_{..}, \beta_j = \mu_{.j} - \mu_{..}, \quad \text{and} \quad \alpha\beta_{ij} = \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j) \quad (10)$$

The “factorial effects” (10) here are particular (estimable) linear combinations of the cell means. It is a consequence of how these are defined that

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \alpha\beta_{ij} = 0 \quad \forall j, \quad \text{and} \quad \sum_j \alpha\beta_{ij} = 0 \quad \forall i \quad (11)$$

An issue of particular interest in two way factorials is whether the hypothesis

$$H_0: \alpha\beta_{ij} = 0 \quad \forall i \text{ and } j \quad (12)$$

is tenable. (If it is, great simplification of interpretation is possible ... changing levels of one factor has the same impact on mean response regardless of which level of the second factor is considered.) This hypothesis can be equivalently written as

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j \quad \forall i \text{ and } j$$

or as

$$(\mu_{ij} - \mu_{ij'}) - (\mu_{i'j} - \mu_{i'j'}) = 0 \quad \forall i, i', j \text{ and } j'$$

and is a statement of “parallelism” on “interaction plots” of means. To test this, one *could* write the hypothesis in terms of $(I - 1)(J - 1)$ statements

$$\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} = 0$$

about the cell means and use the machinery for testing $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ from Example 9. In this case, $\mathbf{d} = \mathbf{0}$ and the test is about $\mathbf{E}\mathbf{Y}$ falling in some subspace of $C(\mathbf{X})$. For thinking about the nature of this subspace and issues related to the hypothesis (12), it is probably best to back up and consider an alternative to the cell means model approach.

Rather than begin with the cell means model, one might instead begin with the non-full-rank “effects model”

$$y_{ijk} = \mu^* + \alpha_i^* + \beta_j^* + \alpha\beta_{ij}^* + \epsilon_{ijk} \quad (13)$$

I have put stars on the parameters to make clear that this is something different from beginning with cell means and defining effects as linear combinations of them. Here there are $k = 1 + I + J + IJ$ parameters for the means and only IJ different means. A model including all of these parameters can not be of full rank. To get simple computations/formulas, one must impose some restrictions. There are several possibilities.

In the first place, the facts (11) suggest the so called “sum restrictions” in the effects model (13)

$$\sum_i \alpha_i^* = 0, \quad \sum_j \beta_j^* = 0, \quad \sum_i \alpha\beta_{ij}^* = 0 \quad \forall j, \quad \text{and} \quad \sum_j \alpha\beta_{ij}^* = 0 \quad \forall i$$

Alternative restrictions are so-called “baseline restrictions.” SAS uses the baseline restrictions

$$\alpha_I^* = 0, \quad \beta_J^* = 0, \quad \alpha\beta_{Ij}^* = 0 \quad \forall j, \quad \text{and} \quad \alpha\beta_{iJ}^* = 0 \quad \forall i$$

while R and Splus use the baseline restrictions

$$\alpha_1^* = 0, \quad \beta_1^* = 0, \quad \alpha\beta_{1j}^* = 0 \quad \forall j, \quad \text{and} \quad \alpha\beta_{i1}^* = 0 \quad \forall i$$

Under any of these sets of restrictions one may write a full rank model matrix as

$$\mathbf{X} = \left(\begin{array}{c|c|c|c} \mathbf{1} & \mathbf{X}_{\alpha^*} & \mathbf{X}_{\beta^*} & \mathbf{X}_{\alpha\beta^*} \\ \hline n \times 1 & n \times (I-1) & n \times (J-1) & n \times (I-1)(J-1) \end{array} \right)$$

and the no interaction hypothesis (12) is the hypothesis $H_0: \mathbf{E}\mathbf{Y} \in C((\mathbf{1} | \mathbf{X}_{\alpha^*} | \mathbf{X}_{\beta^*}))$. So using the full model/reduced model paradigm from the regression discussion, one then has an appropriate numerator sum of squares

$$SS_{H_0} = \mathbf{Y}' \left(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{(\mathbf{1} | \mathbf{X}_{\alpha^*} | \mathbf{X}_{\beta^*})} \right) \mathbf{Y}$$

and numerator degrees of freedom $(I - 1)(J - 1)$ (in complete factorials where every $n_{ij} > 0$).

Other hypotheses sometimes of interest are

$$H_0: \alpha_i = 0 \forall i \text{ or } H_0: \beta_j = 0 \forall j \quad (14)$$

These are the hypotheses that all row averages of cell means are the same and that all column averages of cell means are the same. That is, these hypotheses could be written as

$$H_0: \mu_{i.} - \mu_{i'.} = 0 \forall i, i' \text{ or } H_0: \mu_{.j} - \mu_{.j'} = 0 \forall j, j'$$

It is possible to write the first of these in the cell means model as $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for \mathbf{C} that is $(I - 1) \times k$ and each row of \mathbf{C} specifying $\alpha_i = 0$ for one of $i = 1, 2, \dots, (I - 1)$ (or equality of two row average means). Similarly, the second can be written in the cell means model as $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for \mathbf{C} that is $(J - 1) \times k$ and each row of \mathbf{C} specifying $\beta_j = 0$ for one of $j = 1, 2, \dots, (J - 1)$ (or equality of two column average means). Appropriate numerator sums of squares and degrees of freedom for testing these hypotheses are then obvious using the material of Example 9. These sums of squares are often referred to as “Type III” sums of squares.

How to interpret standard partitions of sums of squares and to relate them to tests of hypotheses (12) and (14) is problematic unless all “cell” sample sizes are the same (all $n_{ij} = m$, the data are “balanced”). That is, depending upon what kind of partition one asks for in a call of a standard two-way ANOVA routine, the program produces the following breakdowns

“Source”	Type I SS (in the order A,B,A × B)	Type II SS	Type III SS
A	$R(\alpha^*{}^{\prime}s \mu^*)$	$R(\alpha^*{}^{\prime}s \mu^*,\beta^{*}{}^{\prime}s)$	SS_{H_0} for type (14) hypothesis in model (9)
B	$R(\beta^{*}{}^{\prime}s \mu^*,\alpha^*{}^{\prime}s)$	$R(\beta^{*}{}^{\prime}s \mu^*,\alpha^*{}^{\prime}s)$	SS_{H_0} for type (14) hypothesis in model (9)
A × B	$R(\alpha\beta^{*}{}^{\prime}s \mu^*,\alpha^*{}^{\prime}s,\beta^{*}{}^{\prime}s)$	$R(\alpha\beta^{*}{}^{\prime}s \mu^*,\alpha^*{}^{\prime}s,\beta^{*}{}^{\prime}s)$	SS_{H_0} for type (12) hypothesis in model (9)

The “A × B” sums of squares of all three types are the same. *When the data are balanced*, the “A” and “B” sums of squares of all three types are also the same. But when the sample sizes are not the same, the “A” and “B” sums of squares of different “types” are generally not the same, only the Type III sums of squares are appropriate for testing hypotheses (14), and exactly what hypotheses could be tested using the Type I or Type II sums of squares is both hard to figure

out and actually quite bizarre when one does figure it out. (They correspond to certain hypotheses about weighted averages of means that use sample sizes as weights. See Koehler’s notes for more on this point.) From Vardeman’s perspective, the Type III sums of squares “have a reason to be” in this context, while the Type I and Type II sums of squares do not. (This is in spite of the fact that the Type I breakdown is an honest partition of an overall sum of squares, while the Type III breakdown is not.)

An issue related to lack of balance in a two-way factorial is the issue of “empty cells” in a two-way factorial. That is, suppose that there are data for only $k < IJ$ combinations of levels of A and B. “Full rank” in this context is “rank k ” and the cell means model (9) can have only $k < IJ$ parameters for the means. However, if the $I \times J$ table of data isn’t “too sparse” (if the number of empty cells isn’t too large and those that are empty are not in a nasty configuration) it is still possible to fit both a cell means model and a no-interaction effects model to the data. This allows the testing of

H_0 : the μ_{ij} for which one has data show no interactions

i.e.

$$H_0: (\mu_{ij} - \mu_{ij'}) - (\mu_{i'j} - \mu_{i'j'}) = 0 \quad \forall \text{ quadruples} \quad (15)$$

$(i, j), (i, j'), (i', j)$ and (i', j') complete in the data set

and the estimation of all cell means *under an assumption that a no-interaction effects model appropriate to the cells where one has data extends to all $I \times J$ cells in the table.*

That is, let \mathbf{X} be the cell means model matrix (for k “full” cells) and

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{1} & | & \mathbf{X}_{\alpha^*} & | & \mathbf{X}_{\beta^*} \\ n \times k & & n \times 1 & & n \times (I-1) & & n \times (J-1) \end{pmatrix}$$

be an appropriate restricted version of an effects model matrix (with no interaction terms). *If the pattern of empty cells is such that \mathbf{X}^* is full rank* (has rank $I + J - 1$), the hypothesis (15) can be tested using

$$F = \frac{\mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}^*})\mathbf{Y} / (k - (I + J - 1))}{\mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} / (n - k)}$$

and an $F_{(k-(I+J-1)),(n-k)}$ reference distribution. Further, every

$$\mu^* + \alpha_i^* + \beta_j^*$$

is estimable in the no interaction effects model. *Provided this model extends to all $I \times J$ combinations of levels of A and B*, this provides estimates of mean responses for all cells. (Note that this is essentially the same kind of extrapolation one does in a regression context to sets of predictors not in the original data set. However, on an intuitive basis, the link supporting extrapolation is probably stronger with quantitative regressors than it is with the qualitative predictors of the present context.)

2 Nonlinear Models

A generalization of the linear model is the (potentially) “nonlinear” model that for $\boldsymbol{\beta}$ a $k \times 1$ vector of (unknown) constants (parameters) and for some function

$$f(\mathbf{x}, \boldsymbol{\beta})$$

that is smooth (differentiable) in the elements of $\boldsymbol{\beta}$, says that what is observed can be represented as

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i \quad (16)$$

for each \mathbf{x}_i a known vector of constants. (The dimension of \mathbf{x} is fixed but basically irrelevant for what follows. In particular, it need not be k .) As is typical in the linear model, one usually assumes that $E\epsilon_i = 0 \forall i$, and it is also common to assume that for an unknown constant (a parameter) $\sigma^2 > 0$, $\text{Var}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$.

2.1 Ordinary Least Squares in the Nonlinear Model

In general (unlike the case when $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i' \boldsymbol{\beta}$ and the model (16) is a linear model) there are typically no explicit formulas for least squares estimation of $\boldsymbol{\beta}$. That is, minimization of

$$g(\mathbf{b}) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \mathbf{b}))^2 \quad (17)$$

is a problem in numerical analysis. There are a variety of standard algorithms used for this purpose. They are all based on the fact that a necessary condition for \mathbf{b}_{OLS} to be a minimizer of $g(\mathbf{b})$ is that

$$\left. \frac{\partial g}{\partial b_j} \right|_{\mathbf{b}=\mathbf{b}_{\text{OLS}}} = 0 \quad \forall j$$

so that in search for an ordinary least squares estimator, one might try to find a simultaneous solution to these k “estimating” equations. A bit of calculus and algebra shows that \mathbf{b}_{OLS} must then solve the matrix equation

$$\mathbf{0} = \mathbf{D}'_{\mathbf{b}} (\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b})) \quad (18)$$

where we use the notations

$$\mathbf{D}_{\mathbf{b}} = \begin{pmatrix} \frac{\partial f(\mathbf{x}_1, \mathbf{b})}{\partial b_j} \\ \frac{\partial f(\mathbf{x}_2, \mathbf{b})}{\partial b_j} \\ \vdots \\ \frac{\partial f(\mathbf{x}_n, \mathbf{b})}{\partial b_j} \end{pmatrix}_{n \times k} \quad \text{and} \quad \mathbf{f}(\mathbf{X}, \mathbf{b}) = \begin{pmatrix} f(\mathbf{x}_1, \mathbf{b}) \\ f(\mathbf{x}_2, \mathbf{b}) \\ \vdots \\ f(\mathbf{x}_n, \mathbf{b}) \end{pmatrix}_{n \times 1}$$

In the case of the linear model

$$\mathbf{D}_{\mathbf{b}} = \left(\frac{\partial}{\partial b_j} \mathbf{x}_i' \mathbf{b} \right) = (x_{ij}) = \mathbf{X} \quad \text{and} \quad \mathbf{f}(\mathbf{X}, \mathbf{b}) = \mathbf{X}\mathbf{b}$$

so that equation (18) is $\mathbf{0} = \mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b})$, i.e. is the set of normal equations $\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. (It is important that for the linear model, the partial derivative matrix $\mathbf{D}_{\mathbf{b}}$ does not depend upon \mathbf{b} .)

One of many iterative algorithms for searching for a solution to the equation (18) is the Gauss-Newton algorithm. It proceeds as follows. For

$$\mathbf{b}^r = \begin{pmatrix} b_1^r \\ b_2^r \\ \vdots \\ b_k^r \end{pmatrix}$$

the approximate solution produced by the r th iteration of the algorithm (\mathbf{b}^0 is some vector of starting values that must be supplied by the user), let

$$\mathbf{D}^r = \mathbf{D}_{\mathbf{b}^r} = \left(\frac{\partial f(\mathbf{x}_i, \mathbf{b})}{\partial b_j} \bigg|_{\mathbf{b}=\mathbf{b}^r} \right)$$

The first order Taylor (linear) approximation to $\mathbf{f}(\mathbf{X}, \boldsymbol{\beta})$ at \mathbf{b}^r is

$$\mathbf{f}(\mathbf{X}, \boldsymbol{\beta}) \approx \mathbf{f}(\mathbf{X}, \mathbf{b}^r) + \mathbf{D}^r (\boldsymbol{\beta} - \mathbf{b}^r)$$

So the nonlinear model $\mathbf{Y} = \mathbf{f}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\epsilon}$ can be written as

$$\mathbf{Y} \approx \mathbf{f}(\mathbf{X}, \mathbf{b}^r) + \mathbf{D}^r (\boldsymbol{\beta} - \mathbf{b}^r) + \boldsymbol{\epsilon}$$

which can be written in linear model form (for the “response vector” $\mathbf{Y}^* = (\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b}^r))$) as

$$(\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b}^r)) \approx \mathbf{D}^r (\boldsymbol{\beta} - \mathbf{b}^r) + \boldsymbol{\epsilon}$$

In this context, ordinary least squares would say that $\boldsymbol{\beta} - \mathbf{b}^r$ could be approximated as

$$(\widehat{\boldsymbol{\beta} - \mathbf{b}^r})_{\text{OLS}} = (\mathbf{D}^{r'}\mathbf{D}^r)^{-1} \mathbf{D}^{r'} (\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b}^r))$$

which in turn suggests that the $(r+1)$ st iterate of \mathbf{b} be taken as

$$\mathbf{b}^{r+1} = \mathbf{b}^r + (\mathbf{D}^{r'}\mathbf{D}^r)^{-1} \mathbf{D}^{r'} (\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b}^r))$$

These iterations are continued until some convergence criterion is satisfied. One type of convergence criterion is based on the “deviance” (what in a linear model would be called the error sum of squares). At the r th iteration this is

$$SSE^r = g(\mathbf{b}^r) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \mathbf{b}^r))^2$$

and in rough terms, the convergence criterion is to stop when this ceases to decrease. Other criteria have to do with (relative) changes in the entries of \mathbf{b}^r (one stops when \mathbf{b}^r ceases to change much). The R function `nls` implements this

algorithm (both in versions where the user must supply formulas for derivatives and where the algorithm itself computes approximate/numerical derivatives). Koehler's note discuss two other algorithms, the Newton-Raphson algorithm and the Fisher scoring algorithm, as they are applied to the (iid normal errors version of the) loglikelihood function associated with the nonlinear model (16). Our fundamental interest here is not in the numerical analysis of how one finds a minimizer of the function (17), \mathbf{b}_{OLS} , but rather how that estimator can be used in making inferences. There are two routes to inference based on complimentary large sample theories connected with \mathbf{b}_{OLS} . We outline those in the next two subsections.

2.2 Inference Methods Based on Large n Theory for the Distribution of MLE's

Just as in the linear model case discussed in Section 1.7, \mathbf{b}_{OLS} and SSE/n are (joint) maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 under the iid normal errors nonlinear model. General theory about the consistency and asymptotic normality of maximum likelihood estimators (outlined, for example, in Section 7.3.1 of the Appendix) suggests the following.

Claim 10 *Precise statements of the approximations below can be made in some large n circumstances.*

1) $\mathbf{b}_{OLS} \sim MVN_k(\boldsymbol{\beta}, \sigma^2 (\mathbf{D}'\mathbf{D})^{-1})$ where $\mathbf{D} = \mathbf{D}_{\boldsymbol{\beta}} = \left(\frac{\partial f(\mathbf{x}_i, \mathbf{b})}{\partial b_j} \Big|_{\mathbf{b}=\boldsymbol{\beta}} \right)$. Further,

$(\mathbf{D}'\mathbf{D})^{-1}$ "typically gets small" with increasing sample size.

2) $MSE = \frac{SSE}{n-k} \approx \sigma^2$.

3) $(\mathbf{D}'\mathbf{D})^{-1} \approx (\widehat{\mathbf{D}}'\widehat{\mathbf{D}})^{-1}$ where $\widehat{\mathbf{D}} = \mathbf{D}_{\mathbf{b}_{OLS}} = \left(\frac{\partial f(\mathbf{x}_i, \mathbf{b})}{\partial b_j} \Big|_{\mathbf{b}=\mathbf{b}_{OLS}} \right)$.

4) For a smooth (differentiable) function \mathbf{h} that maps $\Re^k \rightarrow \Re^q$, claim 1) and the "delta method" (Taylor's Theorem of Section 7.2 of the Appendix) imply that

$\mathbf{h}(\mathbf{b}_{OLS}) \sim MVN_q(\mathbf{h}(\boldsymbol{\beta}), \sigma^2 \mathbf{G}(\mathbf{D}'\mathbf{D})^{-1} \mathbf{G}')$ for $\mathbf{G}_{q \times k} = \left(\frac{\partial h_i(\mathbf{b})}{\partial b_j} \Big|_{\mathbf{b}=\boldsymbol{\beta}} \right)$.

5) $\mathbf{G} \approx \widehat{\mathbf{G}} = \left(\frac{\partial h_i(\mathbf{b})}{\partial b_j} \Big|_{\mathbf{b}=\mathbf{b}_{OLS}} \right)$.

Using this set of approximations, essentially exactly as in Section 1.6, one can develop inference methods. Some of these are outlined below.

Example 11 *(Inference for a single β_j)* From part 1) of Claim 10 we get the approximation

$$\frac{b_{OLSj} - \beta_j}{\sigma \sqrt{\eta_j}} \sim N(0, 1)$$

for η_j the j th diagonal entry of $(\mathbf{D}'\mathbf{D})^{-1}$. But then from parts 2) and 3) of the claim,

$$\frac{b_{OLSj} - \beta_j}{\sigma\sqrt{\eta_j}} \approx \frac{b_{OLSj} - \beta_j}{\sqrt{MSE}\sqrt{\hat{\eta}_j}}$$

for $\hat{\eta}_j$ the j th diagonal entry of $(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}$. In the (normal Gauss-Markov) linear model context, this last random variable is in fact t distributed for any n . Then both so that the nonlinear model formulas reduce to the linear model formulas, and as a means of making the already very approximate inference formulas somewhat more conservative, it is standard to say

$$\frac{b_{OLSj} - \beta_j}{\sqrt{MSE}\sqrt{\hat{\eta}_j}} \sim t_{n-k}$$

and thus to test $H_0:\beta_j = \#$ using

$$T = \frac{b_{OLSj} - \#}{\sqrt{MSE}\sqrt{\hat{\eta}_j}}$$

and a t_{n-k} reference distribution, and to use the values

$$b_{OLSj} \pm t\sqrt{MSE}\sqrt{\hat{\eta}_j}$$

as confidence limits for β_j .

Example 12 (Inference for a univariate function of $\boldsymbol{\beta}$, including a single mean response) For h that maps $\Re^k \rightarrow \Re^1$ consider inference for $h(\boldsymbol{\beta})$ (with application to $f(\mathbf{x}, \boldsymbol{\beta})$ for a given set of predictor variables \mathbf{x}). Facts 4) and 5) of Claim 10 suggest that

$$\frac{h(\mathbf{b}_{OLS}) - h(\boldsymbol{\beta})}{\sqrt{MSE}\sqrt{\hat{\mathbf{G}}(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}\hat{\mathbf{G}}'}} \sim t_{n-k}$$

This leads (as in the previous application/example) to testing $H_0:h(\boldsymbol{\beta}) = \#$ using

$$T = \frac{h(\mathbf{b}_{OLS}) - \#}{\sqrt{MSE}\sqrt{\hat{\mathbf{G}}(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}\hat{\mathbf{G}}'}}$$

and a t_{n-k} reference distribution, and to use of the values

$$h(\mathbf{b}_{OLS}) \pm t\sqrt{MSE}\sqrt{\hat{\mathbf{G}}(\hat{\mathbf{D}}'\hat{\mathbf{D}})^{-1}\hat{\mathbf{G}}'}$$

as confidence limits for $h(\boldsymbol{\beta})$.

For a set of predictor variables \mathbf{x} , this can then be applied to $h(\boldsymbol{\beta}) = f(\mathbf{x}, \boldsymbol{\beta})$ to produce inferences for the mean response at \mathbf{x} . That is, with

$$\mathbf{G}_{1 \times k} = \left(\frac{\partial f(\mathbf{x}, \mathbf{b})}{\partial b_j} \Big|_{\mathbf{b}=\boldsymbol{\beta}} \right) \quad \text{and, as expected,} \quad \hat{\mathbf{G}} = \left(\frac{\partial f(\mathbf{x}, \mathbf{b})}{\partial b_j} \Big|_{\mathbf{b}=\mathbf{b}_{OLS}} \right)$$

one may test $H_0: f(\mathbf{x}, \boldsymbol{\beta}) = \#$ using

$$T = \frac{f(\mathbf{x}, \mathbf{b}_{OLS}) - \#}{\sqrt{MSE} \sqrt{\hat{\mathbf{G}} (\hat{\mathbf{D}}' \hat{\mathbf{D}})^{-1} \hat{\mathbf{G}}'}}$$

and a t_{n-k} reference distribution, and use the values

$$f(\mathbf{x}, \mathbf{b}_{OLS}) \pm t \sqrt{MSE} \sqrt{\hat{\mathbf{G}} (\hat{\mathbf{D}}' \hat{\mathbf{D}})^{-1} \hat{\mathbf{G}}'}$$

as confidence limits for $f(\mathbf{x}, \boldsymbol{\beta})$.

Example 13 (Prediction) Suppose that in the future, y^* normal with mean $h(\boldsymbol{\beta})$ and variance $\gamma\sigma^2$ independent of \mathbf{Y} will be observed. (The constant γ is assumed to be known.) Approximate prediction limits for y^* are then

$$h(\mathbf{b}_{OLS}) \pm t \sqrt{MSE} \sqrt{\gamma + \hat{\mathbf{G}} (\hat{\mathbf{D}}' \hat{\mathbf{D}})^{-1} \hat{\mathbf{G}}'}$$

2.3 Inference Methods Based on Large n Theory for Likelihood Ratio Tests/Shape of the Likelihood Function

Exactly as in Section 1.7 for the normal Gauss-Markov linear model, the (normal) “likelihood function” in the (iid errors) non-linear model is

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}) = (2\pi)^{-\frac{n}{2}} \frac{1}{\sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2 \right)$$

Suppressing display of the fact that this is a function of \mathbf{Y} (i.e. this is a “random function” of the parameters $\boldsymbol{\beta}$ and σ^2) it is common to take a natural logarithm and consider the “log-likelihood function”

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2$$

There is general large n theory concerning the use of this kind of random function in testing (and the inversion of tests to get confidence sets) that leads to inference methods alternative to those presented in Section 2.2. That theory is summarized in Section 7.3.2 of the Appendix. Here we make applications of it in the nonlinear model. In the notation of the Appendix, our applications to the nonlinear model will be to the case where $r = k + 1$ and $\boldsymbol{\theta}$ stands for a vector including both $\boldsymbol{\beta}$ and σ^2 .)

Example 14 (Inference for σ^2) Here let θ_1 in the general theory be σ^2 . For any σ^2 , $l(\beta, \sigma^2)$ is maximized as a function of β by \mathbf{b}_{OLS} . So then

$$l^*(\theta_1) = l(\mathbf{b}_{OLS}, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{SSE}{2\sigma^2}$$

and

$$l(\hat{\theta}_{MLE}) = l\left(\mathbf{b}_{OLS}, \frac{SSE}{n}\right) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \frac{SSE}{n} - \frac{n}{2}$$

so that applying display (61) a large sample approximate confidence interval for σ^2 is

$$\left\{ \sigma^2 \mid -\frac{n}{2} \ln \sigma^2 - \frac{SSE}{2\sigma^2} > -\frac{n}{2} \ln \frac{SSE}{n} - \frac{n}{2} - \frac{1}{2} \chi_1^2 \right\}$$

This is an alternative to simply borrowing from the linear model context the approximation

$$\frac{SSE}{\sigma^2} \sim \chi_{n-k}^2$$

and the confidence limits in Example 6.

Example 15 (Inference for the whole vector β) Here we let θ_1 in the general theory be β . For any β , $l(\beta, \sigma^2)$ is maximized as a function of σ^2 by

$$\widehat{\sigma^2}(\beta) = \frac{\sum_{i=1}^n (y_i - f(\mathbf{x}_i, \beta))^2}{n}$$

So then

$$l^*(\theta_1) = l(\beta, \widehat{\sigma^2}(\beta)) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \widehat{\sigma^2}(\beta) - \frac{n}{2}$$

and (since $l(\hat{\theta}_{MLE})$ is as before) applying display (61) a large sample approximate confidence set for β is

$$\begin{aligned} & \left\{ \beta \mid -\frac{n}{2} \ln \widehat{\sigma^2}(\beta) > -\frac{n}{2} \ln \frac{SSE}{n} - \frac{1}{2} \chi_k^2 \right\} \\ &= \left\{ \beta \mid \ln \widehat{\sigma^2}(\beta) < \ln \frac{SSE}{n} + \frac{1}{n} \chi_k^2 \right\} \\ &= \left\{ \beta \mid \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \beta))^2 < SSE \exp\left(\frac{1}{n} \chi_k^2\right) \right\} \end{aligned}$$

Now as it turns out, in the linear model an exact confidence region for β is of the form

$$\left\{ \beta \mid \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \beta))^2 < SSE \left(1 + \frac{k}{n-k} F_{k, n-k}\right) \right\} \quad (19)$$

for $F_{k, n-k}$ the upper α point. So it is common to carry over formula (19) to the nonlinear model setting. When this is done, the region prescribed by formula (19) is called the ‘‘Beale’’ region for β .

Example 16 (*Inference for a single β_j*) What is usually a better alternative to the methods in Example 11 (in terms of holding a nominal coverage probability) is the following application of the general likelihood analysis. θ_1 from the general theory will now be β_j . Let β_j^* stand for the vector β with its j th entry deleted and let $\widehat{\beta}_j^*(\beta_j)$ minimize $\sum_{i=1}^n (y_i - f(\mathbf{x}_i, \beta_j, \beta_j^*))^2$ over choices of β_j^* . Reasoning like that in Example 15 says that a large sample approximate confidence set for β_j is

$$\left\{ \beta_j \mid \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \beta_j, \widehat{\beta}_j^*(\beta_j)))^2 < SSE \exp\left(\frac{1}{n} \chi_1^2\right) \right\}$$

This is the set of β_j for which there is a β_j^* that makes $\sum_{i=1}^n (y_i - f(\mathbf{x}_i, \beta_j, \beta_j^*))^2$ not too much larger than SSE . It is common to again appeal to exact theory for the linear model and replace $\exp(\frac{1}{n} \chi_1^2)$ with something related to an F percentage point, namely

$$\left(1 + \frac{1}{n-k} F_{1, n-k}\right) = \left(1 + \frac{t_{n-k}^2}{n-k}\right)$$

(here the upper α point of the F distribution and the upper $\frac{\alpha}{2}$ point of the t distribution are under discussion).

3 Mixed Models

A second generalization of the linear model is the model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$$

for \mathbf{Y} , \mathbf{X} , β , and ϵ as in the linear model (1), \mathbf{Z} a $n \times q$ model matrix of known constants, and \mathbf{u} a $q \times 1$ random vector. It is common to assume that

$$\mathbb{E} \begin{pmatrix} \mathbf{u} \\ \epsilon \end{pmatrix} = \mathbf{0} \text{ and } \text{Var} \begin{pmatrix} \mathbf{u} \\ \epsilon \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \\ n \times q & n \times n \end{pmatrix}$$

This implies that

$$\mathbb{E}\mathbf{Y} = \mathbf{X}\beta \text{ and } \text{Var}\mathbf{Y} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

The covariance matrix $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ is typically a function of several variances (called “variance components”). In simple standard models with balanced data, it is often a patterned matrix. In this model, some objects of interest are the entries of β (the so-called “fixed effects”), the variance components, and sometimes (prediction of) the entries of \mathbf{u} (the “random effects”).

3.1 Maximum Likelihood in Mixed Models

We will acknowledge that the covariance matrix $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ is typically a function of several (say p) variances by writing $\mathbf{V}(\boldsymbol{\sigma}^2)$ (thinking that $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)'$). The normal likelihood function here is

$$\begin{aligned} L(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\sigma}^2) &= f(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\sigma}^2)) \\ &= (2\pi)^{-\frac{n}{2}} |\det \mathbf{V}(\boldsymbol{\sigma}^2)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}(\boldsymbol{\sigma}^2)^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right) \end{aligned}$$

Notice that for fixed $\boldsymbol{\sigma}^2$ (and therefore fixed $\mathbf{V}(\boldsymbol{\sigma}^2)$), to maximize this as a function of $\mathbf{X}\boldsymbol{\beta}$, one wants to use

$$\widehat{\mathbf{X}\boldsymbol{\beta}}(\boldsymbol{\sigma}^2) = \widehat{\mathbf{Y}}^*(\boldsymbol{\sigma}^2) = \text{the generalized least squares estimate of } \mathbf{X}\boldsymbol{\beta} \text{ based on } \mathbf{V}(\boldsymbol{\sigma}^2)$$

This is (after a bit of algebra beginning with the generalized least squares formula (5))

$$\widehat{\mathbf{Y}}^*(\boldsymbol{\sigma}^2) = \mathbf{X} \left(\mathbf{X}' \mathbf{V}(\boldsymbol{\sigma}^2)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{V}(\boldsymbol{\sigma}^2)^{-1} \mathbf{Y}$$

Plugging this into the likelihood, one produces a profile likelihood for the vector $\boldsymbol{\sigma}^2$,

$$\begin{aligned} L^*(\boldsymbol{\sigma}^2) &= L(\widehat{\mathbf{Y}}^*(\boldsymbol{\sigma}^2), \boldsymbol{\sigma}^2) \\ &= (2\pi)^{-\frac{n}{2}} |\det \mathbf{V}(\boldsymbol{\sigma}^2)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\mathbf{Y} - \widehat{\mathbf{Y}}^*(\boldsymbol{\sigma}^2) \right)' \mathbf{V}(\boldsymbol{\sigma}^2)^{-1} \left(\mathbf{Y} - \widehat{\mathbf{Y}}^*(\boldsymbol{\sigma}^2) \right)\right) \end{aligned}$$

At least in theory, this can be maximized to find MLE's for the variance components. And with $\widehat{\boldsymbol{\sigma}}_{\text{ML}}^2$ a maximizer of $L^*(\boldsymbol{\sigma}^2)$, the maximum likelihood estimate of the whole set of parameters is $\left(\mathbf{X}\boldsymbol{\beta}(\widehat{\boldsymbol{\sigma}}_{\text{ML}}^2), \widehat{\boldsymbol{\sigma}}_{\text{ML}}^2 \right)$.

While maximum likelihood is supported by large sample theory, the folklore is that in small samples, it tends to underestimate variance components. This is generally attributed to “a failure to account for a reduction in degrees of freedom associated with estimation of the mean vector.” *Restricted Maximum Likelihood* or “REML” estimation is a modification of maximum likelihood that for the estimation of variance components essentially replaces use of a likelihood based on \mathbf{Y} with one based on a vector of residuals that is known to have mean $\mathbf{0}$, and therefore requires no estimation of a mean vector.

More precisely, suppose that \mathbf{B} is $m \times n$ of rank $m = n - \text{rank}(\mathbf{X})$ and $\mathbf{B}\mathbf{X} = \mathbf{0}$. Define

$$\mathbf{r} = \mathbf{B}\mathbf{Y}$$

A REML estimate of $\boldsymbol{\sigma}^2$ is a maximizer of a likelihood based on \mathbf{r} ,

$$L_{\mathbf{r}}(\boldsymbol{\sigma}^2) = f(\mathbf{r}|\boldsymbol{\sigma}^2) = (2\pi)^{-\frac{m}{2}} |\det \mathbf{B}\mathbf{V}(\boldsymbol{\sigma}^2) \mathbf{B}'|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{r}' (\mathbf{B}\mathbf{V}(\boldsymbol{\sigma}^2) \mathbf{B}')^{-1} \mathbf{r}\right)$$

Typically the entries of a REML estimate of $\boldsymbol{\sigma}^2$ are larger than those of a maximum likelihood estimate based on \mathbf{Y} . And (happily) every $m \times n$ matrix \mathbf{B} of rank $m = n - \text{rank}(\mathbf{X})$ with $\mathbf{B}\mathbf{X} = \mathbf{0}$ leads to the same (REML) estimate.

3.2 Estimation of an Estimable Vector of Parametric Functions $\mathbf{C}\boldsymbol{\beta}$

Estimability of a linear combination of the entries of $\boldsymbol{\beta}$ means the same thing here as it always does (estimability has only to do with mean structure, not variance-covariance structure). For \mathbf{C} an $l \times k$ matrix with $\mathbf{C} = \mathbf{A}\mathbf{X}$ for some $l \times n$ matrix \mathbf{A} , the BLUE of $\mathbf{C}\boldsymbol{\beta}$ is

$$\widehat{\mathbf{C}\boldsymbol{\beta}} = \mathbf{A}\widehat{\mathbf{Y}}^*(\boldsymbol{\sigma}^2)$$

This has covariance matrix

$$\text{Var}\widehat{\mathbf{C}\boldsymbol{\beta}} = \mathbf{C} \left(\mathbf{X}'\mathbf{V}(\boldsymbol{\sigma}^2)^{-1}\mathbf{X} \right)^{-} \mathbf{C}'$$

The BLUE of $\mathbf{C}\boldsymbol{\beta}$ depends on the variance components, and is typically unavailable. If one estimates $\boldsymbol{\sigma}^2$ with $\widehat{\boldsymbol{\sigma}}^2$, say via REML or maximum likelihood, then one may estimate $\mathbf{V}(\boldsymbol{\sigma}^2)$ as $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\sigma}}^2) = \mathbf{V}(\widehat{\boldsymbol{\sigma}}^2)$, which makes available the approximation to the BLUE

$$\widehat{\widehat{\mathbf{C}\boldsymbol{\beta}}} = \mathbf{A}\widehat{\mathbf{Y}}^*(\widehat{\boldsymbol{\sigma}}^2) \tag{20}$$

It is then perhaps possible to estimate the variance-covariance matrix of the approximate BLUE (20) as

$$\widehat{\widehat{\text{Var}\mathbf{C}\boldsymbol{\beta}}} = \mathbf{C} \left(\mathbf{X}'\mathbf{V}(\widehat{\boldsymbol{\sigma}}^2)^{-1}\mathbf{X} \right)^{-} \mathbf{C}'$$

(My intuition is that in small to moderate samples this will tend to produce standard errors that are better than nothing, but probably tend to be too small.)

3.3 Best Linear Unbiased Prediction and Related Inference in the Mixed Model

We now consider problems of prediction related to the random effects contained in \mathbf{u} . Under the MVN assumption

$$\text{E}[\mathbf{u}|\mathbf{Y}] = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{21}$$

which, obviously, depends on the fixed effect vector $\boldsymbol{\beta}$. (For what it is worth,

$$\text{Var} \begin{pmatrix} \mathbf{u} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{G}\mathbf{Z}' \\ \mathbf{Z}\mathbf{G} & \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \end{pmatrix}$$

and the $\mathbf{G}\mathbf{Z}'$ appearing in (21) is the covariance between \mathbf{u} and \mathbf{Y} .) Something that is close to the conditional mean (21), but that does not depend on fixed effects is

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \widehat{\mathbf{Y}}^*) \tag{22}$$

where $\widehat{\mathbf{Y}}^*$ is the generalized least squares (best linear unbiased) estimate of the mean of \mathbf{Y} ,

$$\widehat{\mathbf{Y}}^* = \mathbf{X} (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

The predictor (22) turns out to be the Best Linear Unbiased Predictor of \mathbf{u} , and if we temporarily abbreviate

$$\mathbf{B} = \mathbf{X} (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1} \text{ and } \mathbf{P} = \mathbf{V}^{-1} (\mathbf{I} - \mathbf{B}) \quad (23)$$

this can be written as

$$\widehat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1} (\mathbf{Y} - \mathbf{B}\mathbf{Y}) = \mathbf{GZ}'\mathbf{V}^{-1} (\mathbf{I} - \mathbf{B}) \mathbf{Y} = \mathbf{GZ}'\mathbf{P}\mathbf{Y} \quad (24)$$

We consider here predictions based on $\widehat{\mathbf{u}}$ and the problems of quoting appropriate “precision” measures for them.

To begin with the \mathbf{u} vector itself, $\widehat{\mathbf{u}}$ is an obvious approximation and a precision of prediction should be related to the variability in the difference

$$\mathbf{u} - \widehat{\mathbf{u}} = \mathbf{u} - \mathbf{GZ}'\mathbf{P}\mathbf{Y} = \mathbf{u} - \mathbf{GZ}'\mathbf{P} (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon})$$

This random vector has mean $\mathbf{0}$ and covariance matrix

$$\text{Var}(\mathbf{u} - \widehat{\mathbf{u}}) = (\mathbf{I} - \mathbf{GZ}'\mathbf{PZ}) \mathbf{G} (\mathbf{I} - \mathbf{GZ}'\mathbf{PZ})' + \mathbf{GZ}'\mathbf{P}\mathbf{R}\mathbf{PZ}\mathbf{G} = \mathbf{G} - \mathbf{GZ}'\mathbf{PZ}\mathbf{G} \quad (25)$$

(This last equality is not obvious to me, but is what McCulloch and Searle promise on page 170 of their book.)

Now $\widehat{\mathbf{u}}$ in (24) is not available unless knows the covariance matrices \mathbf{G} and \mathbf{V} . If one has estimated variance components and hence has estimates of \mathbf{G} and \mathbf{V} (and for that matter, \mathbf{R} , \mathbf{B} , and \mathbf{P}) the approximate BLUP

$$\widehat{\widehat{\mathbf{u}}} = \widehat{\mathbf{G}}\mathbf{Z}'\widehat{\mathbf{P}}\mathbf{Y}$$

may be used. A way of making a crude approximation to a measure of precision of the approximate BLUP (as a predictor \mathbf{u}) is to plug estimates into the relationship (25) to produce

$$\text{Var}(\widehat{\widehat{\mathbf{u}}}) = \widehat{\mathbf{G}} - \widehat{\mathbf{G}}\mathbf{Z}'\widehat{\mathbf{P}}\mathbf{Z}\widehat{\mathbf{G}}$$

Consider now the prediction of a quantity

$$l = \mathbf{c}'\boldsymbol{\beta} + \mathbf{s}'\mathbf{u}$$

for an estimable $\mathbf{c}'\boldsymbol{\beta}$ (estimability has nothing to do with the covariance structure of \mathbf{Y} and so “estimability” means here what it always does). As it turns out, if $\mathbf{c}' = \mathbf{a}'\mathbf{X}$, the BLUP of l is

$$\widehat{l} = \mathbf{a}'\widehat{\mathbf{Y}}^* + \mathbf{s}'\widehat{\mathbf{u}} = \mathbf{a}'\mathbf{B}\mathbf{Y} + \mathbf{s}'\mathbf{GZ}'\mathbf{P}\mathbf{Y} = (\mathbf{a}'\mathbf{B} + \mathbf{s}'\mathbf{GZ}'\mathbf{P}) \mathbf{Y} \quad (26)$$

To quantify the precision of this as a predictor of l we must consider the random variable

$$\widehat{l} - l$$

The variance of this is the unpleasant (but not impossible) quantity

$$\text{Var}(\widehat{l} - l) = \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{a} + \mathbf{s}'\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}\mathbf{s} - 2\mathbf{a}'\mathbf{B}\mathbf{Z}\mathbf{G}\mathbf{s} \quad (27)$$

(This variance is from page 256 of McCulloch and Searle.)

Now \widehat{l} of (26) is not available unless one knows covariance matrices. But with estimates of variance components and corresponding matrices, what is available as an approximation to \widehat{l} is

$$\widehat{\widehat{l}} = \left(\mathbf{a}'\widehat{\mathbf{B}} + \mathbf{s}'\widehat{\mathbf{G}}\mathbf{Z}'\widehat{\mathbf{P}} \right) \mathbf{Y}$$

A way of making a crude approximation to a measure of precision of the approximate BLUP (as a predictor of l) is to plug estimates into the relationship (27) to produce

$$\widehat{\text{Var}}(\widehat{\widehat{l}} - l) = \mathbf{a}'\mathbf{X}(\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{a} + \mathbf{s}'\widehat{\mathbf{G}}\mathbf{Z}'\widehat{\mathbf{P}}\mathbf{Z}\widehat{\mathbf{G}}\mathbf{s} - 2\mathbf{a}'\widehat{\mathbf{B}}\mathbf{Z}\widehat{\mathbf{G}}\mathbf{s}$$

3.4 Confidence Intervals and Tests for Variance Components

Section 2.4.3 of Pinheiro and Bates indicates that the following is the state of the art for interval estimation of the entries of $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$. (This is an application of the general theory outlined in Appendix 7.3.1) Let

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_p) = (\log \sigma_1^2, \log \sigma_2^2, \dots, \log \sigma_p^2)$$

be the vector of log variance components. (This one-to-one transformation is used because for finite samples, likelihoods for $\boldsymbol{\gamma}$ tend to be “better behaved/more nearly quadratic” than likelihoods for $\boldsymbol{\sigma}^2$.) Let $\widehat{\boldsymbol{\sigma}}^2$ be a ML or REML estimate of $\boldsymbol{\sigma}^2$, and define the corresponding vector of estimated log variance components

$$\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \dots, \widehat{\gamma}_p) = \left(\log \widehat{\sigma}_1^2, \log \widehat{\sigma}_2^2, \dots, \log \widehat{\sigma}_p^2 \right)$$

Consider a full rank version of the fixed effects part of the mixed model and let

$$l(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \log L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$$

be the loglikelihood and

$$l_{\mathbf{r}}(\boldsymbol{\sigma}^2) = \log L_{\mathbf{r}}(\boldsymbol{\sigma}^2)$$

be the restricted loglikelihood. Define

$$l^*(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, (\exp \gamma_1, \exp \gamma_2, \dots, \exp \gamma_p))$$

and

$$l_{\mathbf{r}}^*(\boldsymbol{\gamma}) = l_{\mathbf{r}}(\exp \gamma_1, \exp \gamma_2, \dots, \exp \gamma_p)$$

These are the loglikelihood and restricted loglikelihood for the $\boldsymbol{\gamma}$ parameterization.

To first lay out confidence limits based on maximum likelihood, define a matrix of second partials

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

$\begin{matrix} p \times p & p \times k \\ k \times p & k \times k \end{matrix}$

where

$$\mathbf{M}_{11} = \left(\frac{\partial^2 l^*(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \gamma_i \partial \gamma_j} \Big|_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})_{\text{ML}}} \right), \quad \mathbf{M}_{22} = \left(\frac{\partial^2 l^*(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \beta_i \partial \beta_j} \Big|_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})_{\text{ML}}} \right)$$

$$\text{and } \mathbf{M}_{12} = \left(\frac{\partial^2 l^*(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \gamma_i \partial \beta_j} \Big|_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})_{\text{ML}}} \right) = \mathbf{M}'_{21}$$

Then the matrix

$$\mathbf{Q} = -\mathbf{M}^{-1}$$

functions as an estimated variance-covariance matrix for $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})_{\text{ML}}$, which is approximately normal in large samples. So approximate confidence limits for the i th entry of $\boldsymbol{\gamma}$ are

$$\hat{\gamma}_i \pm z\sqrt{q_{ii}} \quad (28)$$

(where q_{ii} is the i th diagonal element of \mathbf{Q}). Exponentiating, approximate confidence limits for σ_i^2 based on maximum likelihood are

$$\left(\hat{\sigma}_i^2 \exp(-z\sqrt{q_{ii}}), \hat{\sigma}_i^2 \exp(z\sqrt{q_{ii}}) \right) \quad (29)$$

A similar story can be told based on REML estimates. If one defines

$$\mathbf{M}_{\mathbf{r}} = \left(\frac{\partial^2 l_{\mathbf{r}}^*(\boldsymbol{\gamma})}{\partial \gamma_i \partial \gamma_j} \Big|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}_{\text{REML}}} \right)$$

and takes

$$\mathbf{Q}_{\mathbf{r}} = -\mathbf{M}_{\mathbf{r}}^{-1}$$

the formula (29) may be used, where where q_{ii} is the i th diagonal element of $\mathbf{Q}_{\mathbf{r}}$.

Rather than working through approximate normality of $\hat{\boldsymbol{\gamma}}$ and quadratic shape for $l^*(\boldsymbol{\beta}, \boldsymbol{\gamma})$ or $l_{\mathbf{r}}^*(\boldsymbol{\gamma})$ near its maximum, it is possible to work directly with $\hat{\boldsymbol{\sigma}}^2$ and get a different formula parallel to (28) for intervals centered at $\hat{\sigma}_i^2$. Although the two approaches are equivalent in the limit, in finite samples,

method (29) does a better job of holding its nominal confidence level, so it is the one commonly used.

As discussed in Section 2.4.1 of Pinheiro and Bates, it is common to want to compare two random effects structures for a given \mathbf{Y} and fixed effects structure, where the second is more general than the first (the models are “nested”). If λ_1 and λ_2 are the corresponding maximized log likelihoods (or maximized restricted log likelihoods), the test statistic

$$2(\lambda_2 - \lambda_1)$$

can be used to test the hypothesis that the smaller/simpler/less general model is adequate. If there are p_1 variance components in the first model and p_2 in the second, large sample theory (see Appendix 7.3.2) suggests a $\chi_{p_2-p_1}^2$ reference distribution for testing the adequacy of the smaller model.

3.5 Linear Combinations of Mean Squares and the Cochran-Satterthwaite Approximation

It is common in simple mixed models to consider linear combinations of certain independent “mean squares” as estimates of variances. It is thus useful to have some theory for such linear combinations.

Suppose that

$$MS_1, MS_2, \dots, MS_l$$

are independent random variables and for constants df_1, df_2, \dots, df_l each

$$\frac{(df_i) MS_i}{EMS_i} \sim \chi_{df_i}^2 \quad (30)$$

Consider the random variable

$$S^2 = a_1 MS_1 + a_2 MS_2 + \dots + a_l MS_l \quad (31)$$

Clearly,

$$ES^2 = a_1 EMS_1 + a_2 EMS_2 + \dots + a_l EMS_l$$

Further, it is possible to show that

$$\text{Var}S^2 = 2 \sum a_i^2 \frac{(EMS_i)^2}{df_i}$$

and one might estimate this by plugging in estimates of the quantities $(EMS_i)^2$. The most obvious way of estimating $(EMS_i)^2$ is using $(MS_i)^2$, which leads to the estimator

$$\widehat{\text{Var}S^2} = 2 \sum a_i^2 \frac{(MS_i)^2}{df_i}$$

A somewhat more refined (but less conservative) estimator follows from the observation that $E(MS_i)^2 = \frac{df_i+2}{df_i} (EMS_i)^2$, which suggests

$$\widehat{\text{Var}S^2}^* = 2 \sum a_i^2 \frac{(MS_i)^2}{df_i + 2}$$

Clearly, either $\sqrt{\widehat{\text{Var}S^2}}$ or $\sqrt{\widehat{\text{Var}S^2}^*}$ could function as a standard error for S^2 .

Beyond producing a standard error for S^2 , it is common to make a distributional approximation for S^2 . The famous Cochran-Satterthwaite approximation is most often used. This treats S^2 as approximately a multiple of a chi-square random variable. That is, while S^2 does not exactly have any standard distribution, for an appropriate ν one might hope that

$$\frac{\nu S^2}{ES^2} \sim \chi_\nu^2 \tag{32}$$

Notice that the variable $\nu S^2/ES^2$ has mean ν . Setting $\text{Var}(\nu S^2/ES^2) = 2\nu$ (the variance of a χ_ν^2 random variable) and solving for ν produces

$$\nu = \frac{(ES^2)^2}{\sum \frac{(a_i EMS_i)^2}{df_i}} \tag{33}$$

and approximation (32) with degrees of freedom (33) is the Cochran-Satterthwaite approximation.

This approximation leads to (unusable) confidence limits for ES^2 of the form

$$\left(\frac{\nu S^2}{\text{upper } \frac{\alpha}{2} \text{ point of } \chi_\nu^2}, \frac{\nu S^2}{\text{lower } \frac{\alpha}{2} \text{ point of } \chi_\nu^2} \right) \tag{34}$$

These are unusable because ν depends on the unknown expected means squares. But the degrees of freedom parameter may be estimated as

$$\hat{\nu} = \frac{(S^2)^2}{\sum \frac{(a_i MS_i)^2}{df_i}} \tag{35}$$

and plugged into the limits (34) to get even more approximate (but usable) confidence limits for ES^2 of the form

$$\left(\frac{\hat{\nu} S^2}{\text{upper } \frac{\alpha}{2} \text{ point of } \chi_{\hat{\nu}}^2}, \frac{\hat{\nu} S^2}{\text{lower } \frac{\alpha}{2} \text{ point of } \chi_{\hat{\nu}}^2} \right) \tag{36}$$

(Approximation (32) is also used to provide approximate t intervals for various kinds of estimable functions in particular mixed models.)

3.6 Mixed Models and the Analysis of Balanced Two-Factor Nested Data

There are a variety of standard (balanced data) analyses of classical ‘‘ANOVA’’ that are based on particular mixed models. Koehler’s mixed models notes treat several of these. One is the two-factor nested data analysis of Sections 28.1-28.5 of Neter et al. that we proceed to consider.

We assume that Factor A has a levels, within each one of which Factor B has b levels, and that there are c observations from each of the $a \times b$ instances

of a level of A and a level of B within that level of A. Then a possible mixed effects model is that for $i = 1, 2, \dots, a$, $j = 1, 2, \dots, b$, and $k = 1, 2, \dots, c$

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$$

where μ is an unknown constant (the model's only fixed effect) and the a random effects α_i are iid $N(0, \sigma_\alpha^2)$, independent of the ab random effects β_{ij} that are iid $N(0, \sigma_\beta^2)$, and the set of α_i and β_{ij} is independent of the set of ϵ_{ijk} that are iid $N(0, \sigma^2)$. This can be written in the basic mixed model form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$, where

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \boldsymbol{\beta} = \mu, \text{ and } \mathbf{u} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_a \\ \beta_{11} \\ \vdots \\ \beta_{1b} \\ \vdots \\ \beta_{a1} \\ \vdots \\ \beta_{ab} \end{pmatrix}$$

For the $abc \times 1$ vector \mathbf{Y} with the y_{ijk} written down in lexicographical (dictionary) order (as regards the triple numerical subscripts) the covariance matrix here is

$$\mathbf{V} = \sigma_\alpha^2 \mathbf{I}_{a \times a} \otimes \mathbf{J}_{bc \times bc} + \sigma_\beta^2 \mathbf{I}_{ab \times ab} \otimes \mathbf{J}_{c \times c} + \sigma^2 \mathbf{I}_{abc \times abc}$$

For purposes of defining mean squares of interest (and computing them) we may think of doing linear model (all fixed effects) computations under the sum restrictions

$$\sum \alpha_i = 0 \text{ and } \sum_j \beta_{ij} = 0 \quad \forall i$$

Then for a full rank linear model with $(a - 1)$ parameters α_i and $a(b - 1)$ parameters β_{ij} define

$$\begin{aligned} SSA &= R(\alpha's|\mu) \\ SSB(A) &= R(\beta's|\mu, \alpha's) = R(\beta's|\mu) \\ SSE &= \mathbf{e}'\mathbf{e} \end{aligned}$$

and associate with these sums of squares respective degrees of freedom $(a - 1)$,

$a(b-1)$, and $ab(c-1)$. The corresponding mean squares

$$\begin{aligned} MSA &= \frac{SSA}{a-1} \\ MSB(A) &= \frac{SSB(A)}{a(b-1)} \\ MSE &= \frac{SSE}{ab(c-1)} \end{aligned}$$

are independent, and after normalization have chi-squared distributions as in (30).

The expected values of the mean squares are

$$\begin{aligned} EMSA &= \sigma^2 + c\sigma_\beta^2 + bc\sigma_\alpha^2 \\ EMSB(A) &= \sigma^2 + c\sigma_\beta^2 \\ EMSE &= \sigma^2 \end{aligned}$$

and these suggest estimators of variance components

$$\begin{aligned} \hat{\sigma}_\alpha^2 &= \frac{1}{bc}(MSA - MSB(A)) = \frac{1}{bc}MSA - \frac{1}{bc}MSB(A) \\ \hat{\sigma}_\beta^2 &= \frac{1}{c}(MSB(A) - MSE) = \frac{1}{c}MSB(A) - \frac{1}{c}MSE \\ \hat{\sigma}^2 &= MSE \end{aligned}$$

Further, since these estimators of variance components are of the form (31), the Cochran-Satterthwaite material of Section 3.5 (in particular formula (36)) can be used to make approximate confidence limits for the variance components. (The interval for σ^2 is exact, as always.)

Regarding inference for the fixed effect μ , the BLUE is

$$\bar{y}_{...} = \mu + \bar{\alpha}_{..} + \bar{\beta}_{..} + \bar{\epsilon}_{...}$$

It is then easy to see that

$$\text{Var}\bar{y}_{...} = \frac{\sigma_\alpha^2}{a} + \frac{\sigma_\beta^2}{ab} + \frac{\sigma^2}{abc} = \frac{1}{abc}EMSA$$

This, in turn, correctly suggests that confidence limits for μ are

$$\bar{y}_{...} \pm t\sqrt{\frac{MSA}{abc}}$$

for t a percentage point of the $t_{(a-1)}$ distribution.

3.7 Mixed Models and the Analysis of Unreplicated Balanced One-Way Data in Complete Random Blocks

Koehler's Example 10.1 and Neter et al.'s Sections 27.12 and 29.2 consider the analysis of a set of single observations from $a \times b$ combinations of a level of Factor A and a level of a random ("blocking") Factor B. (In Section 29.2 of Neter et al., the second factor is "subjects" and the discussion is in terms of "repeated measures" on subjects under a different treatments.) That is, for $i = 1, 2, \dots, a$ and $j = 1, 2, \dots, b$ in this section we suppose that

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (37)$$

where the fixed effects are μ and the α_i , the b values β_j are iid $N(0, \sigma_\beta^2)$, and the set of β_j is independent of the set of ϵ_{ij} that are iid $N(0, \sigma^2)$. With the $ab \times 1$ vector of observations ordered as

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{a1} \\ y_{12} \\ \vdots \\ y_{a2} \\ \vdots \\ y_{1b} \\ \vdots \\ y_{ab} \end{pmatrix}$$

the model can be written in the standard mixed model form as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1} & \mathbf{I} \\ a \times 1 & a \times a \\ \mathbf{1} & \mathbf{I} \\ a \times 1 & a \times a \\ \vdots & \vdots \\ \mathbf{1} & \mathbf{I} \\ a \times 1 & a \times a \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} + \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ a \times 1 & a \times 1 & a \times 1 & \cdots & a \times 1 \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ a \times 1 & a \times 1 & a \times 1 & \cdots & a \times 1 \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ a \times 1 & a \times 1 & a \times 1 & \cdots & a \times 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \\ a \times 1 & a \times 1 & a \times 1 & \cdots & a \times 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_b \end{pmatrix} + \boldsymbol{\epsilon}$$

and the variance-covariance matrix for \mathbf{Y} is easily be seen to be

$$\mathbf{V} = \sigma_\beta^2 \mathbf{I}_{b \times b} \otimes \mathbf{J}_{a \times a} + \sigma^2 \mathbf{I}_{ab \times ab}$$

For purposes of defining mean squares of interest (and computing them) we may think of doing linear model (all fixed effects) computations under the sum restrictions

$$\sum \alpha_i = 0 \text{ and } \sum \beta_j = 0$$

Then for a full rank linear model with $(a - 1)$ parameters α_i and $(b - 1)$ parameters β_i define

$$\begin{aligned} SSA &= R(\alpha's|\mu) \\ SSB &= R(\beta's|\mu, \alpha's) = R(\beta's|\mu) \\ SSE &= \mathbf{e}'\mathbf{e} \end{aligned}$$

and associate with these sums of squares respective degrees of freedom $(a - 1)$, $(b - 1)$, and $(a - 1)(b - 1)$. The mean squares

$$\begin{aligned} MSB &= \frac{SSB}{b - 1} \\ MSE &= \frac{SSE}{(a - 1)(b - 1)} \end{aligned}$$

are independent, and after normalization have chi-squared distributions as in (30).

The expected values of the two mean squares of most interest are

$$\begin{aligned} EMSB &= \sigma^2 + a\sigma_\beta^2 \\ EMSE &= \sigma^2 \end{aligned}$$

and these suggest estimators of variance components

$$\begin{aligned} \hat{\sigma}_\beta^2 &= \frac{1}{a}(MSB - MSE) = \frac{1}{a}MSB - \frac{1}{a}MSE \\ \hat{\sigma}^2 &= MSE \end{aligned}$$

Note that just as in Section 3.6, the Cochran-Satterthwaite material of Section 3.5 (in particular formula (36)) can be used to make approximate confidence limits for the variance components. (The interval for σ^2 is exact, as always.)

What are usually of more interest in this model are intervals for estimable functions of the fixed effects. In this direction, consider first the estimable function

$$\mu + \alpha_i$$

It is the case that the BLUE of this is

$$\bar{y}_{i.} = \mu + \alpha_i + \bar{\beta}_{.} + \bar{\epsilon}_i.$$

This has

$$\text{Var}\bar{y}_{i.} = \frac{\sigma_\beta^2}{b} + \frac{\sigma^2}{b}$$

A possible estimated variance for the BLUE of $\mu + \alpha_i$ is then

$$S_{\bar{y}_{i.}}^2 = \frac{1}{b}(\hat{\sigma}_\beta^2 + \hat{\sigma}^2) = \frac{1}{ab(b-1)}(SSB + SSE) = \frac{1}{ab}MSB + \frac{(a-1)}{ab}MSE$$

which is clearly of form (31). Applying the Cochran-Satterthwaite approximation, unusable (approximate) confidence limits for $\mu + \alpha_i$ are then easily seen to be

$$\bar{y}_i. \pm t \sqrt{S_{\bar{y}_i.}^2}$$

where t is a t_ν percentage point. And further approximation suggests that usable approximate confidence limits are

$$\bar{y}_i. \pm \hat{t} \sqrt{S_{\bar{y}_i.}^2}$$

for \hat{t} a $t_{\hat{\nu}}$ percentage point, with $\hat{\nu}$ as given in general in (35) and here specifically as

$$\hat{\nu} = \frac{\left(S_{\bar{y}_i.}^2\right)^2}{\frac{\left(\frac{1}{ab}MSE\right)^2}{(b-1)} + \frac{\left(\frac{a-1}{ab}MSE\right)^2}{(a-1)(b-1)}}$$

What is perhaps initially slightly surprising is that *some* estimable functions have obvious exact confidence limits (ones that don't require application of the Cochran-Satterthwaite approximation). Consider, for example, $\alpha_i - \alpha_{i'}$. The BLUE of this is

$$\begin{aligned} \bar{y}_i. - \bar{y}_{i'}. &= (\mu + \alpha_i + \bar{\beta}_. + \bar{\epsilon}_i.) - (\mu + \alpha_{i'} + \bar{\beta}_. + \bar{\epsilon}_{i'}.) \\ &= (\alpha_i - \alpha_{i'}) + (\bar{\epsilon}_i. - \bar{\epsilon}_{i'}.) \end{aligned}$$

This has variance

$$\text{Var}(\bar{y}_i. - \bar{y}_{i'}.) = \frac{2\sigma^2}{b}$$

which can be estimated as a multiple of MSE alone. So exact confidence limits for $\alpha_i - \alpha_{i'}$ are

$$\bar{y}_i. - \bar{y}_{i'}. \pm t \sqrt{\frac{2MSE}{b}} \tag{38}$$

for t a percentage point of the $t_{(a-1)(b-1)}$ distribution.

It is perfectly possible to restate the model $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ in the form $y_{ij} = \mu_i + \beta_j + \epsilon_{ij}$, or in the case that the a treatments themselves have their own factorial structure, even replace the μ_i with something like $\alpha_l + \gamma_k + \alpha\gamma_{lk}$ (producing a model with a factorial treatment structure in random blocks). This line of thinking (represented, for example, in Section 29.3 of Neter et al.) suggests that linear combinations of the μ_i beyond $\mu_i - \mu_{i'} = \alpha_i - \alpha_{i'}$ can be of practical interest in the present context. Many of these will be of the form

$$\sum c_i \mu_i \text{ for constants } c_i \text{ with } \sum c_i = 0 \tag{39}$$

The BLUE of quantity (39) is

$$\sum c_i \bar{y}_i. = \sum c_i \mu_i + \sum c_i \bar{\epsilon}_i.$$

which has variance

$$\text{Var} \sum c_i \bar{y}_i = \frac{\sigma^2}{b} \sum c_i^2$$

(Notice that as for estimating $\alpha_i - \alpha_{i'}$, the condition $\sum c_i = 0$ guarantees that the random block effects do not enter this analysis.) So confidence limits for (39) generalizing the ones in (38) are

$$\sum c_i \bar{y}_i \pm t \sqrt{\frac{\sum c_i^2}{b}} \sqrt{MSE}$$

for t a percentage point of the $t_{(a-1)(b-1)}$ distribution.

3.8 Mixed Models and the Analysis of Balanced Split-Plot/Repeated Measures Data

Something fundamentally different from the imposition of a factorial structure on the μ_i of model (37) is represented by the “split plot model”

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_{jk} + \epsilon_{ijk} \quad (40)$$

for

y_{ijk} = the k th observation at the i th level of A (the split plot treatment)
and the j th level of B (the whole plot treatment)

$i = 1, 2, \dots, a$, $j = 1, 2, \dots, b$, $k = 1, 2, \dots, n$, where the effects μ, α_i, β_j , and $\alpha\beta_{ij}$ are fixed (and, say, subject to the sum restriction so that the fixed effects part of the model is full rank) and the γ_{jk} are iid $N(0, \sigma_\gamma^2)$ independent of the iid $N(0, \sigma^2)$ errors ϵ_{ijk} . If we write

$$\mathbf{y}_{jk} = \begin{pmatrix} y_{1jk} \\ y_{2jk} \\ \vdots \\ y_{ajk} \end{pmatrix}$$

and let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_{11} \\ \vdots \\ \mathbf{y}_{1n} \\ \mathbf{y}_{21} \\ \vdots \\ \mathbf{y}_{2n} \\ \vdots \\ \mathbf{y}_{b1} \\ \vdots \\ \mathbf{y}_{bn} \end{pmatrix}$$

it is the case that

$$\text{Var}\mathbf{Y} = \sigma_\gamma^2 \mathbf{I}_{bn \times bn} \otimes \mathbf{J}_{a \times a} + \sigma^2 \mathbf{I}_{abn \times abn}$$

For purposes of defining mean squares of interest (and computing them) we may think of doing linear model (all fixed effects) computations under the sum restrictions

$$\sum \alpha_i = 0, \sum \beta_j = 0, \sum_i \alpha\beta_{ij} = 0 \quad \forall j, \sum_j \alpha\beta_{ij} = 0 \quad \forall i, \sum_k \gamma_{jk} = 0 \quad \forall j$$

For a full rank linear model with $(a - 1)$ parameters α_i , $(b - 1)$ parameters β_i , $(a - 1)(b - 1)$ parameters $\alpha\beta_{ij}$, and $b(n - 1)$ parameters γ_{jk} define

$$\begin{aligned} SSA &= R(\alpha's|\mu) \\ SSA &= R(\beta's|\mu) \\ SSAB &= R(\alpha\beta's|\mu) \\ SSC(B) &= R(\gamma's|\mu, \beta's) \\ SSE &= \mathbf{e}'\mathbf{e} \end{aligned}$$

and associate with these sums of squares respective degrees of freedom $(a - 1)$, $(b - 1)$, $(a - 1)(b - 1)$, $b(n - 1)$, and $(a - 1)b(n - 1)$. The mean squares

$$\begin{aligned} MSC(B) &= \frac{SSC(B)}{b(n - 1)} \\ MSE &= \frac{SSE}{(a - 1)b(n - 1)} \end{aligned}$$

are independent, and after normalization have chi-squared distributions as in (30). The expected values are

$$\begin{aligned} EMSC(B) &= \sigma^2 + a\sigma_\gamma^2 \\ EMSE &= \sigma^2 \end{aligned}$$

These suggest estimators of variance components

$$\begin{aligned} \hat{\sigma}_\gamma^2 &= \frac{1}{a} (MSC(B) - MSE) \\ \hat{\sigma}^2 &= MSE \end{aligned}$$

and the Cochran-Satterthwaite approximation produces confidence intervals for these.

Regarding inference for the fixed effects, notice that

$$\bar{y}_{i..} - \bar{y}_{i'..} = (\alpha_i - \alpha_{i'}) + (\bar{\epsilon}_{i..} - \bar{\epsilon}_{i'..})$$

so that confidence limits for $\alpha_i - \alpha_{i'}$ are

$$\bar{y}_{i..} - \bar{y}_{i'..} \pm t \sqrt{\frac{2MSE}{bn}} \quad (41)$$

Then note that (for estimating $\beta_j - \beta_{j'}$)

$$\bar{y}_{.j} - \bar{y}_{.j'} = (\beta_j - \beta_{j'}) + (\bar{\gamma}_j - \bar{\gamma}_{j'}) + (\bar{\epsilon}_{.j} - \bar{\epsilon}_{.j'})$$

has variance

$$\text{Var}(\bar{y}_{.j} - \bar{y}_{.j'}) = \frac{2\sigma_\gamma^2}{n} + \frac{2\sigma^2}{an} = \frac{2}{an} \text{EMSC}(B)$$

So confidence limits for $\beta_j - \beta_{j'}$ are

$$\bar{y}_{.j} - \bar{y}_{.j'} \pm t \sqrt{\frac{2\text{MSC}(B)}{an}} \quad (42)$$

Notice from the formulas (41) and (42) that the “error terms” appropriate for comparing (or judging differences in) main effects are different for the two factors. The standard jargon (corresponding to formula (41)) is that the “split plot error” is used in judging the “split plot treatment” and (corresponding to formula (42)) that the “whole plot error” is used in judging the “whole plot treatment.” This is consistent with standard ANOVA analyses where

$$\begin{aligned} H_0 &: \alpha_i = 0 \quad \forall i \text{ is tested using } F = MSA/MSE \\ H_0 &: \beta_j = 0 \quad \forall j \text{ is tested using } F = MSB/MSB(B) \\ H_0 &: \alpha\beta_{ij} = 0 \quad \forall i, j \text{ is tested using } F = MSAB/MSE \end{aligned}$$

The plausibility of the last of these derives from the fact that

$$\begin{aligned} (\bar{y}_{ij} - \bar{y}_{ij'}) - (\bar{y}_{i'j} - \bar{y}_{i'j'}) &= (\alpha\beta_{ij} - \alpha\beta_{ij'}) - (\alpha\beta_{i'j} - \alpha\beta_{i'j'}) + \\ &(\bar{\epsilon}_{ij} - \bar{\epsilon}_{ij'}) - (\bar{\epsilon}_{i'j} - \bar{\epsilon}_{i'j'}) \end{aligned}$$

4 Bootstrap Methods

Much of what has been outlined here (except for the Linear Models material where exact distribution theory exists) has relied on large sample approximations to the distributions of estimators and hopes that when we plug estimates of parameters into formulas for approximate standard deviations we get reliable standard errors. Modern computing has provided an alternative (simulation-based) way of assessing the precision of complicated estimators. This is the “bootstrap” methodology. This section provides a brief introduction.

4.1 Bootstrapping in the iid (Single Sample) Case

Suppose that (possibly vector-valued) random quantities

$$\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$$

are independent, each with marginal distribution F . F here is some (possibly multivariate) distribution. It may be (but need not necessarily be) parameterized by (a possibly vector-valued) θ . (In this case, we might write F_θ .) Suppose that

$$\begin{aligned} T_n &= \text{some function of } \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n \\ G_n^F &= \text{the } (F) \text{ distribution of } T_n \\ H(G_n^F) &= \text{some characteristic of } G_n^F \text{ of interest} \end{aligned}$$

If F is known, evaluation of $H(G_n^F)$ is a probability problem. In simple cases, this may yield to analytical calculations (after the style of Stat 542). When the probability problem is analytically intractable, simulation can often be used. One simulates B samples

$$\mathbf{Y}_1^*, \mathbf{Y}_2^*, \dots, \mathbf{Y}_n^* \tag{43}$$

from F . Each of these is used to compute a simulated value of T_n , say

$$T_n^* \tag{44}$$

The “histogram”/empirical distribution of the B simulated values

$$T_{n1}^*, T_{n2}^*, \dots, T_{nB}^* \tag{45}$$

is used to approximate G_n^F . Finally,

$$H(\text{empirical distribution of } \{T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*\}) \tag{46}$$

is an approximation for $H(G_n^F)$.

If F is unknown (the problem is one of statistics rather than probability) “bootstrap” approximations to the simulation substitute for probability calculations may be made in at least two different ways. First, in parametric contexts, one might estimate θ with $\hat{\theta}$ and then F with $\hat{F} = F_{\hat{\theta}}$. Simulated values (43) are generated not from F , but from $\hat{F} = F_{\hat{\theta}}$, leading to a bootstrapped version of T_n , (44). This can be done B times to produce values (45) and the approximation (46) for $H(G_n^F)$. This way of operating is (not surprisingly) called a “parametric bootstrap.”

Second, in nonparametric contexts, one might estimate F with

$$\hat{F} = \text{the empirical distribution of } \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$$

Then as in the parametric case, simulated values (43) can be generated from \hat{F} . Here, this amounts to random sampling with replacement from $\{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n\}$. A single bootstrap sample (43) then leads to a simulated value of T_n . B such values (45) in turn lead to the approximation (46) for $H(G_n^F)$. Not surprisingly, this methodology is called a “nonparametric bootstrap.”

There are two basic issues of concern in this kind of “bootstrap” approximation for $H(G_n^F)$. There are:

1. The Monte Carlo error in approximating

$$H(G_n^{\widehat{F}})$$

with

$$H(\text{empirical distribution of } T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*)$$

Presumably, this is often handled by making B large so that

$$G_n^{\widehat{F}} \approx \text{empirical distribution of } T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*$$

and relying on some kind of “continuity” of $H(\cdot)$ (so that if the empirical distribution of $T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*$ is nearly $G_n^{\widehat{F}}$, then

$$H(G_n^{\widehat{F}}) \approx H(\text{empirical distribution of } T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*)$$

2. The estimation error in approximating F with \widehat{F} and then approximating

$$G_n^F$$

with

$$G_n^{\widehat{F}}$$

Here we have to rely on an adequately large *original sample size*, n , and hope that

$$G_n^{\widehat{F}} \approx G_n^F \tag{47}$$

so that (again assuming that $H(\cdot)$ is continuous)

$$H(G_n^{\widehat{F}}) \approx H(G_n^F)$$

Notice that only the first of these issues can be handled by the (usually cheap) device of increasing B . One remains at the mercy of the fixed original sample to produce (47).

Several applications/examples now follow.

Example 17 (*Bootstrap standard error for T_n*) *There is, of course, interest in producing standard errors for complicated estimators. That is, one often wants to estimate*

$$H(G_n^F) = \sqrt{\text{Var}_F T_n} \tag{48}$$

Based on a set of B bootstrapped values $T_{n1}^, T_{n2}^*, \dots, T_{nB}^*$, a estimate of (48) is*

$$\sqrt{\frac{1}{B-1} \sum (T_{ni}^* - \overline{T_n^*})^2}$$

Example 18 (*Bootstrap estimate of bias of T_n*) Suppose that it is $\theta = \eta(F)$ that one hopes to estimate with T_n . The bias of T_n is

$$\text{Bias}_F(T_n) = E_F T_n - \eta(F)$$

So a sensible estimate of this bias is

$$\overline{T_n^*} - \eta(\widehat{F})$$

And a possible bias-corrected version of T_n is

$$T_n - \left(\overline{T_n^*} - \eta(\widehat{F}) \right)$$

Notice that in the case that \widehat{F} is the empirical distribution of $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ and $T_n = \eta(\widehat{F})$, this bias correction for T_n amounts to using $2T_n - \overline{T_n^*}$ to estimate θ .

Example 19 (*Percentile bootstrap confidence intervals*) Suppose that a quantity $\theta = \eta(F)$ is of interest and that

$$T_n = \eta(\text{the empirical distribution of } \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n)$$

Based on B bootstrapped values $T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*$, define ordered values

$$T_{n(1)}^* \leq T_{n(2)}^* \leq \dots \leq T_{n(B)}^*$$

Adopt the following convention (to locate lower and upper $\frac{\alpha}{2}$ points for the histogram/empirical distribution of the B bootstrapped values). For

$$k_L = \left\lfloor \frac{\alpha}{2} (B + 1) \right\rfloor \text{ and } k_U = (B + 1) - k_L$$

($\lfloor x \rfloor$ is the largest integer less than or equal to x) the interval

$$\left[T_{n(k_L)}^*, T_{n(k_U)}^* \right] \tag{49}$$

contains (roughly) the “middle $(1 - \alpha)$ fraction of the histogram of bootstrapped values.” This interval is called the (uncorrected) “ $(1 - \alpha)$ level bootstrap percentile confidence interval” for θ .

The standard argument for why interval (49) might function as a confidence interval for θ is as follows. Suppose that there is an increasing function $m(\cdot)$ such that with

$$\phi = m(\theta) = m(\eta(F))$$

and

$$\widehat{\phi} = m(T_n) = m(\eta(\text{the empirical distribution of } \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n))$$

for large n

$$\widehat{\phi} \sim N(\phi, w^2)$$

Then a confidence interval for ϕ is

$$\left[\widehat{\phi} - zw, \widehat{\phi} + zw \right]$$

and a corresponding confidence interval for θ is

$$\left[m^{-1}(\widehat{\phi} - zw), m^{-1}(\widehat{\phi} + zw) \right] \quad (50)$$

The argument is then that the bootstrap percentile interval (49) for large n and large B approximates this interval (50). The plausibility of an approximate correspondence between (49) and (50) might be argued as follows. Interval (50) is approximately

$$\begin{aligned} & \left[m^{-1}(\phi - zw), m^{-1}(\phi + zw) \right] \\ = & \left[m^{-1} \left(\text{lower } \frac{\alpha}{2} \text{ point of the dsn of } \widehat{\phi} \right), m^{-1} \left(\text{upper } \frac{\alpha}{2} \text{ point of the dsn of } \widehat{\phi} \right) \right] \\ = & \left[m^{-1} \left(m \left(\text{lower } \frac{\alpha}{2} \text{ point of } T_n \text{ dsn} \right) \right), m^{-1} \left(m \left(\text{lower } \frac{\alpha}{2} \text{ point of } T_n \text{ dsn} \right) \right) \right] \\ = & \left[\text{lower } \frac{\alpha}{2} \text{ point of the dsn of } T_n, \text{upper } \frac{\alpha}{2} \text{ point of the dsn of } T_n \right] \end{aligned}$$

and one may hope that interval (49) approximates this last interval. The beauty of the bootstrap argument in this context is that one doesn't need to know the correct transformation m (or the standard deviation w) in order to apply it.

Example 20 (*Bias corrected and accelerated bootstrap confidence intervals (BC_a and ABC intervals)*) In the set-up of Example 19, the statistical folklore is that for small to moderate samples, the intervals (49) may not have actual coverage probabilities close to nominal. Various modifications/improvements on the basic percentile interval have been suggested. Two such are the “bias corrected and accelerated” (BC_a) intervals and the “approximate biased corrected” (ABC) intervals.

The basic idea in the first of these cases is to modify the percentage points of the histogram of bootstrapped values one uses. That is, in place of the intervals (49), the BC_a idea employs

$$\left[T_{n(\alpha_1(B+1))}^*, T_{n(\alpha_2(B+1))}^* \right] \quad (51)$$

for appropriate fractions α_1 and α_2 . (Round $\alpha_1(B+1)$ down and $\alpha_2(B+1)$ up in order to get integers.) For Φ the standard normal cdf,

$$z = \text{the upper } \frac{\alpha}{2} \text{ point of the standard normal distribution}$$

$$\widehat{z}_0 = \Phi^{-1}(\text{the fraction of the } T_{ni}^* \text{ that are smaller than } T_n)$$

$T_{n\check{j}}$ = T_n computed dropping the j th observation from the data set
= η (the empirical distribution of $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{j-1}, \mathbf{Y}_{j+1}, \dots, \mathbf{Y}_n$)

$$\overline{T_{n\cdot}^*} = \frac{1}{n} \sum_{j=1}^n T_{n\check{j}}$$

and

$$\hat{a} = \frac{\sum_{j=1}^n (\overline{T_{n\cdot}^*} - T_{n\check{j}})^3}{6 \left(\sum_{j=1}^n (\overline{T_{n\cdot}^*} - T_{n\check{j}})^2 \right)^{3/2}}$$

the BC_a method (51) uses

$$\alpha_1 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 - z}{1 - \hat{a}(\hat{z}_0 - z)} \right) \text{ and } \alpha_2 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z}{1 - \hat{a}(\hat{z}_0 + z)} \right)$$

The quantity \hat{z}_0 is commonly called “a measure of median bias of T_n in normal units” and is 0 when half of the bootstrapped values are smaller than T_n . The quantity \hat{a} is commonly called an “acceleration factor” and intends to correct the percentile intervals for the fact that the standard deviation of T_n may not be constant in θ .

Evaluation of the BC_a intervals may be computationally burdensome when n is large. There is an analytical approximation to these known as the ABC intervals that requires less computation time. Both the BC_a and ABC intervals are generally expected to do a better job of holding their nominal confidence levels than the uncorrected percentile bootstrap intervals.

4.2 Bootstrapping in Nonlinear Models

It is possible to extend the bootstrap idea much beyond the simple context of iid (single sample) models. Here we consider its application in the nonlinear models context of Section 2. That is, suppose that as in equation (16)

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i \text{ for } i = 1, 2, \dots, n$$

and T_n is some statistic of interest computed for the (\mathbf{x}_i, y_i) data. (T_n might, for example, be the j th coordinate of \mathbf{b}_{OLS} .) There are at least two commonly suggested methods for using the bootstrap idea in this context. These are bootstrapping (\mathbf{x}, y) pairs and bootstrapping residuals.

When bootstrapping (\mathbf{x}, y) pairs, one makes up a bootstrap sample

$$(\mathbf{x}, y)_1^*, (\mathbf{x}, y)_2^*, \dots, (\mathbf{x}, y)_n^*$$

by sampling with replacement from

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

For each of B such bootstrap samples, one computes a value T_n^* and proceeds as in the iid/single sample context. Indeed, the clearest rationale for this method

is in those cases where it makes sense to think of $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ as iid from some joint distribution F for (\mathbf{x}, y) .

To bootstrap residuals, one computes

$$e_i = y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \quad \text{for } i = 1, 2, \dots, n$$

and makes up bootstrap samples

$$(\mathbf{x}_1, f(\mathbf{x}_1, \hat{\boldsymbol{\beta}}) + e_1^*), (\mathbf{x}_2, f(\mathbf{x}_2, \hat{\boldsymbol{\beta}}) + e_2^*), \dots, (\mathbf{x}_n, f(\mathbf{x}_n, \hat{\boldsymbol{\beta}}) + e_n^*) \quad (52)$$

by sampling from

$$\{e_1, e_2, \dots, e_n\}$$

with replacement to get the e_i^* values.

One might improve this second plan somewhat by taking account of the fact that residuals don't have constant variance. For example, in the Gauss-Markov linear model one has $\text{Vare}_i = (1 - h_{ii})\sigma^2$ and one might think of sampling not residuals, but rather analogues of the values

$$\frac{e_1}{\sqrt{1 - h_{11}}}, \frac{e_2}{\sqrt{1 - h_{22}}}, \dots, \frac{e_n}{\sqrt{1 - h_{nn}}}$$

for addition to the $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$. (Or even going a step further, one might take account of the fact that analogues of the $e_i/\sqrt{1 - h_{ii}}$ don't (sum or) average to zero, and sample from values created by subtracting from the residuals adjusted for non-constant variance the arithmetic mean of those.)

The statistical folklore says that one must be very careful that the constant variance assumption makes sense before putting much faith in the method (52). Where there is a trend in the error variance, the method of bootstrapping residuals will produce synthetic data sets that are substantially unlike the original one. In such cases, it is unreasonable to expect the bootstrap methodology to produce reliable inferences.

5 Generalized Linear Models

Another kind of generalization of the Linear Model is meant to allow “regression type” analysis for even discrete response data. It is built on the assumption that an individual response/observation, y , has pdf or pmf

$$f(y|\theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi) \right] \quad (53)$$

for some functions $a(\cdot), b(\cdot), c(\cdot, \cdot)$, a “canonical parameter” θ , and a “dispersion parameter” ϕ . (We will soon let θ depend upon a k -vector of explanatory variables \mathbf{x} . But for the moment, we need to review some properties of the exponential family (53).) Notable distributions that can be written in the

form (53) include the normal, Poisson, binomial, gamma and inverse normal distributions.

General statistical theory (whose validity extends beyond the exponential family) implies that

$$\mathbb{E}_{\theta_0, \phi_0} \left. \frac{\partial}{\partial \theta} \log f(y|\theta, \phi) \right|_{\theta_0, \phi_0} = 0$$

and that

$$\text{Var}_{\theta_0, \phi_0} \left(\left. \frac{\partial}{\partial \theta} \log f(y|\theta, \phi) \right|_{\theta_0, \phi_0} \right) = -\mathbb{E}_{\theta_0, \phi_0} \left. \frac{\partial^2}{\partial \theta^2} \log f(y|\theta, \phi) \right|_{\theta_0, \phi_0}$$

Here these facts imply that

$$\mu \doteq \mathbb{E}y = b'(\theta) \tag{54}$$

and that

$$\text{Vary} = a(\phi) b''(\theta) \tag{55}$$

From (54) one has

$$\theta = b'^{-1}(\mu)$$

If we write

$$v(\mu) = b''(b'^{-1}(\mu))$$

we have

$$\text{Vary} = a(\phi) v(\mu)$$

and unless v is constant, Vary varies with μ .

The heart of a generalized linear model is then the decision to model some function of $\mu = \mathbb{E}y$ as linear in \mathbf{x} . That is, for some so-called “link function” $h(\cdot)$ and some k -vector of parameters $\boldsymbol{\beta}$, we suppose that

$$h(\mu) = \mathbf{x}'\boldsymbol{\beta} \tag{56}$$

Immediately from (56) one has then assumed that

$$\mu = h^{-1}(\mathbf{x}'\boldsymbol{\beta}) \tag{57}$$

Then for n fixed vectors of predictors \mathbf{x}_i , we suppose that y_1, y_2, \dots, y_n are independent with $\mathbb{E}y_i = \mu_i$ satisfying

$$h(\mu_i) = \mathbf{x}'_i \boldsymbol{\beta}$$

(Probably to avoid ambiguities, we should also assume that

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix}$$

is of full rank.) The likelihood function for this model is then

$$f(\mathbf{Y}|\boldsymbol{\beta}, \phi) = \prod_{i=1}^n \exp \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi) \right]$$

where

$$\theta_i = b'^{-1}(\mu_i) = b'^{-1}(h^{-1}(\mathbf{x}'_i \boldsymbol{\beta}))$$

and inference proceeds based on large sample theory for maximum likelihood estimation and likelihood ratio testing.

5.1 Inference for the Generalized Linear Model Based on Large Sample Theory for MLEs

If $(\hat{\boldsymbol{\beta}}, \hat{\phi})$ is a maximum likelihood estimate of the parameter vector $(\boldsymbol{\beta}, \phi)$ in the generalized linear model, it must satisfy

$$\left. \frac{\partial}{\partial \beta_j} \log f(\mathbf{Y}|\boldsymbol{\beta}, \phi) \right|_{(\hat{\boldsymbol{\beta}}, \hat{\phi})} = 0 \quad \forall j$$

As it turns out, these k equations can be written as

$$0 = \sum_{i=1}^n \frac{x_{ij}}{v(h^{-1}(\mathbf{x}'_i \boldsymbol{\beta})) h'(h^{-1}(\mathbf{x}'_i \boldsymbol{\beta}))} [y_i - h^{-1}(\mathbf{x}'_i \boldsymbol{\beta})] \quad \forall j \quad (58)$$

which doesn't involve the dispersion parameter or the form of the function a . Solving the equations (58) for an estimate $\hat{\boldsymbol{\beta}}$ is a numerical problem handled by various clever algorithms.

If $\hat{\boldsymbol{\beta}}$ solves the equations (58), standard large sample theory suggests that for large n

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}_k \left(\boldsymbol{\beta}, a(\phi) (\mathbf{X}' \mathbf{W}(\boldsymbol{\beta}) \mathbf{X})^{-1} \right)$$

for

$$\mathbf{W}(\boldsymbol{\beta}) = \text{diag} \left(\frac{1}{v(h^{-1}(\mathbf{x}'_1 \boldsymbol{\beta})) (h'(h^{-1}(\mathbf{x}'_1 \boldsymbol{\beta})))^2}, \dots, \frac{1}{v(h^{-1}(\mathbf{x}'_n \boldsymbol{\beta})) (h'(h^{-1}(\mathbf{x}'_n \boldsymbol{\beta})))^2} \right)$$

So, for example, if

$$\mathbf{Q} = \widehat{a(\phi)} \left(\mathbf{X}' \mathbf{W}(\hat{\boldsymbol{\beta}}) \mathbf{X} \right)^{-1}$$

and q_j is the j th diagonal entry of \mathbf{Q} , approximate confidence limits for β_j are

$$\hat{\beta}_j \pm z \sqrt{q_j}$$

and more generally, approximate confidence limits for $\mathbf{c}'\boldsymbol{\beta}$ are

$$\mathbf{c}'\hat{\boldsymbol{\beta}} \pm z \sqrt{\mathbf{c}'\mathbf{Q}\mathbf{c}}$$

(which, for example, gives limits for the linear functions $\mathbf{x}'_i\boldsymbol{\beta}$, which can then be transformed to limits for the $\mu_i = h^{-1}(\mathbf{x}'_i\boldsymbol{\beta})$).

In the event that one is working in a model where $a(\cdot)$ is not constant, it is possible to do maximum likelihood for ϕ , that is try to solve

$$\frac{\partial}{\partial\phi} \log f(\mathbf{Y}|\widehat{\boldsymbol{\beta}}, \phi) = 0$$

for ϕ . This is

$$\sum_{i=1}^n \left[\frac{a'(\phi)}{a(\phi)^2} (y_i\widehat{\theta}_i - b(\widehat{\theta}_i)) + \frac{\partial}{\partial\phi} c(y_i, \phi) \right] = 0$$

for $\widehat{\theta}_i = b'^{-1}(\widehat{\mu}_i) = b'^{-1}(h^{-1}(\mathbf{x}'_i\widehat{\boldsymbol{\beta}}))$. However, it appears that it is more common to simply estimate $a(\phi)$ directly as

$$\widehat{a(\phi)} = \frac{1}{n-k} \sum_{i=1}^n \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)}$$

for $\widehat{\mu}_i = h^{-1}(\mathbf{x}'_i\widehat{\boldsymbol{\beta}})$.

5.2 Inference for the Generalized Linear Model Based on Large Sample Theory for Likelihood Ratio Tests/Shape of the Likelihood Function

The standard large sample theory for likelihood ratio tests/shape of the likelihood function can be applied here. Let

$$l(\boldsymbol{\beta}, \phi) = \log f(\mathbf{Y}|\boldsymbol{\beta}, \phi)$$

With $\widehat{\phi}(\boldsymbol{\beta})$ a maximizer of $l(\boldsymbol{\beta}, \cdot)$, the set

$$\left\{ \boldsymbol{\beta} \mid l(\widehat{\boldsymbol{\beta}}, \widehat{\phi}) - l(\boldsymbol{\beta}, \widehat{\phi}(\boldsymbol{\beta})) \leq \frac{1}{2}\chi_k^2 \right\}$$

is an approximate confidence set for $\boldsymbol{\beta}$. Similarly, with $\boldsymbol{\beta}_{\bar{j}}$ the $(k-1)$ -vector with β_j deleted, and $(\widehat{\boldsymbol{\beta}}_{\bar{j}}, \phi)(\beta_j)$ a maximizer of l subject to the given value of β_j , the set

$$\left\{ \beta_j \mid l(\widehat{\boldsymbol{\beta}}, \widehat{\phi}) - l(\beta_j, (\widehat{\boldsymbol{\beta}}_{\bar{j}}, \phi)(\beta_j)) \leq \frac{1}{2}\chi_1^2 \right\}$$

functions as an approximate confidence set for β_j . And in those cases where $a(\cdot)$ is not constant and one needs an inference method for ϕ , the set

$$\left\{ \phi \mid l(\widehat{\boldsymbol{\beta}}, \widehat{\phi}) - l(\widehat{\boldsymbol{\beta}}, \phi) \leq \frac{1}{2}\chi_1^2 \right\}$$

is an approximate confidence set for ϕ .

5.3 Common GLMs

In Generalized Linear Model parlance, the link function that makes $\theta = \mathbf{x}'\boldsymbol{\beta}$ is called the “canonical link function.” This is $h = b'^{-1}$. Usually, statistical software for GLMs defaults to the canonical link function unless a user specifies some other option. Here we simply list the three most common GLMs, the canonical link function and common alternatives.

The Gauss-Markov normal linear model is the most famous GLM. Here $\theta = \mu = \mathbb{E}y$, the canonical link function is the identity, $h(\mu) = \mu$, and this makes $\mu = \mathbf{x}'\boldsymbol{\beta}$.

Poisson responses make a second standard GLM. Here, $\mu = \mathbb{E}y = b'(\theta) = \exp(\theta)$, so that $\theta = \log \mu$. The canonical link is the log function, $h(\mu) = \log \mu$, and

$$\mu = \exp(\mathbf{x}'\boldsymbol{\beta})$$

This is the “log-linear model.” Other possibilities for links are the identity $h(\mu) = \mu$ (giving $\mu = \mathbf{x}'\boldsymbol{\beta}$), and the square root link, $h(\mu) = \sqrt{\mu}$ (giving $\mu = (\mathbf{x}'\boldsymbol{\beta})^2$).

Bernoulli or binomial responses constitute a third important case. Here, binomial (m, p) responses have $\mu = \mathbb{E}y = mp$. For Bernoulli (binomial with $m = 1$) cases the canonical link is the logit function

$$h(\mu) = \log \frac{\mu}{1 - \mu} = \log \frac{p}{1 - p}$$

from which

$$p = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}$$

This is so-called “logistic regression.” And if one wants to think not about individual Bernoulli outcomes, but rather the numbers of “successes” in m trials (binomial observations), the corresponding canonical link is $h(\mu) = \log((\mu/n) / (1 - (\mu/n)))$.

Other possible links for the Bernoulli/binomial case are the “probit” and “complimentary log log” links. The first of these is

$$h(\mu) = \Phi^{-1}(p)$$

which gives $p = \Phi(\mathbf{x}'\boldsymbol{\beta})$ and the case of “probit analysis.” The second is

$$h(\mu) = \log(-\log(1 - p))$$

which gives $p = 1 - \exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))$.

6 Smoothing Methods

There are a variety of methods for fitting “curves” through n data pairs

$$(x_i, y_i)$$

that are more flexible than, for example, polynomial regression based on the linear models material of Section 1 or the fitting of some parametric nonlinear function $y = f(x, \beta)$ via nonlinear least squares as in Section 2. Here we briefly describe a few of these methods. (A good source for material on this topic is the book *Generalized Additive Models* by Hastie and Tibshirani.)

Bin smoothers partition the part of the x axis of interest into disjoint intervals (bins) and use as a “smoothed” y value at x ,

$$\hat{y}(x) = \text{“operation O” applied to those } y_i \text{ whose } x_i \text{ are in the same bin as } x$$

Here “operation O” can be “median,” “mean,” “linear regression,” or yet something else. A choice of “median” or “mean” will give the same value across an entire bin, and therefore produce a step function for $\hat{y}(x)$.

Running smoothers, rather than applying a fixed finite set of bins across the interval of values of x of interest, “use a different bin for each x ,” with bin defined in terms of “ k nearest neighbors.” (So the bin length changes with x , but the number of points considered in producing $\hat{y}(x)$ does not.) Symmetric versions apply “operation O” to the $\frac{k}{2}$ values y_i whose x_i are less than or equal to x and closest to x , together with the $\frac{k}{2}$ values y_i whose x_i are greater than than or equal to x and closest to x . Possibly non-symmetric versions simply use the k values y_i whose x_i are closest to x . Where “operation O” is (simple linear) regression, this is a kind of local regression that treats all points in a neighborhood of x equally. By virtue of the way least squares works, this gives neighbors furthest from x the greatest influence on $\hat{y}(x)$. And like bin smoothers, these running smoothers will typically produce fits that have discontinuities.

Locally weighted running (polynomial regression) smoothers can be thought of as addressing deficiencies of the running regression smoother. For a given x , one

1. identifies the k values x_i closest to x (call this set $N_k(x)$)
2. computes the distance from x to the element of $N_k(x)$ furthest from x ,

$$\Delta_k(x) = \max_{x_i \in N_k(x)} |x - x_i|$$

3. assigns weights to elements of $N_k(x)$ as

$$\begin{aligned} w_i &= \text{weight on } x_i \\ &= W\left(\frac{|x - x_i|}{\Delta_k(x)}\right) \end{aligned}$$

for some appropriate “weight function” $W : [0, 1] \rightarrow [0, \infty)$

4. fits a polynomial regression using weighted least squares (For example, in the linear case one chooses $b_{0,x}$ and $b_{1,x}$ to minimize

$$\sum_{i \text{ s.t. } x_i \in N_k(x)} w_i (y_i - (b_{0,x} + b_{1,x}x_i))^2$$

There are simple closed forms for such $b_{0,x}$ and $b_{1,x}$ in terms of the x_i, y_i and w_i .)

5. uses

$\hat{y}(x)$ = the value at x of the polynomial fit at x

The size of k makes a difference in how smooth such a $\hat{y}(x)$ will be (the larger is k , the smoother the fit). It is sometimes useful to write $k = \lambda n$ so that the parameter λ stands for the fraction of the data set used to do the fitting at any x . And a standard choice for weight function is the “tricube” function

$$W(u) = \begin{cases} (1 - u^3)^3 & 0 \leq u \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

When this is applied with simple linear regression, the smoother is typically called the *loess* smoother.

Kernel smoothers are locally weighted means with weights coming from a “kernel” function. A kernel $K(t) \geq 0$ is smooth, possibly symmetric about 0, and decreases as t moves away from 0. For example, one might use the standard normal density, $\phi(t)$, as a kernel function. One defines weights at x by

$$w_i(x) = K\left(\frac{x_i - x}{b}\right)$$

for $b > 0$ a “bandwidth” parameter. The kernel smoother is then

$$\hat{y}(x) = \frac{\sum y_i w_i(x)}{\sum w_i(x)}$$

The choice of the Gaussian kernel is popular, as are the choices

$$K(t) = \begin{cases} \frac{3}{4}(1 - t^2) & -1 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$K(t) = \begin{cases} \frac{3}{8}(3 - 5t^2) & -1 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(Note that this last choice can produce negative weights.)

The idea behind *polynomial regression splines* is to fit a function via least squares that is piecewise a polynomial and is smooth (has a couple of continuous derivatives). That is, if smoothing over the interval $[a, b]$ is desired, one picks positions

$$a \leq \xi_1 \leq \xi_2 \leq \dots \leq \xi_k \leq b$$

for k “knots” (points at which $k+1$ polynomials will be “tied together”). Then, for example, in the cubic case one fits by OLS the relationship

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^k \theta_j (x - \xi_j)_+^3$$

where

$$(x - \xi_j)_+^3 = \begin{cases} (x - \xi_j)^3 & \text{if } x \geq \xi_j \\ 0 & \text{otherwise} \end{cases}$$

This can be done using a regression program (there are $k + 4$ regression parameters to fit). The fitted function is cubic on each interval $[a, \xi_1], [\xi_1, \xi_2], \dots, [\xi_{k-1}, \xi_k], [\xi_k, b]$ with 2 continuous derivatives (and therefore smoothness) at the knots.

Finally (in this short list of possible methods) are the *cubic smoothing splines*. It turns out to be possible to solve the optimization problem of minimizing

$$\sum (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt \quad (59)$$

over choices of function f with 2 continuous derivatives. In criterion (59), the sum is a penalty for “missing” the data values with the function, the integral is a penalty for f wiggling (a linear function would have second derivative 0 and incur no penalty here), and λ weights the two criteria against each other (and is sometimes called a “stiffness” parameter). Even such widely available and user-friendly software as JMP provides such a smoother.

Theory for choosing amongst the possible smoothers and making good choices of parameters (like bandwidth, stiffness, knots, kernel, etc.), and doing approximate inference is a whole additional course.

7 Appendix

Here we collect some general results that are used repeatedly in Stat 511.

7.1 Some Useful Facts About Multivariate Distributions (in Particular Multivariate Normal Distributions)

Here are some important facts about multivariate distributions in general and multivariate normal distributions specifically.

1. If a random vector \mathbf{X} has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{d}$ has mean vector $E\mathbf{Y} = \mathbf{B}\boldsymbol{\mu} + \mathbf{d}$ and covariance matrix $\text{Var}\mathbf{Y} = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'$.
2. The MVN distribution is most usefully defined as the distribution of $\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$, for \mathbf{Z} a vector of independent standard normal random variables. Such a random vector has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$. (This definition turns out to be unambiguous. Any dimension p and any matrix \mathbf{A} giving a particular $\boldsymbol{\Sigma}$ end up producing the same k -dimensional joint distribution.)

a multivariate Taylor's Theorem argument and fact 1 in Section 7.1 provide an approximate mean vector and an approximate covariance matrix for \mathbf{Y} . That is, if the functions g_i are differentiable, let

$$\mathbf{D} = \begin{pmatrix} \frac{\partial g_i}{\partial x_j} \Big|_{\mu_1, \mu_2, \dots, \mu_k} \end{pmatrix}_{p \times k}$$

A multivariate Taylor approximation says that for each x_i near μ_i

$$\mathbf{y} = \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_p(\mathbf{x}) \end{pmatrix} \approx \begin{pmatrix} g_1(\boldsymbol{\mu}) \\ g_2(\boldsymbol{\mu}) \\ \vdots \\ g_p(\boldsymbol{\mu}) \end{pmatrix} + \mathbf{D}(\mathbf{x} - \boldsymbol{\mu})$$

So if the variances of the X_i are small (so that with high probability \mathbf{Y} is near $\boldsymbol{\mu}$, that is that the linear approximation above is usually valid) it is plausible that \mathbf{Y} has mean vector

$$\begin{pmatrix} \text{E}Y_1 \\ \text{E}Y_2 \\ \vdots \\ \text{E}Y_k \end{pmatrix} \approx \begin{pmatrix} g_1(\boldsymbol{\mu}) \\ g_2(\boldsymbol{\mu}) \\ \vdots \\ g_k(\boldsymbol{\mu}) \end{pmatrix}$$

and variance-covariance matrix

$$\text{Var}\mathbf{Y} \approx \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'$$

7.3 Large Sample Inference Methods

Here is a summary of two variants of large sample theory that are used extensively in Stat 511 to produce approximate inference methods where no exact methods are available. These are theories of the behavior of maximum likelihood estimators (actually, of solutions to the likelihood equation(s)) and of the shape of the likelihood function/behavior of likelihood ratio tests.

7.3.1 Large n Theory for Maximum Likelihood Estimation

Let $\boldsymbol{\theta}$ be a generic $r \times 1$ parameter and let $l(\boldsymbol{\theta})$ stand for the natural log of the likelihood function. The likelihood equations are the r equations

$$\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}) = 0$$

Suppose that $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ that is “consistent for $\boldsymbol{\theta}$ ” (for large samples “is guaranteed to be close to $\boldsymbol{\theta}$ ”) and solves the likelihood equations, i.e. has

$$\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \quad \forall i$$

Typically, maximum likelihood estimators have these properties (they solve the likelihood equations and are consistent). There is general statistical theory that then suggests that for large samples

1. $\widehat{\boldsymbol{\theta}}$ is approximately MVN_r ,
2. $E\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}$
3. an estimated variance-covariance matrix for $\widehat{\boldsymbol{\theta}}$ may be obtained from second partials of $l(\boldsymbol{\theta})$ evaluated at $\widehat{\boldsymbol{\theta}}$. That is, one may use

$$\widehat{\text{Var}}\widehat{\boldsymbol{\theta}} = \left(- \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right)_{r \times r}^{-1} \quad (60)$$

These are very important facts, and for example allow one to make approximate confidence limits for the entries of $\boldsymbol{\theta}$. If d_i is the i th diagonal entry of the estimated variance-covariance matrix (60), θ_i is the i th entry of $\boldsymbol{\theta}$, and $\widehat{\theta}_i$ is the i th entry of $\widehat{\boldsymbol{\theta}}$, approximate confidence limits for θ_i are

$$\widehat{\theta}_i \pm z\sqrt{d_i}$$

for z an appropriate standard normal quantile.

7.3.2 Large n Theory for Inference Based on the Shape of the Likelihood Function/Likelihood Ratio Testing

There is a second general approach (beyond maximum likelihood estimation) that is applied to produce (large sample) inferences in Stat 511 (in Sections 2,3 and 5, where no exact theory is available). That is the large sample theory of likelihood ratio testing and its inversion to produce confidence sets.

Again let $\boldsymbol{\theta}$ be a generic $r \times 1$ parameter vector that we now partition as

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ p \times 1 \\ \boldsymbol{\theta}_2 \\ (r-p) \times 1 \end{pmatrix}$$

(We may take $p = r$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ here.) Further, again let $l(\boldsymbol{\theta})$ stand for the natural log of the likelihood function. There is a general large sample theory of likelihood ratio testing/shape of $l(\boldsymbol{\theta})$ that suggests that if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$

$$2 \left(\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) - \max_{\boldsymbol{\theta} \text{ with } \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0} l(\boldsymbol{\theta}) \right) \sim \chi_p^2$$

Now if for each $\boldsymbol{\theta}_1$, $\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ maximizes $l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ over choices of $\boldsymbol{\theta}_2$, this may be rewritten as

$$2 \left(l(\widehat{\boldsymbol{\theta}}_{\text{MLE}}) - l(\boldsymbol{\theta}_1^0, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0)) \right) \sim \chi_p^2$$

This leads to the testing of $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ using the statistic

$$2 \left(l \left(\hat{\boldsymbol{\theta}}_{\text{MLE}} \right) - l \left(\boldsymbol{\theta}_1^0, \hat{\boldsymbol{\theta}}_2 \left(\boldsymbol{\theta}_1^0 \right) \right) \right)$$

and a χ_p^2 reference distribution.

A $(1 - \alpha)$ level confidence set for $\boldsymbol{\theta}_1$ can be made by assembling all of those $\boldsymbol{\theta}_1^0$'s for which the corresponding α level test of $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ does not reject. This amounts to

$$\left\{ \boldsymbol{\theta}_1 \mid l \left(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2 \left(\boldsymbol{\theta}_1 \right) \right) > l \left(\hat{\boldsymbol{\theta}}_{\text{MLE}} \right) - \frac{1}{2} \chi_p^2 \right\} \quad (61)$$

where χ_p^2 is the upper α point. The function

$$l^* \left(\boldsymbol{\theta}_1 \right) = l \left(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2 \left(\boldsymbol{\theta}_1 \right) \right)$$

is the loglikelihood maximized out over $\boldsymbol{\theta}_2$ and is called the “profile loglikelihood function” for $\boldsymbol{\theta}_1$. The formula (61) is a very important one and says that an approximate confidence set for $\boldsymbol{\theta}_1$ consists of all those vectors whose values of the profile loglikelihood are within $\frac{1}{2} \chi_p^2$ of the maximum loglikelihood. Statistical folklore says that such confidence sets typically do better in terms of holding their nominal confidence levels than the kind of methods outline in the previous section.