Koopmans T (ed.) 1951 *Activity Analysis of Production and Allocation*. Wiley, New York

Kuhn H W, Tucker A W 1951 Nonlinear programming. *Econometrica* **19**: 50–1

Mangasarian O L 1969 *Nonlinear Programming*. McGraw-Hill, New York

Rockafellar R T 1970 *Convex Analysis*. Princeton University Press, Princeton, NJ

Rockafellar R T, Wets R J-B 1998 *Variational Analysis*. Springer, Berlin

Roos C, Terlaky T, Vial J-Ph 1997 *Theory and Algorithms for Linear Optimization: An Interior Point Approach*. Wiley, New York

Smale S 1983 On the average number of steps of the simplex method of linear programming. *Mathematical Programming* **27**: 241–62

Von Neumann J, Morgenstern O 1944 *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ

A. N. Iusem

Linear Hypothesis

1. Introduction

The term 'linear hypothesis' is often used interchangeably with the term 'linear model.' Statistical methods using linear models are widely used in the behavioral and social sciences, e.g., regression analysis, analysis of variance, analysis of covariance, multivariate analysis, time series analysis, and spatial data analysis. Linear models provide a flexible tool for data analysis and useful approximations for more complex models.

A common object of linear modeling is to find the most precise linear model that explains the data, to use that model to predict future observations, and to interpret that model in the context of the data collection. Traditionally, analysis of variance models have been used to analyze data from designed experiments while regression analysis has been used to analyze data from observational studies but the techniques of both analysis methods apply to both kinds of data. See also *Experimental Design: Overview* and *Observational Studies: Overview.*

2. Definition

The linear hypothesis is that the mean (average) of a random observation can be written as a linear combination of some observed predictor variables. For example, Coleman et al. (1996) provides observations on various schools. The dependent variable y consists of the average verbal test score for sixth-grade students. The report also presents predictor variables. A composite measure of socioeconomic status x_1 is based on father's and mother's education, family size and intactness, home items, and percent of fathers who are white collar. Staff salaries per pupil is x_2. The average score on a verbal test given to the school's teachers is x_3. Denoting different schools using the subscript i and the mean of y_i by m_i, a linear hypothesis states that for some unknown numbers (parameters) β_0, β_1, β_2, and β_3,

$$m_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_3 x_{i3}$$

Mosteller and Tukey (1977, pp. 326, 566) and Mosteller et al. (1983, pp. 408–20) give excerpts and analysis of the data.

In other applications, the predictors only identify whether an observation is in some group. For example, in 1978 observations y_{ij} were collected on the age at which people in Albuquerque committed suicide, see Koopmans (1987, p. 409). Here i is used to identify the person's group membership (Hispanic, Native American, non-Hispanic Caucasian), and j identifies individuals within a group. The three categories are taken to be mutually exclusive for the present discussion (although the US government now allows for individuals to identify themselves with multiple races in various surveys and the decennial census). We can define group identifier predictor variables. Let δ_{1i} take the value 1 if an individual belongs to group 1 (Hispanic) and 0 otherwise, with similar predictors to identify other groups, say δ_{2i} and δ_{3i} for Native Americans and non-Hispanic Caucasians. Note that the predictor variables do not depend on the value of j identifying individuals within a group. Denoting the mean of y_{ij} by m_{ij}, a linear hypothesis states that for some unknown parameters μ, α_1, α_2, α_3,

$$m_{ij} = \mu + \alpha_1 \delta_{1i} + \alpha_2 \delta_{2i} + \alpha_3 \delta_{3i}$$

Since two of the δs are always zero, this model is often written more succinctly as

$$m_{ij} = \mu + \alpha_i$$

A linear hypothesis is usually combined with other assumptions about the observations y. Most commonly, the assumptions are that the observations are independent, have the same (unknown) variance σ^2, and have normal (Gaussian) distributions. For the two examples, these assumptions are written

$$y_i \text{ indep. } N(m_i, \sigma^2) \quad \text{and} \quad y_{ij} \text{ indep. } N(m_{ij}, \sigma^2)$$

where, for example, $N(m_i, \sigma^2)$ indicates a normal distribution with mean m_i and variance σ^2. Incorporating these additional assumptions, the linear hypothesis becomes a full-fledged linear model, traditionally written

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,\ \varepsilon_i \text{ indep. } N(0, \sigma^2) \tag{1}$$

and

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \varepsilon_{ij} \text{ indep. } N(0, \sigma^2) \tag{2}$$

where the εs are unobservable random errors.

Both theoretical and computational work with linear hypotheses and linear models is facilitated by the use of matrices. If $i = 1, 2, 3, 4$, the model (1) can be written as the equality of two 4×1 matrices

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{33} + \varepsilon_2 \\ \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \varepsilon_3 \\ \beta_0 + \beta_1 x_{41} + \beta_2 x_{42} + \beta_3 x_{43} + \varepsilon_4 \end{bmatrix}$$

Using properties of matrix algebra, this can be rewritten as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Similarly, model (6) for $i = 1, 2, 3$, $j = 1, 2$ can be written using δ as

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

In general, a linear model is traditionally written using matrices as

$$Y = X\beta + e \tag{3}$$

where Y is a $n \times 1$ matrix containing the observable random variables y, X is a $n \times p$ matrix containing observed predictors, β is a $p \times 1$ matrix containing unobservable parameters, and e is a $n \times 1$ matrix containing unobservable random errors. To have a full-fledged model, assumptions must be made about the distribution of e. One standard assumption is that the mean value of each element of e is 0. Denote the $n \times 1$ matrix of mean (expected) values as $\mathrm{E}(e)$. This assumption is written

$$\mathrm{E}(e) = 0$$

The corresponding linear hypothesis involves the mean values of the entries of Y, i.e., $\mathrm{E}(Y)$. The linear hypothesis is that

$$\mathrm{E}(Y) = X\beta$$

for some matrix β. Another standard assumption about e is that the covariance matrix, the $n \times n$ matrix of variances and covariances between the ε_is or equivalently between the y_is is

$$\mathrm{Cov}(e) = \mathrm{Cov}(Y) = \sigma^2 I$$

where I indicates the $n \times n$ identity matrix that has 1s on the diagonal and 0 elsewhere.

An alternative way to write model (3) in terms of each y_i using some matrices is

$$y_i = x_i'\beta + \varepsilon_i$$

$i = 1, \ldots, n$ where β is as before and x_i' is the ith row of the X matrix.

See also *Analysis of Variance and Generalized Linear Models* and *Linear Hypothesis: Regression (Basics)*.

3. *Estimation*

The unknown parameters of the linear model are the elements of β, the βs, and the variance of an observation, σ^2; they must be estimated from the data. The β_js are generally estimated using least squares estimation. For given data Y, the least squares estimates are the values of β_j that minimize

$$\sum_{i=1}^{n} (y_i - x_i'\beta)^2 \quad \text{or equivalently} \quad (Y - X\beta)'(Y - X\beta)$$

where the prime on $(Y - X\beta)'$ indicates the matrix transpose of $(Y - X\beta)$. (The transpose of a matrix turns the columns of the original matrix into the rows of a new matrix.)

Under the standard assumptions, least squares estimates have optimal statistical properties. If the ε_is are independent $N(0, \sigma^2)$, then the least squares estimates provide maximum likelihood estimates (MLEs) and minimum variance unbiased estimates (MVUEs). A MVUE is an estimate that has less variability than any other unbiased estimate. If we drop the assumption of normality but retain independence (actually 0 covariances) and equal variances, least squares estimates are best linear unbiased estimates (BLUEs). A BLUE is an estimate based on a linear combination of the observations that has less variability than any other unbiased estimate based on a linear combination of the observations.

In matrix terms, the least squares estimates, say, $\hat{\beta}$, are obtained as solutions to the normal equations

$$X'X\beta = X'Y$$

Typically, in regression problems, the unique solution is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

In general, if $X'X$ is singular, so that $(X'X)^{-1}$ does not exist, an infinite number of solutions exist, giving an infinite number of least squares estimates. Any one of these estimates is just as valid as any other, but fortunately, for all practical purposes, the estimates are unique. This means that whenever a linear parameter $\lambda'\beta$ is identifiable in the model (estimable), the least squares estimate of the parameter is uniquely defined. See also *Statistical Identification and Estimability* and *Analysis of Variance and Generalized Linear Models.*

The other unknown parameter to be estimated is σ^2, the variance of an observation. To estimate the variance, first rearrange the model into

$$\varepsilon_i = y_i - x_i'\beta \quad \text{or} \quad e = Y - X\beta$$

Then estimate β to get the residuals

$$\hat{\varepsilon}_i = y_i - x_i'\hat{\beta} \quad \text{or} \quad \hat{e} = Y - X\hat{\beta}$$

The sum of the squared residuals is called the sum of squares error,

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

The degrees of freedom error (dfE) is n minus the rank of the matrix X, i.e.,

$$df\text{E} = n - r(X)$$

The rank of X is the number of free parameters that can be fitted to the model by least squares. The variance σ^2 is estimated by the mean squared error,

$$MSE = \frac{SS\text{E}}{df\text{E}}$$

See also *Estimation: Point and Interval.*

4. Testing

The basic idea of testing is to create some model for the process of generating the data, and evaluate whether the observed data are consistent with that model. For example, such a model might be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

One can think of this model has having four parts: (a) ε_i are independent, (b) ε_i have $\mathrm{E}(\varepsilon_i) = 0$, i.e.,

$$\mathrm{E}(y_i) = m_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

(c) ε_i have common variance σ^2, (d) ε_i are normally distributed. Data that are inconsistent with the model may occur whenever any of these four parts is inappropriate.

In the context of testing the linear hypothesis, it is assumed that these four specifications are correct and a further refinement of the model is then imposed by placing an additional restriction on

$$\mathrm{E}(y_i) = m_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

The test is usually thought of as a test of this additional restriction, generally called the null hypothesis. For example, the additonal restriction could be the null hypothesis H_0: $\beta_1 = \beta_3 = 0$. In the sixth grade verbal test example, this hypothesis indicates that the variables x_1, socioeconomic status, and x_3, teachers mean verbal test score, have no ability to help predict sixth graders' mean verbal test scores. The test of whether the data are consistent with a model that incorporates $\beta_1 = \beta_3 = 0$ is based on assuming that the null hypothesis and items (a) through (d) are all correct. If assumptions (a) through (d) are correct, then data that are inconsistent with the model suggest that either $\beta_1 \neq 0$ or $\beta_3 \neq 0$ or both, i.e., that the null hypothesis is not true. However, in reality, data that are inconsistent with the model can result if any of the assumptions are violated, i.e., any of the four itemized assumptions or the null hypothesis. Because of this, if one is interested in testing only the null hypothesis, it is crucial to do everything possible to validate the assumptions embodied in items (a) through (d). See also *Hypothesis Testing in Statistics* and *Significance, Tests of.*

There are two ways to think about testing a linear hypothesis. One is to think in terms of testing models and the other is to think in terms of testing parameters.

4.1 Models

In testing models one assumes the validity of a (full) model and tests to see if some smaller (reduced) model is consistent with the data. For example, with the verbal test score data a full model might be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

incorporating all three predictors. The reduced model may incorporate the idea that the socioeconomic and mean teacher verbal score variables x_{i1} and x_{i3} are not important, yielding

$$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

In this context, the null hypothesis is that the reduced model provides an adequate explanation for the data. In any particular test, the full model need not incorporate all of the variables that have been measured. It need only be a model that is plausibly correct.

Both the full model and the reduced model have sums of squares error, degrees of freedom error, and mean squared errors associated with them. The reduced model is rejected as being inconsistent with the data if the following F statistic is too large,

$$F = \frac{SSE(\text{Red}) - SSE(\text{Full})}{dfE(\text{Red}) - dfE(\text{Full})} / MSE(\text{Full})$$

If the reduced model is true, then both the full model and the reduced model are true, so both $MSE(\text{Red})$ and $MSE(\text{Full})$ are estimates of σ^2. It can then be shown that $(SSE(\text{Red}) - SSE(\text{Full}))/(dfE(\text{Red}) - dfE(\text{Full}))$ is also an estimate of σ^2, so both the numerator and the denominator of F are estimates of σ^2, and the ratio should be about 1. Of course, there is sampling variability associated with the F statistic. The theoretical F distribution describes the usual distribution of values one will encounter when making such calculations. It depends on two parameters: $dfE(\text{Red}) - dfE(\text{Full})$ degrees of freedom for the numerator and $dfE(\text{Full})$ degrees of freedom for the denominator.

On the other hand, if the reduced model is an inadequate explanation of the data, it can be shown that $(SSE(\text{Red}) - SSE(\text{Full}))/(dfE(\text{Red}) - dfE(\text{Full}))$ estimates σ^2 plus a positive number, so that if the reduced model is bad, the F statistic should be larger than 1. If the observed value of the F statistic is so much larger than 1 as to be inconsistent with the theoretical F distribution, one can reasonably conclude that the reduced model is an inadequate explanation of the data, i.e., one rejects the null hypothesis.

4.2 *Parameters*

The other way to perform tests in the linear hypothesis is to specify a null hypothesis in terms of the parameters. For example, if we assume the verbal test score model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

we might be interested in testing the null hypothesis H_0: $\beta_1 = \beta_3 = 0$ in which socioeconomic status and teachers mean verbal test score have no ability to predict. One way to test this hypothesis is to identify the reduced model associated with it, i.e.,

$$y_i = \beta_0 + 0x_{i1} + \beta_2 x_{i2} + 0x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

or

$$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

For a more complicated example, suppose two previous studies had suggested the hypothesis H_0: $\beta_1 + \beta_3 = 5$ and $\beta_2 = 0$. We can rewrite the hypothesis as H_0: $\beta_1 = 5 - \beta_3$ and $\beta_2 = 0$, and create a reduced model by substitution,

$$y_i = \beta_0 + (5 - \beta_3)x_{i1} + 0x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$
$$\varepsilon_i, \text{ indep. } N(0, \sigma^2)$$

Rearranging terms gives a reduced model

$$y_i - 5x_{i1} = \beta_0 + \beta_3(x_{i3} - x_{i1}) + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

This is just a new linear model with a new dependent variable $y_i - 5x_{i1}$ and one predictor variable $(x_{i3} - x_{i1})$. One can compute $SSE(\text{Red})$ and $dfE(\text{Red})$ in the usual way for this new model, and simply substitute these quantities into the usual formula for the F statistic to get the test.

If the data are consistent with the reduced model, we have a more precise model than the one we started with. While testing provides no assurance that the reduced model is correct, the data are at least consistent with it. This more precise model can be investigated for the validity of its predictions and its usefulness in explaining the data collection process.

In addition, (estimable) parametric hypotheses can be tested directly using matrices, without identifying the reduced model. For example, the hypothesis H_0: $\beta_1 + \beta_3 = 5$ and $\beta_2 = 0$ can be written in matrix form by creating a matrix Λ' that, when multiplied by β, isolates $\beta_1 + \beta_3$ and β_2. The hypothesis is

$$\Lambda'\beta \equiv \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_4 \end{bmatrix}$$

$$= \begin{bmatrix} \beta_1 + \beta_3 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix} \equiv d$$

In matrix terms the F statistic for a null hypothesis $H_0 : \Lambda'\beta = d$ is

$$F = \frac{(\Lambda'\hat{\beta}-d)'[\Lambda'(X'X)^{-}\Lambda]^{-}(\Lambda'\hat{\beta}-d)}{r(\Lambda)MSE} \qquad (4)$$

where $\hat{\beta}$ is a least squares estimate of β, $r(\Lambda)$ is the rank of the matrix Λ, and A^{-} is a generalized inverse of the matrix A. If A^{-} exists, it is the unique generalized inverse. The reduced model test and this direct test give the same F statistic and use the same degrees of freedom.

Confidence regions can be constructed using the F distribution. For the example, we can construct a simultaneous confidence region for $\beta_1+\beta_3$ and β_2 as follows. Let $F(0.95, r(\Lambda), n-r(X))$ be the number so that with probability 0.95

$$\frac{(\Lambda'\hat{\beta}-\Lambda'\beta)'[\Lambda'(X'X)^{-}\Lambda]^{-}(\Lambda'\hat{\beta}-\Lambda'\beta)}{r(\Lambda)MSE} \leqslant F(0.95, r(\Lambda), n-r(X))$$

Here, $MSE(X'X)^{-}$ is an estimate of the covariance matrix of $\hat{\beta}$ and $MSE[\Lambda'(X'X)^{-}\Lambda]$ is an estimate of the covariance matrix of $\Lambda'\hat{\beta}$. After substituting the observed values of $\Lambda'\hat{\beta}$ and MSE into the inequality, the only thing that is unknown in this inequality is the value of $\Lambda'\beta$. A confidence region for $\Lambda'\beta$ consists of all the values that can be substituted for $\Lambda'\beta$ while maintaining the truth of the inequality. A precise definition of the 95 percent confidence region is as the collection of all values d in (4) that would not be rejected by an $\alpha = 0.05$ level F test. If Λ' has only one row, the confidence region simplifies to a confidence interval. See also *Estimation: Point and Interval.*

5. *Bayesian Estimation*

An alternative to least squares estimation is Bayesian estimation. In Bayesian estimation the parameters of the linear model are considered to be random and probability distributions reflecting the investigator's prior knowledge about the parameters are specified. Using Bayes theorem, the prior probability distributions are updated into posterior distributions that reflect the information embodied in the data. All statistical inferences are then based on the posterior distributions. Sometimes, a noninformative (improper) prior is used in which case, interval estimates arrived at using Bayesian methods are numerically equal to the interval estimates obtained from classical methods, only the interpretations of the intervals change. Alternatively, real prior information is often specified by taking the prior distribution of β given σ^2 to be a multivariate normal distribution and the prior distribution of σ^2 to be an inverse gamma distribution. This leads to modified least squares estimates that are shrunk towards the mean of the prior distribution β with corresponding changes to interval estimation. See also *Bayesian Statistics*; *Distributions, Statistical: Special and Continuous*; *Elicitation of Probabilities and Probability Distributions*.

6. *Model Checking*

Valid testing of a null hypothesis requires that all of the four itemized assumptions embodied in the model be true. Confidence intervals and regions are collections of parameter values that cannot be rejected by tests, so they also require the validity of the basic assumptions in the model as do Bayesian and non-Bayesian estimation. Model checking is considered an integral part of all statistical inference. A variety of methods have been developed for model checking. Most are based on plotting the residuals, see *Linear Hypothesis: Regression (Graphics)*. These methods include checking the normality assumption by plotting the ordered (standardized (Studentized)) residuals against (some approximation to) the expected order statistics of a standard normal distribution. Such a plot should be approximately linear if the data are normal.

The (standardized (Studentized)) residuals are also plotted against their corresponding predicted variables or any other variable (that does not depend on y) for which a measurement is available for each observation. In these plots, any identifiable systematic structure is an indication of problems with the assumptions.

Heteroscedasticity refers to having observations with different variances. In some models, formal tests for equal variances are available, e.g., Bartlett's test and the model-based tests discussed in Carroll and Ruppert (1988).

Lack of fit refers to the problem of having specified an incorrect linear mean structure, i.e., an incorrect linear hypothesis. Formal tests for lack of fit are often based on identifying clusters of observations that have similar predictor variables, i.e., similar rows of the X matrix, see Christensen (1996a, Sect. 6.6). Miller et al. (1998) have provided optimal methods for grouping observations into near replicate clusters.

The independence assumption is perhaps the hardest to evaluate. If the observations are taken at equally spaced time intervals, standard time series methods are often applied to the residuals. More generally, Christensen and Bedrick (1997) proposed creating rational subgroups (clusters) of the observations in which one suspects the observations may be more similar within clusters than between clusters. The methods of near replicate lack of fit tests can then be applied to test the independence assumption.

7. Weighted Models and Covariance Estimation

The standard assumption in linear models is that the observations all have the same variance and that the observations are all independent. If either or both of those assumptions is invalidated, different methods of estimation and testing are required. The usual linear model has

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I$$

Writing a model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 V$$

where $\sigma^2 V$ is the variance–covariance matrix of the observations (and also of the unobservable errors), V can incorporate unequal variances for observations as well as correlations between observations. Assume that V has an inverse matrix V^{-1}.

Rather than using least squares estimates of β, we use weighted least squares estimates that minimize

$$(Y - X\beta)' V^{-1} (Y - X\beta)$$

for given Y and X. The estimates can be obtained by solving the normal equations

$$X' V^{-1} X\beta = X' V^{-1} Y$$

Typically, in regression problems, the unique solution is

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} Y$$

These estimates are also BLUEs. Moreover, under the assumption that the observations have a multivariate normal distribution, the weighted least squares estimates are MLEs and MVUEs. The variance σ^2 is estimated by $MSE = SSE/dfE$ as before, except now

$$SSE = (Y - X\hat{\beta})' V^{-1} (Y - X\hat{\beta})$$

Testing a full model and a reduced model, works as before, with the F statistic still being

$$F = \frac{SSE(\mathrm{Red}) - SSE(\mathrm{Full})}{dfE(\mathrm{Red}) - dfE(\mathrm{Full})} / MSE(\mathrm{Full})$$

Note that incorporating the more general matrix V has no effect on the process of identifying reduced models—which only depend on the mean structure and not on the variance–covariance structure of the data.

All of these methods presume that V is known. In practice, V is often unknown and must be estimated from the data. This greatly complicates the entire process of data analysis. Typically, when V is estimated, weighted least squares estimates are not BLUEs or MVUEs, nor does the F statistic have an F distribution under the null hypothesis that the reduced model is true. Nonetheless, the ideas behind weighted regression are fundamental to many applications of the linear hypothesis.

A common application of weighted regression is when the observations are independent but have different variances. If the relative sizes of the variances are known, then V is known. Often the variance of each observation is modeled as a function of some predictor variables, see Carroll and Ruppert (1988).

Split plot models involve groups of observations that all have the same correlation. For example, in an experiment that involves different ways of scheduling classes in schools and different ways of assigning homework to students, the behavior of the schools might be considered independent of each other, but a group of five students taken from the same school should have similarities with each other, while being independent of students from other schools. This correlation structure must be incorporated into V. If the same number of students are measured from each school, i.e., if the groups of observations having the same correlation are all of the same size, many linear models still allow one to use least squares estimates and their associated F tests.

Mixed models involve the use of random effects, Partition X and β into two parts, so as to write

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

Instead of considering β_1 and β_2 as fixed unknown parameters, treat β_2 as random. Assume that $\mathrm{E}(\beta_2) = 0$ and $\mathrm{Cov}(\beta_2) = \sigma^2 D$ with, say, e and β_2 independent. Consolidate all of the random parts into $\xi = X_2\beta_2 + e$ and write a linear model

$$Y = X_1\beta_1 + \xi, \quad \mathrm{E}(x) = 0, \quad \mathrm{Cov}(\xi) = \sigma^2(X_2 D X_2' + I)$$

where $V = X_2 D X_2' + I$. Typically, D has to be estimated. To estimate β_1, simply use the estimate of D to get an estimate of V which is then treated as the true V.

Mixed models get much more complicated. X_2 is often partitioned into submatrices $Z_1, \ldots, Z_r$ and β_2 is correspondingly partitioned into γ_k, so as to write

$$Y = X_1\beta_1 + \sum_{k=1}^{r} Z_k\gamma_k + e$$

where $\mathrm{E}(\gamma_k) = 0$, $\mathrm{Cov}(\gamma_k) = \sigma_k^2 I$, and the γ_k are independent. The σ_k^2 are unknown parameters, and various methods such as residual (restricted) maximum likelihood (REML) and minimum norm quadratic unbiased estimation (MINQUE) are used to estimate them. See also *Hierarchical Models: Random and Fixed Effects*; *Longitudinal Data*.

Another application of weighted estimation is in the analysis of spatial data. With spatial data, observations that are taken close together spatially are usually more correlated than observations taken far apart. V reflects the spatial correlations of the observations and X may reflect the locations of the observations. Models for V are developed based on the distances between observations. A common application in geographic information systems is map making, in which predictions are made for a very fine grid of locations. An individual best linear unbiased predictor for an observation on the grid, say, y_0 corresponding to location x_0' is

$$\hat{y}_0 = x_0'\hat{\beta} + V_{OY}V^{-1}(Y - X\hat{\beta})$$

where V_{0Y} is a matrix of covariances between y_0 and the original observations in Y. In practice, both V and V_{0Y} must be estimated from the data. See also *Spatial Statistical Methods*.

Multivariate linear models involve creating linear models for several variables at once. For example, y_{i1} might measure verbal ability and y_{i2} mathematical ability. Obviously, for any individual i, y_{i1} and y_{i2} may be correlated. The multivariate linear model fits

$$y_{i1} = x_i'\beta_1 + \varepsilon_{i1} \quad \text{and} \quad y_{i2} = x_i'\beta_2 + \varepsilon_{i2}$$

simultaneously. See also *Multivariate Analysis: Overview*.

Time series data involve observations $y_1, y_2, y_3, \ldots,$ taken at regular time intervals. The correlation between two observations depends on the amount of time between them. One simple linear model for explaining the behavior of such data is a first order autoregression model.

$$y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i$$

in which the observation at time $i-1$ is used to predict the observation at time i. See also *Time Series: General* and *Time Series: ARIMA Methods*

Generalized linear models involve transforming the mean before making the linear hypothesis. For example, if observations y_i have means m_i one might make the hypothesis that

$$\log(m_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

This hypothesis together with assumptions about the distribution of y_i form a generalized linear model. In particular, since the linear hypothesis is based on the log transformation, this would be called a log-linear model. Parameter estimates are typically found by performing a series of weighted least squares analyses. Generalized linear models allow for the analysis of a variety of nonnormally distributed data. See also *Analysis of Variance and Generalized Linear Models*; *Multivariate Analysis: Discrete Variables (Loglinear Models)*; *Multivariate Analysis: Discrete Variables (Logistic Regression)*

8. Conclusion

Models based on the linear hypothesis, and their generalizations provide tools for analyzing a wide variety of data. Linear hypotheses allow one to address numerous interesting questions about the mean structure of the data. Moreover, the models can adjust for very general forms of dependence among the data. Linear models and their generalizations continue to be the most important and widely used models for the statistical analysis of data.

See also: Analysis of Variance and Generalized Linear Models; Bayesian Statistics; Distributions, Statistical: Special and Continuous; Elicitation of Probabilities and Probability Distributions; Estimation: Point and Interval; Experimental Design: Overview; Hierarchical Models: Random and Fixed Effects; Hypothesis Testing in Statistics; Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Longitudinal Data; Multivariate Analysis: Discrete Variables (Logistic Regression); Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Overview; Observational Studies: Overview; Significance, Tests of; Spatial Statistical Methods; Statistical Identification and Estimability; Time Series: ARIMA Methods; Time Series: General

Bibliography

Carroll R J, Ruppert D 1988 *Transformation and Weighting in Regression*. Chapman and Hall, New York

Christensen R 1991 *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer, New York

Christensen R 1996a *Plain Answers to Complex Questions: The Theory of Linear Models*, 2nd edn. Springer, New York

Christensen R 1996b *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, New York

Christensen R 1997 *Log-linear Models and Logistic Regression*, 2nd Edn. Springer, New York

Christensen R, Bedrick E J 1997 Testing the independence assumption in linear models. *Journal of the American Statistical Association* **92**: 1006–16

Coleman J S, Campbell E Q, Hobson C J, McPartland J, Mood A M, Weinfeld F D, York R L 1996 *Equality of Educational Opportunity*. US Dept. of Health, Education, and Welfare, Office of Education, Washington, DC

Cook R D, Weisberg S 1994 *An Introduction to Regression Graphics*. Wiley, New York

Cressie N A C 1993 *Statistics for Spatial Data*, rev. edn. Wiley, New York

Koopmans L H 1987 *Introduction to Contemporary Statistical Methods*, 2nd edn. Duxbury Press, Boston, MA

McCullagh P, Nelder J A 1989 *Generalized Linear Models*, 2nd edn. Chapman and Hall, London

Miller F R, Neill J W, Sherfey B W 1998 Maximin clusters for near-replicate regression lack of fit tests. *Annals of Statistics* **26**: 1411–33

Mosteller F, Fienberg S E, Rourke R E K 1983 *Beginning Statistics with Data Analysis*. Addison-Wesley, Reading, MA

Mosteller F, Tukey J W 1977 *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, MA

Searle S R 1971 *Linear Models*. Wiley, New York

Searle S R, Casella G, McCulloch C E 1992 *Variance Components*. Wiley, New York

Seber G A F 1977 *Linear Regression Analysis*. Wiley, New York

Seber G A F 1984 *Multivariate Observations*. Wiley, New York

Shumway R H, Stoffer D S 2000 *Time Series Analysis and its Applications*. Springer, New York

R. Christensen

Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox)

The study of relationships between two quantities has been important since earliest times: from the connection between weather and the hunted animals, down to cigarettes causing lung cancer. In the laboratory sciences it is possible to determine the relationship by controlling the conditions, changing one quantity and observing the effect of the change on the other. In the social and behavioral sciences this determination is not possible for three, related, reasons. First, conditions can rarely be controlled. Second, it is often not possible to set a quantity at an assigned value. Third, the relationship may not be exact but possess a random element, as with the heights of parents and their children. As a result, it is not easy to determine the nature of any possible relationship between quantities. Often an apparent connection has been found, on further study, to be spurious. We begin by looking at one of the more bizarre possibilities, usually known as Simpson's paradox (Simpson, 1951), but known to earlier writers. A more recent reference is Lindley and Novick (1981). Chapter 6 of Pearl (2000) contains a fine discussion in terms of causation. It is explained using some illustrative, medical data.

Patients were given either a treatment T, or a placebo T', to alleviate a condition. After a fixed time, it was observed whether they had recovered, R, or not, R'. Table 1 gives the data in the form of a 2×2 contingency table: thus 16 patients recovered, even with the placebo. The table also provides the marginal totals and the two recovery rates. The treatment appears to have been effective in increasing the recovery rate over the placebo by 10 percent. (As an aside, the numbers here are small, 80 patients in all; but the phenomenon could persist with 8,000. It is not the effect of small samples that is at issue.) There is an apparent relationship between treatment and recovery, but it has a random element and outside conditions have not been controlled.

One of these conditions is the patient's sex. Table 2, in the same form, gives the data for the men who participated in the study. Again the treatment has an effect but in the opposite direction, reducing the recovery rate by 10 percent. Common sense might suggest that a treatment which was overall beneficial, but harmful to men, would be highly beneficial for women. The data for them can be found by subtracting the cell entries in Table 2 for men from those for all patients in Table 1. Thus no new information is needed. The results are given in Table 3. Common sense misleads: the recovery rate for women also drops by 10 percent as a result of the treatment, just as with the men. This is a treatment which is bad for men, bad for women, but good for all of us. This is the paradox. It is here in an extreme form of complete reversal when an additional factor is included. It often arises in the form where there is some change in relationship when extra information is included. How can this happen?

It is easy to see what has gone wrong in the medical example. The condition is more serious for the women, with lower recovery rates than the men. Yet the treatment has been mainly given to the men—30 of them, against only 10 women. Perhaps the doctor was suspicious of the treatment and felt it was too

Table 1

	R	R′		Rate
T	20	20	40	50%
T′	16	24	40	40%
	36	44	80	

Table 2

Males	R	R′		Rate
T	18	12	30	60%
T′	7	3	10	70%
	25	15	40	

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7