4 Inferences about parameters in Normal Linear Models

4.1 Introduction

In the general *Normal Linear Model* of §3.3, an n-vector of responses $\mathbf{Y}$ follows the distribution $N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, given the value of an $n \times p$ matrix $\mathbf{X}$. Here $\mathbf{X}$ is assumed to have full rank, so that $\mathbf{X}^T\mathbf{X}$ is invertible, although this condition can be relaxed. We now consider making inferences about one or more of the coefficients in the p-vector β.

The Normality of $\mathbf{Y}$ given $\mathbf{X}$ implies that the least-squares estimator $\widehat{\beta}$ is also Normal: from (3.5.2) and (3.5.3), its distribution is $N_p(\beta, \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1})$. Also (from §3.8) the residual sum of squares (*RSS*) is independent of $\widehat{\beta}$ with the distribution $\sigma^2\chi^2(n-p)$. Hence, if $\widehat{\sigma}^2$ denotes the residual *mean* square (given by $RSS/(n-p)$), the corresponding estimator is unbiased for σ^2.

When the R regression function `lm` is used, an intercept is included in the model by default, i.e. $\mathbf{X}$ is assumed to include a column of 1's. To omit the intercept use `-1` in the formula, e.g. `lm(y~-1+x)`.

4.2 Single linear function of β

For inferences about a *single* linear function of β, such as β_1 or $\beta_1 - \beta_2$, we can use tests and confidence intervals based on Student's t. If $\mathbf{c}$ is a constant p-vector then the linear function

$$\mathbf{c}^T\beta = c_1\beta_1 + \ldots + c_p\beta_p \tag{4.2.1}$$

is estimated by $\mathbf{c}^T\widehat{\beta}$. From (3.5.4), the corresponding least squares estimator has estimated standard deviation

$$\widehat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}, \tag{4.2.2}$$

which is sometimes called its *estimated standard error*. Confidence intervals and tests of hypotheses for $\mathbf{c}^T\beta$ can thus be based on the random variable

$$\frac{\mathbf{c}^T\widehat{\beta} - \mathbf{c}^T\beta}{\widehat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}, \tag{4.2.3}$$

which has a t_{n-p} distribution. For example, if $t_{0.025}$ denotes the upper 2.5% point of this distribution, a 95% confidence interval for $\mathbf{c}^T\beta$ is given by

$$\mathbf{c}^T\widehat{\beta} \pm t_{0.025}\,\widehat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}. \tag{4.2.4}$$

4.2.1 Individual coefficients

Inference about an individual element of β, β_r say, corresponds to taking $\mathbf{c}^T$ equal to $(0 \ldots 0\,1\,0 \ldots 0)$ in (4.2.1) with $c_r = 1$, so that (4.2.2) gives the estimated standard error of $\widehat{\beta}_r$ as

$$\widehat{\sigma}\sqrt{r\text{-th diagonal element of } (\mathbf{X}^T\mathbf{X})^{-1}}. \tag{4.2.5}$$

The output from R's regression analysis `summary(lm(y~x))` includes the estimated standard error for each regression coefficient (called '`Std. Error`') with a corresponding t-statistic and significance probability for a two-sided test of the hypothesis that the true coefficient is zero. It is sometimes convenient to parametrize a linear model so that a linear function of interest (such as a difference between two regression slopes) becomes an individual parameter: see §4.3.3 for an example.

4.2.2 Future responses

If $\mathbf{x}_*$ denotes a future possible value of the vector of explanatory variables, then the corresponding expected response is $\mathrm{E}(Y\,|\,\mathbf{x}_*) = \mathbf{x}_*^T\beta$. From (4.2.2), the estimator $\mathbf{x}_*^T\widehat{\beta}$ of this expectation has estimated standard error

$$\widehat{\sigma}\sqrt{\mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}\,. \tag{4.2.6}$$

While, the estimated standard error for an *individual* future response is [following the same argument as in Question 2(b) of Problem Sheet 1],

$$\widehat{\sigma}\sqrt{1+\mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}\,. \tag{4.2.7}$$

In R the function `predict` can be used to calculate standard errors, as well as confidence intervals and prediction intervals,
e.g. for confidence intervals `predict(lm(y~x),se.fit = TRUE, interval = "confidence")`.

4.3 Single linear function of $\dot{\beta}$ when the model has an intercept

If the model includes an intercept and q explanatory variables, then the alternative formulation given in §3.9 is usually assumed, under which the explanatory variables are measured from their mean values. Thus we have

$$\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \gamma\mathbf{1}_n + \dot{\mathbf{X}}\dot{\beta}\,, \tag{4.3.1}$$

where $\dot{\mathbf{X}}$ has elements $x_{ij}-\overline{x}_j$. If $\dot{\mathbf{X}}$ has full rank then the residual sum of squares, the mean response $\widehat{\gamma}=\overline{Y}$ and the estimator $\widehat{\dot{\beta}}$ of $\dot{\beta}$ are independent of each other with respective distributions $\sigma^2\chi^2(n-q-1)$, $N(\gamma,n^{-1}\sigma^2)$ and $N_q(\dot{\beta},\sigma^2(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1})$. If $\mathbf{c}$ is now a constant q-vector and c_0 a scalar, then a linear function $c_0\gamma+\mathbf{c}^T\dot{\beta}$ has the estimator $c_0\widehat{\gamma}+\mathbf{c}^T\widehat{\dot{\beta}}$ with estimated standard error

$$\widehat{\sigma}\sqrt{n^{-1}c_0^2+\mathbf{c}^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\mathbf{c}}\,.$$

4.3.1 Individual coefficients

The estimated standard error of $\widehat{\beta}_r$ is expressible as

$$\widehat{\sigma}\sqrt{r\text{-th diagonal element of }(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}}\,. \tag{4.3.2}$$

4.3.2 Future responses

The expected response corresponding to a future possible value $\mathbf{x}_*$ of the q-vector of explanatory variables is $\mathrm{E}(Y\,|\,\mathbf{x}_*) = \gamma+(\mathbf{x}_*-\bar{\mathbf{x}})^T\dot{\beta}$. The estimated standard errors for the estimator $\widehat{\gamma}+(\mathbf{x}_*-\bar{\mathbf{x}})^T\widehat{\dot{\beta}}$ and for an *individual* future response are respectively

$$\widehat{\sigma}\sqrt{n^{-1}+(\mathbf{x}_*-\bar{\mathbf{x}})^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}(\mathbf{x}_*-\bar{\mathbf{x}})}\,, \tag{4.3.3}$$

$$\widehat{\sigma}\sqrt{1+n^{-1}+(\mathbf{x}_*-\bar{\mathbf{x}})^T(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}(\mathbf{x}_*-\bar{\mathbf{x}})}\,. \tag{4.3.4}$$

4.3.3 Two simple linear regressions

The model described in §3.2(d) for comparing two simple linear regressions is

$$\mathrm{E}(Y_i\,|\,x_i) = \alpha_1 + \beta_1 x_i \ (i = 1,\ldots,m), \quad \mathrm{E}(Y_i\,|\,x_i) = \alpha_2 + \beta_2 x_i \ (i = m+1,\ldots,n) \tag{4.3.5}$$

with $\mathrm{var}(Y_i\,|\,x_i) = \sigma^2$. This model does not include a common intercept, since there is no parameter included in all the expectations. It could be fitted in R by omitting the default intercept and using vectors $(1\ldots1\,0\ldots0)^T$, $(x_1\ldots x_m\,0\ldots0)^T$, $(0\ldots0\,1\ldots1)^T$ and $(0\ldots0\,x_{m+1}\ldots x_n)^T$. To compare the slopes of the two fitted lines, we would divide $\widehat{\beta}_1 - \widehat{\beta}_2$ by its estimated standard error and compare this t-statistic with $t(n-4)$.

It is more convenient to redefine the model, replacing (4.3.5) by, say,

$$\mathrm{E}(Y_i\,|\,x_i) = \alpha + \beta x_i \ (i = 1,\ldots,m), \quad \mathrm{E}(Y_i\,|\,x_i) = \alpha + \gamma + \beta x_i + \delta x_i \ (i = m+1,\ldots,n), \tag{4.3.6}$$

so that α, β replace α_1, β_1, and parameters γ and δ are respectively the difference between the intercepts $(\alpha_2 - \alpha_1)$ and the difference between the slopes $(\beta_2 - \beta_1)$. This model may be fitted by *including* the common intercept (corresponding to α) and using vectors $(x_1\ldots x_n)^T$, $(0\ldots0\,1\ldots1)^T$ and $(0\ldots0\,x_{m+1}\ldots x_n)^T$ (corresponding to β, γ and δ respectively). The t-statistic for the final vector tests the difference in the slopes. The difference in the intercepts could also be tested, but this is sensible only if the expected responses at $x = 0$ have particular significance.

If the final vector in the above model is omitted, we assume a model with a common slope β but different intercepts α and $\alpha+\gamma$, i.e. the two regression lines are assumed parallel. Under this model, the hypothesis of a common intercept (and hence a common regression) is tested using the t-statistic for γ, i.e. for the vector $(0\ldots0\,1\ldots1)^T$.

Example — Cavity-wall insulation

4.4 Tests of the hypotheses $\beta = \mathbf{0}$ and $\dot{\beta} = \mathbf{0}$

Suppose that we fit the linear model $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$, $\mathrm{var}(\mathbf{Y}\,|\,\mathbf{X}) = \sigma^2\mathbf{I}_n$ assuming that the response vector $\mathbf{Y}$ is Normally distributed, and then want to test the hypothesis that β equals $\mathbf{0}$, i.e. that $\beta_1 = \ldots = \beta_p = 0$ or $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{0}$. Separate Student-$t$ tests on the estimated coefficients do not provide an appropriate method, since they test only one coefficient at a time: instead we compare the *model sum of squares* with the *residual sum of squares* for the model.

From §3.8, the model SS is

$$\widehat{\beta}^T\mathbf{X}^T\mathbf{Y} = \mathbf{Y}^T\mathbf{P}_\mathbf{X}\mathbf{Y}, \tag{4.4.1}$$

and has the distribution $\sigma^2\chi^2(p, \sigma^{-2}\beta^T\mathbf{X}^T\mathbf{X}\beta)$; it is independent of the residual SS,

$$(n-p)\widehat{\sigma}^2 = \textstyle\sum_i Y_i^2 - \widehat{\beta}^T\mathbf{X}^T\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}, \tag{4.4.2}$$

which has the distribution $\sigma^2\chi^2(n-p)$. If β *is* equal to $\mathbf{0}$ then the model SS has the distribution $\sigma^2\chi^2(p)$, and the random variable

$$F = \frac{\text{model MS}}{\text{residual MS}} = \frac{\text{model SS}/p}{\widehat{\sigma}^2} \tag{4.4.3}$$

has the distribution $F(p, n-p)$. [If β does *not* equal $\mathbf{0}$ then the model SS has a *non-central* $\sigma^2\chi^2$ distribution with expectation *larger* than $p\sigma^2$.] If the hypothesis is false we expect *large* values for

the model SS and hence for F. So we can test the hypothesis that β is $\mathbf{0}$ by comparing the value of F in (4.4.3) with the upper percentage points of $F(p, n-p)$.

More usually, we consider the alternative formulation defined in §3.9, and want to test the model $\mathrm{E}(Y_i \,|\, \mathbf{X}) = \gamma$ (with no dependence on the explanatory variables) against the model

$$\mathrm{E}(Y_i \,|\, \mathbf{X}) = \gamma + \beta_1 (x_{i1} - \overline{x}_1) + \ldots + \beta_q (x_{iq} - \overline{x}_q) \quad (i = 1, \ldots, n), \tag{4.4.4}$$

which includes q explanatory variables. This is equivalent to testing the hypothesis that $\dot{\beta}$ is $\mathbf{0}$. From §3.9.2, the regression SS, $\widehat{\dot{\beta}}^T \dot{\mathbf{X}}^T \mathbf{Y}$, has distribution $\sigma^2 \chi^2(q, \sigma^{-2} \dot{\beta}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \dot{\beta})$, and is independent of the residual SS, which is distributed as $\sigma^2 \chi^2(n-q-1)$. Thus the expectation of the regression SS is larger than $q\sigma^2$ unless $\dot{\beta}$ is $\mathbf{0}$, so we test the hypothesis $\dot{\beta} = \mathbf{0}$ by comparing the value of

$$F = \frac{\text{regression MS}}{\text{residual MS}} \tag{4.4.5}$$

with the distribution $F(q, n-q-1)$. A *large* value provides evidence against the hypothesis.

4.5 Testing a linear hypothesis using the 'Extra Sum of Squares'

We now generalize from the hypotheses $\beta = \mathbf{0}$ and $\dot{\beta} = \mathbf{0}$ considered in §4.4 to a linear hypothesis H_0 specifying a set of c linear constraints on β or $\dot{\beta}$. In general, these have the form

$$\mathbf{C}\beta = \mathbf{d} \qquad \text{or} \qquad \mathbf{C}\dot{\beta} = \mathbf{d} \tag{4.5.1}$$

where $\mathbf{C}$ is a specified $c \times p$ (or $c \times q$) matrix of rank c and $\mathbf{d}$ is a specified c-vector. As in the comparison of two simple linear regressions in §4.3.3, we can redefine our linear model so that the c constraints specify the values of c individual parameters rather than of linear combinations.

To suggest the form of the statistic to be used for testing (4.5.1), first note that the null hypothesis in the test of $\dot{\beta} = \mathbf{0}$ is that the responses have a common expectation γ. The least squares estimate of γ under this hypothesis is $\overline{y}$, so the total SS about the mean, $\sum_i (y_i - \overline{y})^2$, is the residual SS under the *null* hypothesis. The regression SS is therefore the *increase* in the residual SS due to imposing the constraint that $\dot{\beta}$ is $\mathbf{0}$: this is called the *extra sum of squares* for this hypothesis, and the magnitude of this extra SS is compared with the residual SS using the F-statistic of (4.4.5).

We can extend the idea of calculating an F-statistic to compare the extra SS for a linear hypothesis with the residual SS to include the more general form of hypothesis H_0 given in (4.5.1). We calculate the statistic

$$F = \frac{(\text{extra SS for } \mathrm{H}_0)/c}{\text{residual MS}} \tag{4.5.2}$$

$$= \frac{(\text{residual SS under } \mathrm{H}_0 - \text{residual SS under full model})/c}{\text{residual MS}}, \tag{4.5.3}$$

and compare its value with the upper percentage points of $F(c, n-p)$ or $F(c, n-q-1)$; *large* values of F provide evidence against H_0. The distributions given for F are based on the results that

(a) the residual SS under the full model has the distribution $\sigma^2 \chi^2(n-p)$ or $\sigma^2 \chi^2(n-q-1)$;

(b) the extra SS is independent of the residual SS;

(c) the extra SS has the distribution $\sigma^2 \chi^2(c)$ under H_0 and is non-central $\sigma^2 \chi^2(c)$ otherwise.

Assertion (a) comes from §3.8 and §3.9.2. Assertions (b) and (c) are not proved in detail here, but an indication of the method of proof is as follows:

- The random c-vector $\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}$ has the distribution $N_c(\mathbf{C}\boldsymbol{\beta} - \mathbf{d}, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T)$ and is independent of the residual SS under the linear model of §3.3.
- The expectation of $\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}$ is $\mathbf{0}$ under H_0, so the quadratic form

$$(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d})^T \left\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\right\}^{-1} (\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}) \tag{4.5.4}$$

  has the distribution $\sigma^2\chi^2(c)$ independently of the residual SS under H_0.
- The realised value of the quadratic form in (4.5.4) is equal to the extra SS in (4.5.2).
- Under the alternative formulation for models with an intercept, $\dot{\mathbf{X}}$ and $\widehat{\dot{\boldsymbol{\beta}}}$ replace $\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}$.

4.5.1 Calculating an extra sum of squares as a quadratic form

To illustrate the use of the extra SS, we test the linear hypothesis $\beta_1 = \beta_2 = 1$ for a plum-leaf data example. [This hypothesis means that leaf area is approximately proportional to lamina length times breadth.] The residual MS is $\widehat{\sigma}^2 = 0.000176$ and $\left(\widehat{\beta}_1\ \widehat{\beta}_2\right) = (1.132 \quad 0.836)$, while $\mathbf{C} = \mathbf{I}_2$ and $\mathbf{d} = \mathbf{1}_2$ in (4.5.1). Thus

$$\mathbf{C}\widehat{\dot{\boldsymbol{\beta}}} - \mathbf{d} = (1.132 - 1 \quad 0.836 - 1)^T = (0.132 \quad -0.164)^T,$$

and (4.5.4) takes the value

$$\begin{aligned}(\mathbf{C}\widehat{\dot{\boldsymbol{\beta}}} - \mathbf{d})^T \left\{\mathbf{C}(\dot{\mathbf{X}}^T\dot{\mathbf{X}})^{-1}\mathbf{C}^T\right\}^{-1} (\mathbf{C}\widehat{\dot{\boldsymbol{\beta}}} - \mathbf{d}) &= \begin{pmatrix} 0.132 & -0.164 \end{pmatrix} \begin{pmatrix} 0.6615 & 0.6926 \\ 0.6926 & 0.8268 \end{pmatrix} \begin{pmatrix} 0.132 \\ -0.164 \end{pmatrix} \\ &= 0.0039.\end{aligned}$$

Hence the F-statistic for the hypothesis is

$$f = \frac{0.0039/2}{0.000176} = 11.06.$$

The 0.5% point of $F(2,12)$ is 8.51, so the data appear *not* to be consistent with this hypothesis.

4.5.2 Calculating an extra sum of squares as a difference of sums of squares

Rather than calculate the extra SS using (4.5.4), it is usually more convenient to find the extra SS for the hypothesis by fitting the full model and the model constrained by the hypothesis separately and finding the difference between the two residual SS.

For example, suppose a total of n responses Y_{jk} are independent with distributions $N(\mu_j, \sigma^2)$ $(j = 1,\ldots,g;\, k = 1,2,\ldots,n_j)$: this structure is sometimes called a *one-way classification*. We might want to test the hypothesis that the expected values $\mu_1,\ldots,\mu_g$ are equal. This amounts to imposing $g-1$ linear constraints on the μ_j, which might be written as $\mu_j - \mu_g = 0\ (j = 1,\ldots,g-1)$, so that $\boldsymbol{\beta} = \left(\mu_1 \ldots \mu_g\right)^T$, $\mathbf{d} = \mathbf{0}$ and $\mathbf{C}$ is $(g-1) \times g$ in (4.5.1).

Rather than find the extra SS from (4.5.4), we note that the residual SS under the full model and under the hypothesis of a common expectation μ, say, are respectively the within-groups SS, $\sum_j \sum_k (y_{jk} - \overline{y}_j)^2$ (with $n-g$ degrees of freedom), and the total SS about the mean, $\sum_j \sum_k (y_{jk} - \overline{y})^2$, so that the extra SS for the hypothesis is their difference, the between-groups SS, $\sum_j n_j (\overline{y}_j - \overline{y})^2$ (with $g-1$ df). The hypothesis is therefore tested by comparing

$$F = \frac{\text{between-groups MS}}{\text{within-groups MS}} = \frac{\text{between-groups SS} \,/\, (g-1)}{\text{within-groups SS} \,/\, (n-g)} \tag{4.5.5}$$

with the distribution $F(g-1, n-g)$.

To apply this analysis in R use `anova(y~as.factor(x))`, where `x` contains the levels of the factor and `y` is the response.

4.5.3 Calculating an extra sum of squares from sequential sums of squares

In R we might use `anova(lm(y~x1))` and then `anova(lm(y~x1),lm(y~x1+x2),lm(y~x1+x2+x3))` to give us 'sequential sums of squares' (*or* this may be read directly from the ANOVA table `anova(lm(y~x1+x2+x3))`). For regression on variables $x_1, \ldots, x_q$, say, these comprise

- the regression SS for fitting x_1 alone,
- the extra SS for fitting x_2 *as well as* x_1, equal to

$$\text{regression SS for fitting } x_1, x_2 \quad - \quad \text{regression SS for fitting } x_1\,,$$

- the extra SS for fitting x_3 in addition to x_1, x_2,

and so on. These can be used to find the extra SS for a hypothesis which sets some of the regression coefficients $\beta_1, \ldots, \beta_q$ to zero, i.e. which omits a subset of the explanatory variables from the model.

For example, if we fit a cubic polynomial to a set of n responses, we might test the hypothesis that a linear polynomial is adequate, putting the explanatory variables in the natural order x, x^2, x^3, and calculating the extra SS for the quadratic and cubic terms as the sum of their sequential SS.

4.5.4 Analysis of variance tables for tests of linear hypotheses

An analysis of variance (or ANOVA) table can be used to present the decomposition of the total SS about the mean into the regression SS and residual SS. This sort of table can be expanded to display the calculations required for a test of a linear hypothesis. The table below shows the general form for a test of such a hypothesis when (as is usual) the model includes an intercept which is not constrained by the hypothesis; the hypothesis is assumed to impose c linear constraints on $\dot{\beta}$ given by $\mathbf{C}\dot{\beta} = \mathbf{d}$, so that there are effectively only $q-c$ parameters in $\dot{\beta}$ to be estimated. Below the dashed line, the total SS about the mean is split into the usual regression SS and residual SS from fitting the unconstrained model. Above the line, this regression SS is itself split into the SS for fitting the model under the hypothesis and the extra SS arising from it. [Here $\widehat{\dot{\beta}}_c$ denotes the vector of estimates under the hypothesis.] The hypothesis can be tested by comparing

$$\frac{\text{hypothesis MS}}{\text{residual MS}} \tag{4.5.6}$$

with the distribution $F(c, n-q-1)$.

Source	DF	SS	MS
Regression under hypothesis	$q-c$	$\hat{\dot{\beta}}_c^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \hat{\dot{\beta}}_c$	
Deviations from hypothesis	c	extra SS	hypothesis MS
Regression for unconstrained model	q	$\hat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \hat{\dot{\beta}}$	
Residual for unconstrained model	$n-q-1$	$S_{yy} - \hat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \hat{\dot{\beta}}$	residual MS ($\hat{\sigma}^2$)
Total about mean	$n-1$	S_{yy}	

Under the more general linear model $\mathrm{E}(\mathbf{Y}\,|\,\mathbf{X}) = \mathbf{X}\beta$ with $\mathbf{X}$ $n \times p$ and of rank p, there is a similar analysis of variance table for a test of a hypothesis $\mathbf{C}\beta = \mathbf{0}$ in which $\mathbf{C}$ is $c \times p$ and has rank c. The raw total SS (with n df) is first decomposed into the residual SS (with $n-p$ df) and the model SS. The model SS is then split into the SS for the model under the hypothesis (with $p-c$ df) and the extra SS for the hypothesis (with c df).

For an example of this type of table, consider again the example of §4.5.3 in which a cubic polynomial is fitted to n responses, and we want to test the hypothesis that the data follow a linear polynomial. The analysis of variance table would then have the following form.

Source	DF	SS	MS
Linear regression on x	1	S_{xy}^2/S_{xx}	
Deviations from linear regression on x	2	extra SS	hypothesis MS
Cubic regression on x	3	$\hat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \hat{\dot{\beta}}$	
Residual from cubic regression on x	$n-4$	$S_{yy} - \hat{\dot{\beta}}^T \dot{\mathbf{X}}^T \dot{\mathbf{X}} \hat{\dot{\beta}}$	residual MS ($\hat{\sigma}^2$)
Total about mean	$n-1$	S_{yy}	

4.6 Testing the fit of a linear regression model with replicate data ('Lack of Fit' and 'Pure Error')

To further illustrate the test of a linear hypothesis and its analysis of variance table, suppose that responses Y_{jk} are independent with distributions $N(\mu_j, \sigma^2)$ $(j = 1, \ldots, g\,;\, k = 1, 2, \ldots, n_j)$ (one-way ANOVA model) and that there is an explanatory variable x taking the value x_j for group j. The hypothesis to be tested is that the responses have linear regression on x, so that, for some unknown values β_0 and β_1, the expectations $\mu_1, \ldots, \mu_g$ are related by

$$\mu_j = \beta_0 + \beta_1 x_j \quad (j = 1, \ldots, g)\,. \tag{4.6.1}$$

The residual SS under the full model is the within-groups SS. The residual SS under the constraints (4.6.1) equals the total SS about the mean minus the regression SS. The extra SS is thus

$$\begin{aligned} & (\text{within-groups SS} + \text{between-groups SS} - \text{regression SS}) - \text{within-groups SS} \\ = \;& \text{between-groups SS} - \text{regression SS}\,. \end{aligned}$$

This extra SS, with $g-2$ degrees of freedom, measures how much of the variation in the response is *not* explained by linear regression, i.e. the extent of deviations from linearity. The analysis of variance table has the following form.

Source	DF	SS	MS
Linear regression on x	1	$\left\{\sum_j n_j(x_j-\overline{x})\overline{y}_j\right\}^2 / \sum_j n_j(x_j-\overline{x})^2$	
Deviations from linearity	$g-2$	extra SS for non-linearity	hypothesis MS
Between groups	$g-1$	$\sum_j n_j(\overline{y}_j-\overline{y})^2$	
Within groups	$n-g$	$\sum_j \sum_k (y_{jk}-\overline{y}_j)^2$	within-groups MS
Total about mean	$n-1$	$\sum_j \sum_k (y_{jk}-\overline{y})^2$	

The hypothesis of linear regression is tested by comparing the ratio of the hypothesis MS to the within-groups MS with $F(g-2, n-g)$. Note that

(a) In this context the within-groups SS is called 'pure error', since its distribution does not depend on the linear regression assumption in (4.6.1).

(b) To apply this test in R, we can use `anova(lm(y~x),lm(y~as.factor(x)))`.

(c) The n_j responses at each x_j are called 'replicates'.

(d) The null hypothesis being tested is that the responses have linear regression on the x_j (with the alternative of arbitrary dependence on j) *not* that there is no dependence on x_j.

(e) Assuming that the regression model in (4.6.1) is true gives more precise inferences about the μ_j, but introduces bias if the model is wrong.