STATS 3001

STATISTICAL MODELLING III

Lecture Notes

Semester 1, 2018

Lecturer: Professor Patty Solomon

THE UNIVERSITY *of* ADELAIDE

STATS 3001 Statistical Modelling III

Semester 1, 2018

Contents

Lecture 1, 2018

1 Introduction

1.1 Notation

In this course we will make extensive use of random vectors and occasional use of random matrices. We will use the convention of representing random variables by uppercase letters, e.g. Y, and realisations of random variables by the corresponding lowercase letters, e.g. y. Throughout, we will consider variables for which means and variances exist.

Definition 1.1 *A random vector is a vector of random variables. For example,*

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

For the random vector $\boldsymbol{Y}$ we define the mean vector, $\boldsymbol{\eta}$, by

$$\boldsymbol{\eta} = E(\boldsymbol{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}.$$

The variance matrix is defined by

$$\mathrm{Var}(\boldsymbol{Y}) = \Sigma = [\sigma_{ij}]$$

where

$$\sigma_{ij} = \begin{cases} \mathrm{cov}(Y_i, Y_j) & \textit{for } i \neq j, \\ \mathrm{var}(Y_i) & \textit{for } i = j. \end{cases}$$

□

A random matrix can also be defined to be a matrix of random variables,

$$\boldsymbol{\mathcal{Y}} = [Y_{ij}]$$

and we will use the convention

$$E(\boldsymbol{\mathcal{Y}}) = [E(Y_{ij})].$$

Note that we will not need to define the variance structure for random matrices.

We will use the following result concerning linear transformation of random vectors and matrices.

Lemma 1.1 *Suppose $\boldsymbol{Y}$ is a random vector with $E(\boldsymbol{Y}) = \boldsymbol{\eta}$ and $\mathrm{Var}(\boldsymbol{Y}) = \Sigma$ and let $A_{m\times n}$ and $\boldsymbol{b}_{m\times 1}$ be fixed. Then*

$$E(A\boldsymbol{Y} + \boldsymbol{b}) = A\boldsymbol{\eta} + \boldsymbol{b} \text{ and } \mathrm{Var}(A\boldsymbol{Y} + \boldsymbol{b}) = A\Sigma A^T.$$

If $\boldsymbol{\mathcal{Y}}$ is a random matrix and A is a fixed matrix then

$$E(A\boldsymbol{\mathcal{Y}}) = AE(\boldsymbol{\mathcal{Y}}).$$

□

In this course, we will use the notation,

$$\boldsymbol{Y} \sim N_r(\boldsymbol{\mu}, \Sigma)$$

to indicate that the r-dimensional random vector $\boldsymbol{Y}$ has the r-dimensional *multivariate normal distribution* with mean vector $\boldsymbol{\mu}$ and variance matrix Σ. Note that a detailed discussion of the multivariate normal distribution is given in Mathematical Statistics III. We will use the following results.

Lemma 1.2 *If $\boldsymbol{Y} \sim N_r(\boldsymbol{\mu}, \Sigma)$ and $A_{k\times r}$ and $\boldsymbol{b}_{k\times 1}$ are fixed then*

$$A\boldsymbol{Y} + \boldsymbol{b} \sim N_k(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T).$$

If $\boldsymbol{Y} \sim N_r(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{a}_{r\times 1}$ is fixed, then

$$\boldsymbol{a}^T\boldsymbol{Y} \sim N(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\Sigma\boldsymbol{a}).$$

□

1.2 Multiple regression

The regression model is used to model the dependence between a predictor variable x and a response variable Y. In general, there may be several predictor variables $x_1, x_2, \ldots, x_r$ and single response Y. In this case the *multiple regression* model may be used to model the simultaneous influence of the predictors.

Consider data,

$$(y_1, x_{11}, x_{12}, \ldots, x_{1r}),\ (y_2, x_{21}, x_{22}, \ldots, x_{2r}),\ \ldots,\ (y_n, x_{n1}, x_{n2}, \ldots, x_{nr}).$$

The multiple regression model postulates

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir} + e_i$$

where $e_1, e_2, \ldots, e_n$ are realisations of independent random variables $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ with

$$E(\mathcal{E}_i) = 0 \text{ and } \mathrm{var}(\mathcal{E}_i) = \sigma^2.$$

Often, we assumed in addition that the errors are normally distributed. That is, $\mathcal{E}_i$ are *iid* $N(0, \sigma^2)$.

For the purpose of discussing statistical inference it is important to maintain the distinction between the data and the model and also between random variables and realisations of random variables. Nevertheless, it is often convenient to simply write the regression model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir} + e_i$$

with $e_1, e_2, \ldots, e_n$ *iid* $N(0, \sigma^2)$ as an abbreviation for the random variable formulation given above. This convention will be used where appropriate in this course. In some theoretical contexts, it is more convenient to use random variable notation

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir} + \mathcal{E}_i$$

and this notation will also be used when needed.

For many purposes it is most convenient to use the following matrix formulation of the multiple regression model. Let

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1r} \\ 1 & x_{21} & x_{22} & \ldots & x_{2r} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nr} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} \text{ and } \boldsymbol{\mathcal{E}} = \begin{pmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_n \end{pmatrix}.$$

The multiple regression model can then be formulated as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$$

or, in terms of random variables,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$$

with

$$E(\boldsymbol{\mathcal{E}}) = \mathbf{0} \text{ and } \operatorname{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I_{n \times n}.$$

The additional assumption of normality is then formulated as

$$\boldsymbol{\mathcal{E}} \sim N_n(\mathbf{0}, \sigma^2 I).$$

Formulation of the multiple regression model entails certain technical considerations.

Definition 1.2 *A set of vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p\}$ is said to be* ***linearly independent*** *if*

$$\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \ldots + \alpha_p \boldsymbol{v}_p = \mathbf{0} \quad \Rightarrow \quad \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0.$$

Otherwise it is said to be ***linearly dependent****.* □

Remark When $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ are linearly dependent it means that one of the $\boldsymbol{v}_i$'s is expressible as a linear combination of the remaining $\boldsymbol{v}$'s.

Consider the multiple regression model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}.$$

We require that the columns of X be linearly independent. To see why this is necessary, suppose the columns were linearly dependent. Then we could find a non-zero vector

$$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_r)^T$$

such that

$$X\boldsymbol{\alpha} = \mathbf{0}.$$

In this case, $\boldsymbol{\beta}$ would not uniquely identified since we would have

$$X\boldsymbol{\beta} = X(\boldsymbol{\beta} + \boldsymbol{\alpha}).$$

On the other hand, if the columns of X are linearly independent, we have

$$X\boldsymbol{\alpha} = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{\alpha} = \mathbf{0}$$

so $\boldsymbol{\beta}$ is uniquely identified.

Lecture 2, 2018

2 The linear regression model

2.1 Least squares estimation

Definition 2.1

1. *The least squares estimate,* $\hat{\boldsymbol{\beta}}$ *is the vector that minimises the sum of squares*

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

2. *The variance* σ^2 *is estimated by*

$$\begin{aligned} s_e^2 &= \frac{1}{n-p}\sum_{i=1}^{n}\{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_r x_{ir})\}^2 \\ &= \frac{1}{n-p}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2. \end{aligned}$$

where $p = r + 1$ *is the number of columns of* X*.*

□

Theorem 2.1 *If the columns of* X *are linearly independent then the least squares estimates are given uniquely by*

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

This result is proved in Statistical Modelling and Inference II. □

The vector of fitted values is defined by

$$\hat{\boldsymbol{\eta}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \boldsymbol{y} = P\boldsymbol{y}$$

where $P = X(X^T X)^{-1} X^T$. The alternative notation $H = X(X^T X)^{-1} X^T$ is sometimes used. For reasons to be discussed later, the $n \times n$ matrix P is called an orthogonal projection matrix. The elementary statistical properties of $\hat{\boldsymbol{\beta}}$ and s_e^2 are summarised in Theorem 2.2

Theorem 2.2 *Suppose*

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$$

where

$$E(\boldsymbol{\mathcal{E}}) = \mathbf{0} \text{ and } \mathrm{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I.$$

1. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
2. $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$
3. $E(s_e^2) = \sigma^2$.

Moreover, if $\boldsymbol{\mathcal{E}} \sim N_n(\mathbf{0}, \sigma^2 I)$, then:

4. $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$.

5. $\dfrac{(n-p)s_e^2}{\sigma^2} \sim \chi^2_{n-p}$ *independently of* $\hat{\boldsymbol{\beta}}$.

Proof:

1.

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((X^T X)^{-1} X^T \boldsymbol{Y}) \\ &= (X^T X)^{-1} X^T E(\boldsymbol{Y}) \\ &= (X^T X)^{-1} X^T X \boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

2.

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}((X^T X)^{-1} X^T \boldsymbol{Y}) \\ &= (X^T X)^{-1} X^T \,\mathrm{Var}(\boldsymbol{Y}) \{(X^T X)^{-1} X^T\}^T \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

3. Observe first that if $P = X(X^T X)^{-1} X^T$, then

   (a) $P^2 = P^T = P$;

   (b) $(I-P)^2 = (I-P)^T = I - P$;

   (c) If $\boldsymbol{\eta} = E(\boldsymbol{Y}) = X\boldsymbol{\beta}$ then $(I-P)\boldsymbol{\eta} = \mathbf{0}$.

   Next, observe

$$\begin{aligned} (n-p)s_e^2 &= \|(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\|^2 \\ &= \|(I - X(X^T X)^{-1} X^T)\boldsymbol{Y}\|^2 \\ &= \|(I-P)\boldsymbol{Y}\|^2 \\ &= \|(I-P)(\boldsymbol{Y} - \boldsymbol{\eta})\|^2 \\ &= \{(I-P)(\boldsymbol{Y} - \boldsymbol{\eta})\}^T \{(I-P)(\boldsymbol{Y} - \boldsymbol{\eta})\} \\ &= (\boldsymbol{Y} - \boldsymbol{\eta})^T (I-P)^T (I-P)(\boldsymbol{Y} - \boldsymbol{\eta}) \\ &= (\boldsymbol{Y} - \boldsymbol{\eta})^T (I-P)(\boldsymbol{Y} - \boldsymbol{\eta}) \\ &= \mathrm{tr}\{(\boldsymbol{Y} - \boldsymbol{\eta})^T (I-P)(\boldsymbol{Y} - \boldsymbol{\eta})\} \\ &= \mathrm{tr}\{((I-P)(\boldsymbol{Y} - \boldsymbol{\eta})(\boldsymbol{Y} - \boldsymbol{\eta})^T\}. \end{aligned}$$

Hence,

$$\begin{aligned} E\left((n-p)s_e^2\right) &= E\left(\mathrm{tr}\{((I-P)(\boldsymbol{Y}-\boldsymbol{\eta})(\boldsymbol{Y}-\boldsymbol{\eta})^T\}\right) \\ &= \mathrm{tr}\{((I-P)E\left((\boldsymbol{Y}-\boldsymbol{\eta})(\boldsymbol{Y}-\boldsymbol{\eta})^T\right)\} \\ &= \mathrm{tr}\{(I-P)\sigma^2 I\} \\ &= \sigma^2\,\mathrm{tr}\{I-P\} \\ &= \sigma^2\{\mathrm{tr}(I)-\mathrm{tr}(P)\}. \end{aligned}$$

Finally, observe $\mathrm{tr}(I) = n$ and

$$\mathrm{tr}(P) = \mathrm{tr}\{X(X^TX)^{-1}X^T\} = \mathrm{tr}\{(X^TX)^{-1}X^TX\} = \mathrm{tr}(I_{p\times p}) = p$$

so that

$$E\left((n-p))s_e^2\right) = (n-p)\sigma^2 \text{ and hence } E\left(s_e^2\right) = \sigma^2$$

as required.

4. Follows from Lemma 1.2, using the same calculations as in 1, and 2.
5. Omitted. □

Lecture 3, 2018

2.2 Estimable functions and best linear unbiased estimates

Theorem 2.2 describes the sampling distribution of the least squares estimates. The least squares estimate $\hat{\boldsymbol{\eta}}$ is also the best linear unbiased estimate (BLUE) of $\boldsymbol{\eta}$. To explain this property, we begin with the following definition.

Definition 2.2 *A quantity of the form $\boldsymbol{\lambda}^T\boldsymbol{\eta}$ where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T$ is a fixed vector is said to be an estimable function.*

□

Examples Consider the multiple regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$ where the columns of X are assumed to be linearly independent.

1. Each individual regression coefficient β_j is an estimable function. To see this, take

$$\boldsymbol{\lambda}^T = (0, \ldots, 0, 1, 0, \ldots, 0)(X^TX)^{-1}X^T.$$

2. The point prediction, $\beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r$ is an estimable function for any fixed values $x_1, x_2, \ldots, x_r$. To see this, take

$$\boldsymbol{\lambda}^T = (1, x_1, x_2, \ldots, x_r)(X^TX)^{-1}X^T.$$

Theorem 2.3 *(The Gauss-Markov Theorem)*
Suppose $E(\boldsymbol{Y}) = \boldsymbol{\eta} = X\boldsymbol{\beta}$ and $\mathrm{Var}(\boldsymbol{Y}) = \sigma^2 I$. If $\boldsymbol{a}^T\boldsymbol{Y}$ is an unbiased linear estimator for $\boldsymbol{\lambda}^T\boldsymbol{\eta}$ then

$$\mathrm{var}(\boldsymbol{a}^T\boldsymbol{Y}) \geq \mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}})$$

with equality if and only if

$$\boldsymbol{a} = X(X^TX)^{-1}X^T\boldsymbol{\lambda}.$$

Proof Observe first that for $\boldsymbol{\eta} = X\boldsymbol{\beta}$,

$$E(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) = \boldsymbol{\lambda}^T\boldsymbol{\eta} \text{ and } \mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) = \sigma^2\boldsymbol{\lambda}^T P\boldsymbol{\lambda}.$$

Next, observe that $\boldsymbol{a}^T\boldsymbol{Y}$ unbiased for $\boldsymbol{\lambda}^T\boldsymbol{\eta}$

$$\begin{aligned}
&\Leftrightarrow \quad E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{\lambda}^T\boldsymbol{\eta} \\
&\Leftrightarrow \quad (\boldsymbol{a} - \boldsymbol{\lambda})^T\boldsymbol{\eta} = 0 \text{ for all } \boldsymbol{\eta} = X\boldsymbol{\beta} \\
&\Leftrightarrow \quad (\boldsymbol{a} - \boldsymbol{\lambda})^T X = \boldsymbol{0} \\
&\Leftrightarrow \quad (\boldsymbol{a} - \boldsymbol{\lambda})^T P = \boldsymbol{0} \text{ where } P = X(X^TX)^{-1}X^T
\end{aligned}$$

and

$$\mathrm{var}(\boldsymbol{a}^T\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}} + \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) = \mathrm{var}(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) + \mathrm{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) + 2\,\mathrm{cov}(\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}, \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}).$$

Now,

$$\begin{aligned}\operatorname{cov}(\boldsymbol{a}^T\boldsymbol{Y}-\boldsymbol{\lambda}^{\boldsymbol{T}}\hat{\boldsymbol{\eta}},\boldsymbol{\lambda}^{\boldsymbol{T}}\hat{\boldsymbol{\eta}}) &= \operatorname{cov}\left((\boldsymbol{a}^T-\boldsymbol{\lambda}^TP)\boldsymbol{Y},\boldsymbol{\lambda}^TP\boldsymbol{Y}\right)\\ &= (\boldsymbol{a}^T-\boldsymbol{\lambda}^TP)\sigma^2IP\boldsymbol{\lambda}\\ &= \sigma^2(\boldsymbol{a}^T-\boldsymbol{\lambda}^T)P\boldsymbol{\lambda}=0.\end{aligned}$$

Hence,

$$\begin{aligned}\operatorname{var}(\boldsymbol{a}^T\boldsymbol{Y}) &= \operatorname{var}(\boldsymbol{a}^T\boldsymbol{Y}-\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}})+\operatorname{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}})\\ &\geq \operatorname{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}})\end{aligned}$$

with equality if and only if $\operatorname{var}(\boldsymbol{a}^T\boldsymbol{Y}-\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}})=0$. That is, if and only if, $\boldsymbol{a}=P\boldsymbol{\lambda}$. □

2.3 Inference for multiple regression

2.3.1 Inference for regression coefficients

Theorem 2.2 allows for inference concerning individual regression coefficients. Suppose the multiple regression model $M:\boldsymbol{Y}=X\boldsymbol{\beta}+\boldsymbol{\mathcal{E}}$ with $\boldsymbol{\mathcal{E}}\sim N_n(\mathbf{0},\sigma^2I)$ holds. A *pivotal quantity* for β_j is given by

$$\frac{\hat{\beta}_j-\beta_j}{s_e\sqrt{c_{jj}}}\sim t_{n-p}$$

where c_{jj} is the jth diagonal element of $(X^TX)^{-1}$. Hence,

$$\hat{\beta}_j\pm t_{n-p}(\alpha/2)s_e\sqrt{c_{jj}} \qquad (1)$$

is a $100(1-\alpha)\%$ confidence interval for β_j. To test $H_0:\beta_j=0$ we use the test statistic

$$t=\frac{\hat{\beta}_j-0}{s_e\sqrt{c_{jj}}} \qquad (2)$$

and the rule,

Reject H_0 for $|t|\geq t_{n-p}(\alpha/2)$,

to define a two-sided test with significance level α. The P-value can be calculated in the usual way.

Hypotheses of the form

$$H_0:\beta_p=\beta_{p-1}=\ldots=\beta_{p_0+1}=0$$

may be tested using an F-test. The ANOVA table is

Source	SS	DF	MS	F
H_0 vs M	$Q(\hat{\boldsymbol{\beta}}_0)-Q(\hat{\boldsymbol{\beta}})$	$p-p_0$	$MS_H=\frac{Q(\hat{\boldsymbol{\beta}}_0)-Q(\hat{\boldsymbol{\beta}})}{p-p_0}$	$F=\frac{MS_H}{MS_E}$
Error	$Q(\hat{\boldsymbol{\beta}})$	$n-p$	$MS_E=\frac{Q(\hat{\boldsymbol{\beta}})}{n-p}$	
Total	$Q(\hat{\boldsymbol{\beta}}_0)$	$n-p_0$		

If H_0 is true then F has the F-distribution, $F \sim F_{p-p_0,n-p}$ so we reject H_0, with significance level α, if $F \geq F_{p-p_0,n-p}(\alpha)$. The P-value can also be calculated in the usual way.

Remarks

1. $\hat{\boldsymbol{\beta}}_0$ is obtained by applying least squares calculation to the reduced regression model

$$\boldsymbol{Y} = X_0\boldsymbol{\beta}_0 + \boldsymbol{\mathcal{E}}$$

   where X_0 is the $n \times p_0$ matrix comprising the first p_0 columns of X. That is, we have $\hat{\boldsymbol{\beta}}_0 = (X_0^T X_0)^{-1} X_0^T \boldsymbol{Y}$.

2. It can be proved that the F-statistic is given equivalently by

$$F = \frac{\hat{\boldsymbol{\beta}}_1^T C_{11}^{-1} \hat{\boldsymbol{\beta}}_1}{(p - p_0)s_e^2}$$

   where $\hat{\boldsymbol{\beta}}_1^T = (\hat{\beta}_{p_0+1}, \ldots, \hat{\beta}_p)$ and C_{11} is the lower right $(p - p_0) \times (p - p_0)$ block of $(X^T X)^{-1}$.

3. This form of the F-statistic reduces to $F = t^2$ when the test concerns a single parameter, $H_0 : \beta_p = 0$.

Lecture 4, 2018

2.3.2 Prediction

Consider the multiple regression model $M : \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$ with $\boldsymbol{\mathcal{E}} \sim N_n(\mathbf{0}, \sigma^2 I)$ and suppose a new observation $(Y_0, \boldsymbol{x}_0)$ is to be made independently with $Y_0 \sim N(\boldsymbol{x}_0^T\boldsymbol{\beta}, \sigma^2)$. As in the case of simple linear regression, there are two problems associated with prediction.

1. Obtain a $100(1-\alpha)\%$ confidence interval for $\eta_0 = \boldsymbol{x}_0^T\boldsymbol{\beta}$.

   Solution: It follows from the Gauss-Markov theorem that $\hat{\eta}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ is the BLUE for η_0. Moreover, since

   $$\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}\right),$$

   it can be shown that

   $$\hat{\eta_0} \sim N(\eta_0, \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)$$

   independently of s_e^2. The pivotal quantity is

   $$\frac{\hat{\eta}_0 - \eta_0}{s_e\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}} \sim t_{n-p}.$$

   The $100(1-\alpha)\%$ confidence interval for η_0 is therefore

   $$\hat{\eta}_0 \pm t_{n-p}(\alpha/2)s_e\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}$$

2. Obtain a $100(1-\alpha)\%$ prediction interval for Y_0.

   Solution: Since $Y_0 \sim N\left(\eta_0, \sigma^2\right)$ independently of $\boldsymbol{Y}$ it follows that Y_0 is independent of $\hat{\eta}_0$ so that

   $$Y_0 - \hat{\eta}_0 \sim N\left(0, \sigma^2(1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0\right)$$

   and hence

   $$\frac{Y_0 - \hat{\eta}_0}{s_e\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}} \sim t_{n-p}.$$

   The $100(1-\alpha)\%$ prediction interval for Y_0 is therefore

   $$\hat{\eta}_0 \pm t_{n-p}(\alpha/2)s_e\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}$$

Remark: The confidence interval in (1) is simply a confidence interval for the estimable function $\eta_0 = \boldsymbol{x}_0^T\boldsymbol{\beta}$. The prediction interval is different in nature in the sense that it is predicting a likely range of values for the random variable Y_0. If the parameters $\boldsymbol{\beta}$ and σ^2 were known without error, it would still be relevant to consider the question of prediction and resulting interval would simply be

$$\eta_0 \pm z(\alpha/2)\sigma.$$

The interval given above differs from this simple form because the parameters have been replaced by estimates and adjustments to allow for this fact have been made. In particular, we use $t_{n-p}(\alpha/2)$ instead of $z(\alpha/2)$ to allow for the fact that we are using s_e in place of σ. The term $\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0$ is introduced to allow for errors in the estimation of η_0 by $\hat{\eta}_0$.

2.4 Symbolic specification of linear models

Most high-level statistical packages provide for the automatic construction of the model matrix X in the formulation of the linear model, $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$. We will consider the system of model formulæ implemented in R.

2.4.1 Multiple regression models

Recall that the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

can be fitted in R using the lm function. In particular, we use a command of the form lm(y~x) where y is a vector containing the y-values and x is a vector containing the x-values. The first argument of the lm function is the string "y~x" and is called the model formula. In general, a single term on the left hand side of the ~ operator is used to specify the response variable. The expression on the right hand side is expanded and used to construct the *model matrix*.

As a simple illustration, consider the following example in R. Note that the model.matrix function can be used to display the model matrix that is actually used in the regression calculations. Note also, that the column labels displayed in the model.matrix output correspond to coefficient labels in the regression summary.

```
x<-1:5
y<-c(0.1339862,-0.6302166,0.8446705,-0.1718445,-0.3497024)
x

## [1] 1 2 3 4 5

y

## [1]  0.1339862 -0.6302166  0.8446705 -0.1718445 -0.3497024

model.matrix(~x)

##   (Intercept) x
## 1           1 1
## 2           1 2
## 3           1 3
## 4           1 4
## 5           1 5
## attr(,"assign")
## [1] 0 1
```

```
summary(lm(y~x))
```

```
## 
## Call:
## lm(formula = y ~ x)
## 
## Residuals:
##        1        2        3        4        5 
##  0.06681 -0.64650  0.87929 -0.08632 -0.21328 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1181     0.6766   0.175    0.873
## x            -0.0509     0.2040  -0.250    0.819
## 
## Residual standard error: 0.6451 on 3 degrees of freedom
## Multiple R-squared:  0.02033,Adjusted R-squared:  -0.3062 
## F-statistic: 0.06225 on 1 and 3 DF,  p-value: 0.8191
```

Multiple regression models can be specified by including additional terms in the right hand side of the model formula, separated by the + symbol. For example,

```
x1<-1:5
x2<-seq(3,11,by=2)
x1
```

```
## [1] 1 2 3 4 5
```

```
x2
```

```
## [1]  3  5  7  9 11
```

```
model.matrix(~x1+x2)
```

```
##   (Intercept) x1 x2
## 1           1  1  3
## 2           1  2  5
## 3           1  3  7
## 4           1  4  9
## 5           1  5 11
## attr(,"assign")
## [1] 0 1 2
```

or `FVC~Height+Weight`. Functions can be applied directly to any terms in the model formula. For, example `log(FVC)~Height+Weight` or `log(FVC)~log(Height)+Weight`.

Arithmetic calculations, involving addition, subtraction, multiplication, division can also be incorporated into the model formula but must be enclosed within the `I` (insulate) function. For example, to fit a cubic regression model, we use `y~x+I(x^2)+I(x^3)` and not `y~x+x^2+x^3`. The reason is that the operators +, -, *, / and ^ have different interpretations (that will be explained later) in the context of a model formula.

Lecture 5, 2018

2.4.2 Factors

A categorical predictor variable is called a *factor*.

Example In a certain experiment, animals were exposed to one of three different poisons and four different treatments and the survival time recorded for each animal. Four animals were observed at each treatment/poison combination so that the experiment would be described as a 3×4 factorial experiment with replication 4. Box and Cox(1964) have previously determined that if

$$y = \frac{1}{\text{Survival Time}}$$

then the additive model

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

with $i \equiv$ poison, $j \equiv$ treatment and $k \equiv$ replicate, provides an adequate description of the data. □

The additive formulation shown above cannot be used directly because the parameters α_i and β_j are not identifiable. That is, if we tried to construct the model matrix X, the columns would not be linearly independent. For this reason, it is necessary to impose certain constraints on the parameters. The classical approach is to impose the side conditions,

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = 0.$$

For most computing packages (including R) the default parameterisation is to use the *treatment contrasts* defined by

$$\alpha_1 = \beta_1 = 0.$$

It can be checked mathematically that both approaches are equivalent in terms of the way that the model constrains the fitted values η_{ijk}. For practical purposes it is essential to understand that the interpretation of the numerical parameter estimates depends upon the constraints used.

Zero Sum Constraints: α_i indicates how far the mean for treatment i lies above or below the average for all treatments.

Treatment Contrasts: α_i how far the mean for treatment i lies above or below the mean for the reference level, treatment 1.

For models involving two or more factors, it is necessary to consider also *interactions*. For example, in the poison data, the most general model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \mathcal{E}_{ijk}$$

where

$$\alpha_1 = \beta_1 = \gamma_{i1} = \gamma_{1j} = 0.$$

The term γ_{ij} is called the interaction and allows for possible non-additive treatment and poison effects.

Models involving factors, such as the two factor model with and without interaction, are all expressible in the general form

$$\boldsymbol{\eta} = X\boldsymbol{\beta}.$$

The system of model formulæ implemented in `R` provides for the automatic generation of blocks of columns in the model matrix corresponding to individual factors (main effects) and interactions. Factors may be entered into `R` either as character data where the possible values are the names of the factor levels or as numerical data, where the numbers are codes for the factor levels. Such variables must be declared as factors using the `factor` command.

```
# Generate some artificial data
A<-rep(c("Level1","Level2","Level3"),3)
B<-rep(c(1,2,3),c(3,3,3))
A

## [1] "Level1" "Level2" "Level3" "Level1" "Level2" "Level3" "Level1" "Level2" "Level3"

B

## [1] 1 1 1 2 2 2 3 3 3

# See what happens if B is not declared as a factor.
model.matrix(~B)

##   (Intercept) B
## 1           1 1
## 2           1 1
## 3           1 1
## 4           1 2
## 5           1 2
## 6           1 2
## 7           1 3
## 8           1 3
## 9           1 3
## attr(,"assign")
## [1] 0 1

# Now define A and B as factors
A<-factor(A)
B<-factor(B)
model.matrix(~B)

##   (Intercept) B2 B3
## 1           1  0  0
## 2           1  0  0
## 3           1  0  0
## 4           1  1  0
## 5           1  1  0
## 6           1  1  0
## 7           1  0  1
## 8           1  0  1
## 9           1  0  1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$B
## [1] "contr.treatment"
```

2.4.3 Model formula operators

In this section, the R model formula operators +, :, $*$, ^, / and - are discussed. For the purpose of illustration, we consider two factors A, B and a covariate, x as shown below.

```
  A B x
1 1 1 1
2 1 2 2
3 1 3 3
4 2 1 4
5 2 2 5
6 2 3 6
7 3 1 7
8 3 2 8
9 3 3 9
```

The + operator:

The addition operator operator is used to add extra terms to a model. Its action is to concatenate the columns corresponding to the expressions on either side of the + operator. For example, the model

$$\eta_{ij} = \mu + \alpha_i + \beta_j$$

with $\alpha_1 = \beta_1 = 0$ is represented as ~A+B.

```r
model.matrix(~A)
```

```
##   (Intercept) A2 A3
## 1           1  0  0
## 2           1  0  0
## 3           1  0  0
## 4           1  1  0
## 5           1  1  0
## 6           1  1  0
## 7           1  0  1
## 8           1  0  1
## 9           1  0  1
```

```r
model.matrix(~B)
```

```
##   (Intercept) B2 B3
## 1           1  0  0
## 2           1  1  0
## 3           1  0  1
## 4           1  0  0
## 5           1  1  0
## 6           1  0  1
## 7           1  0  0
## 8           1  1  0
## 9           1  0  1
```

```
model.matrix(~A+B)
```

```
##   (Intercept) A2 A3 B2 B3
## 1           1  0  0  0  0
## 2           1  0  0  1  0
## 3           1  0  0  0  1
## 4           1  1  0  0  0
## 5           1  1  0  1  0
## 6           1  1  0  0  1
## 7           1  0  1  0  0
## 8           1  0  1  1  0
## 9           1  0  1  0  1
```

Similarly, the parallel regression model

$$\eta_{ij} = \mu + \alpha_i + \beta x_{ij}$$

with $\alpha_1 = 0$ is represented by `~A+x`.

```
model.matrix(~A+x)
```

```
##   (Intercept) A2 A3 x
## 1           1  0  0 1
## 2           1  0  0 2
## 3           1  0  0 3
## 4           1  1  0 4
## 5           1  1  0 5
## 6           1  1  0 6
## 7           1  0  1 7
## 8           1  0  1 8
## 9           1  0  1 9
```

Finally, it should be noted that the `+` operator automatically eliminates terms that are algebraically redundant. For example, `~A+A` is automatically recognised and reduced to `~A`.

```
model.matrix(~A+A)
```

```
##   (Intercept) A2 A3
## 1           1  0  0
## 2           1  0  0
## 3           1  0  0
## 4           1  1  0
## 5           1  1  0
## 6           1  1  0
## 7           1  0  1
## 8           1  0  1
## 9           1  0  1
```

The : operator:

The interaction operator, : is used to generate interactions between factors or between factors and covariates. For example, the two-factor model with interaction

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

with $\alpha_1 = \beta_1 = \gamma_{i1} = \gamma_{1j} = 0$ is represented by `~A+B+A:B`.

```
model.matrix(~A+B+A:B)
```

```
##   (Intercept) A2 A3 B2 B3 A2:B2 A3:B2 A2:B3 A3:B3
## 1           1  0  0  0  0     0     0     0     0
## 2           1  0  0  1  0     0     0     0     0
## 3           1  0  0  0  1     0     0     0     0
## 4           1  1  0  0  0     0     0     0     0
## 5           1  1  0  1  0     1     0     0     0
## 6           1  1  0  0  1     0     0     1     0
## 7           1  0  1  0  0     0     0     0     0
## 8           1  0  1  1  0     0     1     0     0
## 9           1  0  1  0  1     0     0     0     1
```

The separate regressions model,

$$\eta_{ij} = \mu + \alpha_i + \beta x_{ij} + \beta_i x_{ij}$$

with $\alpha_1 = \beta_1 = 0$ is represented by `A+x+A:x`.

```
model.matrix(~A+x+A:x)
```

```
##   (Intercept) A2 A3 x A2:x A3:x
## 1           1  0  0 1    0    0
## 2           1  0  0 2    0    0
## 3           1  0  0 3    0    0
## 4           1  1  0 4    4    0
## 5           1  1  0 5    5    0
## 6           1  1  0 6    6    0
## 7           1  0  1 7    0    7
## 8           1  0  1 8    0    8
## 9           1  0  1 9    0    9
```

Note that this model allows for regressions with different slopes and intercepts within each group. It is parameterised so that β is the slope for group 1, β_2 is the difference in slopes for group 2 and group 1, and β_3 is the difference in slopes for group 3 and group 1. The hypothesis $H_0 : \beta_2 = \beta_3 = 0$ is therefore the hypothesis of parallel regressions.

In general the `A:B` operation works by pointwise multiplication of each column associated with `A` with each column associated with `B`. In the preceding example, where `A` and `B` are both factors with three levels, there are two columns associated with each and therefore 4 columns associated with the interaction term. The same principle can also be seen to apply to the parallel regression example. The interaction operator can be applied to an arbitrary number of terms.

As with the + operator, the : operator also recognises and eliminates certain redundancies. For example, `~A:A` is automatically reduced to `~A` and `~A:B:A` is reduced to `~A:B`.

The $*$ operator:
The interaction operator : is not usually used directly as it is laborious to include all main effects and interactions explicitly. The $*$ operator is used to generate those terms automatically. For example, `A*B` expands to `A+B+A:B` and `A*x` expands to `A+x+A:x`. The $*$ operator can be applied to more than two arguments so, for example, `A*B*C` expands to `A+B+C+A:B+A:C+B:C+A:B:C`.

The ^ operator:
The exponentiation operator provides a convenient notation for iteration of the * operator so, for example, `(A+B+C)^2` is expanded to `(A+B+C)*(A+B+C)`. Recalling that, redundant terms are automatically eliminated, this expands to `A+B+C+A:B+A:C+B:C`. The primary purpose of this operator is therefore to specify models including all interactions up to a given order. As a further example, `(A+B+C+D)^2` expands to `A+B+C+D+A:B+A:C+A:D+B:C+B:D+C:D`.

The - operator:
The subtraction operator `-` is used to remove terms already included in the model formula. The most common applications are:

- To perform regression through the origin by removal of the intercept. In particular, the model $\eta_j = \beta x_j$ is specified as `~x-1`.
- To remove higher order interaction terms. For example, the model of no three factor interaction `A+B+C+A:B+A:C+B:C` can be specified as `A*B*C-A:B:C`.

It should be noted that the `-` operator is interpreted in context as the formula is parsed from left to right, so that `A+B-A+A` is equivalent to `B+A`.

If the term being subtracted is not present then the operation has no effect and does not generate an error.

The / operator:
The nesting operator `/` specifies models that are equivalent to the * operator but parameterised differently. Consider the separate regression model `A*x`.

```
model.matrix(~A*x)
```

```
##   (Intercept) A2 A3 x A2:x A3:x
## 1           1  0  0 1    0    0
## 2           1  0  0 2    0    0
## 3           1  0  0 3    0    0
## 4           1  1  0 4    4    0
## 5           1  1  0 5    5    0
## 6           1  1  0 6    6    0
## 7           1  0  1 7    0    7
## 8           1  0  1 8    0    8
## 9           1  0  1 9    0    9
```

As discussed previously, this model allows for a different slope within each group and the slopes are parameterised in terms of the slope for group 1 and then the differences between each other group and group 1.

The same model can be specified using the nesting operator. In this case the three slope parameters correspond directly to the three groups and are therefore said to be nested within `A`.

```
model.matrix(~A/x)
```

```
##   (Intercept) A2 A3 A1:x A2:x A3:x
## 1           1  0  0    1    0    0
## 2           1  0  0    2    0    0
## 3           1  0  0    3    0    0
## 4           1  1  0    0    4    0
## 5           1  1  0    0    5    0
## 6           1  1  0    0    6    0
## 7           1  0  1    0    0    7
## 8           1  0  1    0    0    8
## 9           1  0  1    0    0    9
```

Lecture 6, 2018

2.4.4 The marginality principle

Consider an experiment with a single factor A having r levels and the model ~A or, equivalently,

$$M : \eta_{ij} = \mu + \alpha_i.$$

As previously discussed, the parameters α_i are not identifiable unless a suitable constraint is imposed and that constraint can be chosen in many different ways. For example, $\alpha_1 = 0$ or $\sum_i \alpha_i = 0$.

Although the meanings of the parameter values depends critically upon the choice of constraint, the model is not affected. In particular, each such specification is equivalent to the statement that η_{ij} does not depend on j.

Suppose now that we remove the intercept term and consider the model,

$$M : \eta_{ij} = \alpha_i.$$

If, in addition, we impose a constraint such as $\alpha_1 = 0$ or $\sum_i \alpha_i = 0$ different models are obtained. For this reason, it is required that the intercept term μ be included in the model whenever the factor A is present.

The generalisation of this requirement is called the *marginality principle*. For factorial experiments, it states that:

> Whenever an interaction term is included in the model, all implied lower order interactions and main effects must also be included.

For example, suppose the interaction term A:B:C is to be included in a model.The marginality principle requires that the implied two-way interaction terms, A:B, A:C, B:C, the main effects A, B, C and the intercept 1 all be included.

The reason for this restriction is similar to the one-factor illustration.

- It is necessary to impose constraints on the parameters of the model to achieve identifiability;
- The constraints will not affect the model provided that the marginality principle is observed.

The marginality principle is enforced by the model formulæ in R as follows:

1. The standard interaction operator $*$ expands according to the marginality principle: A*B=1+A+B+A:B.
2. If terms are explicitly removed from a model formula in R, the corresponding constraints are also removed. For example:

```
model.matrix(~A)
```

```
##   (Intercept) A2 A3
## 1           1  0  0
## 2           1  0  0
## 3           1  0  0
## 4           1  1  0
## 5           1  1  0
## 6           1  1  0
## 7           1  0  1
## 8           1  0  1
## 9           1  0  1
```

```
model.matrix(~A-1)
```

```
##   A1 A2 A3
## 1  1  0  0
## 2  1  0  0
## 3  1  0  0
## 4  0  1  0
## 5  0  1  0
## 6  0  1  0
## 7  0  0  1
## 8  0  0  1
## 9  0  0  1
```

3. To illustrate further, consider the following example involving two factors, `A` and `B` and the two-way interaction model

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

The $*$ operator expands `A*B` to `A+B+A:B` as discussed previously and the constraints $\alpha_1 = \beta_1 = \gamma_{i1} = \gamma_{1j} = 0$ are imposed as is apparent in the model matrix.

```
model.matrix(~A*B)
```

```
##   (Intercept) A2 A3 B2 B3 A2:B2 A3:B2 A2:B3 A3:B3
## 1           1  0  0  0  0     0     0     0     0
## 2           1  0  0  1  0     0     0     0     0
## 3           1  0  0  0  1     0     0     0     0
## 4           1  1  0  0  0     0     0     0     0
## 5           1  1  0  1  0     1     0     0     0
## 6           1  1  0  0  1     0     0     1     0
## 7           1  0  1  0  0     0     0     0     0
## 8           1  0  1  1  0     0     1     0     0
## 9           1  0  1  0  1     0     0     0     1
```

If the `A` main effect is removed (in violation of the marginality principle) `R` compensates by removing the constraint $\gamma_{i1} = 0$.

```
model.matrix(~A*B-A)
```

```
##   (Intercept) B2 B3 A2:B1 A3:B1 A2:B2 A3:B2 A2:B3 A3:B3
## 1           1  0  0     0     0     0     0     0     0
## 2           1  1  0     0     0     0     0     0     0
## 3           1  0  1     0     0     0     0     0     0
## 4           1  0  0     1     0     0     0     0     0
## 5           1  1  0     0     0     1     0     0     0
## 6           1  0  1     0     0     0     0     1     0
## 7           1  0  0     0     1     0     0     0     0
## 8           1  1  0     0     0     0     1     0     0
## 9           1  0  1     0     0     0     0     0     1
```

If the interaction term A:B is specified in isolation, i.e. without the necessary main effects, R tries to compensate by removing all of the constraints on γ_{ij}. Note in this case that the columns of the model matrix are not linearly independent.

```
model.matrix(~A:B)
```

```
##   (Intercept) A1:B1 A2:B1 A3:B1 A1:B2 A2:B2 A3:B2 A1:B3 A2:B3 A3:B3
## 1           1     1     0     0     0     0     0     0     0     0
## 2           1     0     0     0     1     0     0     0     0     0
## 3           1     0     0     0     0     0     0     1     0     0
## 4           1     0     1     0     0     0     0     0     0     0
## 5           1     0     0     0     0     1     0     0     0     0
## 6           1     0     0     0     0     0     0     0     1     0
## 7           1     0     0     1     0     0     0     0     0     0
## 8           1     0     0     0     0     0     1     0     0     0
## 9           1     0     0     0     0     0     0     0     0     1
```

Polynomial regressions A weaker version of the marginality principle also applies to polynomial regression models. In particular, it is usually argued that we should not include the cubic term x^3 in a polynomial regression unless the lower order terms 1, x and x^2 are also included. This requirement is not driven by considerations of identifiability. Models such as $\eta = \beta x^3$ are unambiguously defined but tend not to be very useful in general.

Models with factors and covariates The marginality principle is applied to models such as the analysis of covariance model, A∗x=A+x+A:x. In particular the usual rules for factors are applied to any terms containing factors.

Lecture 7, 2018

3 Regression diagnostics

3.1 The regression assumptions

An important part of good statistical practice is assumption checking. When a model, such as the multiple regression model, is applied to data, it is important to understand that the use of the model involves making certain assumptions about the data. If the assumptions are not valid, then conclusions based on the model have the potential to be misleading. It is therefore important to check the assumptions as far as possible before using a statistical model to make conclusions.

The linear regression model can be written succinctly as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$$

where $e_1, e_2, \ldots, e_n$ are IID $N(0, \sigma^2)$ *realisations*.

More explicitly, it is assumed that:

1. The e_i all have mean zero;
2. The e_i all have variance σ^2;
3. The e_i are normally distributed;
4. The e_i are independent.

In addition, it is also assumed that

5. The x variables are known without error;
6. The errors e_i are not correlated with the x variables.

In practice, most attention is focused on using the data to investigate the plausibility of assumptions 1-3. Assumptions 4-6, are more fundamental but cannot be verified from the observed data. It is important to understand the context in which the data arose in order to determine whether these assumptions are reasonable.

3.1.1 Residuals

Model checking in linear models is usually based on analysis of the *residuals*. The *ordinary residuals* are defined by

$$\hat{e}_i = y_i - \hat{\eta}_i$$

or, in vector notation,

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{\eta}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}.$$

The intuitive justification for consideration of residuals is that if the model

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

where $e_1, e_2, \ldots, e_n$ are independent $N(0, \sigma^2)$ realisations holds, then we should have

$$\hat{e}_i \approx e_i.$$

In this case $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ should look roughly like a sample of independent $N(0, \sigma^2)$ observations.

However, a more careful analysis can be applied to the distribution of the residuals. In particular, suppose the model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \text{ with } \boldsymbol{\mathcal{E}} \sim N_n(\boldsymbol{0}, \sigma^2 I)$$

holds and consider

$$\hat{\boldsymbol{\mathcal{E}}} = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}} = (I - H)\boldsymbol{Y}$$

where

$$H = X(X^T X)^{-1} X^T.$$

Note that H is the orthogonal projection matrix also denoted P. It follows that $E(\hat{\boldsymbol{\mathcal{E}}}) = \boldsymbol{0}$ and

$$\operatorname{Var}(\hat{\boldsymbol{\mathcal{E}}}) = \operatorname{Var}\{(I - H)\boldsymbol{Y}\} = (I - H)\operatorname{Var}(\boldsymbol{Y})(I - H)^T = \sigma^2(I - H).$$

This shows that unlike an IID $N(0, \sigma^2)$ sample, the residuals are typically not independent and do not have the same variances. In particular, it follows that

$$\operatorname{var}(\hat{\mathcal{E}}_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} is the ith diagonal element of H.

To correct for the unequal variances, the *standardized residuals* are defined by

$$\hat{e}'_i = \frac{\hat{e}_i}{s_e\sqrt{1 - h_{ii}}}.$$

The standardized residuals thus have the same variance but are still not independent.

A second difficulty with the raw and standardized residuals is that an aberrant y_i-value will influence the corresponding fitted value, $\hat{\eta}_i$, and this may lead to an unremarkable residual, thus masking the fact that the data point was in fact aberrant. To avoid this difficulty, the *studentized residuals* are defined by

$$\hat{e}^*_i = \frac{y_i - \hat{\eta}^{(i)}}{\sqrt{\widehat{\operatorname{var}}(Y_i - \hat{\eta}^{(i)})}}$$

where $\hat{\eta}^{(i)}$ and $\widehat{\operatorname{var}}(Y_i - \hat{\eta}^{(i)})$ are calculated from the data with the ith observation omitted. It can be shown that

$$\hat{e}^*_i = \hat{e}'_i \left(\frac{n - p - \hat{e}'^2_i}{n - p - 1} \right)^{-1/2}.$$

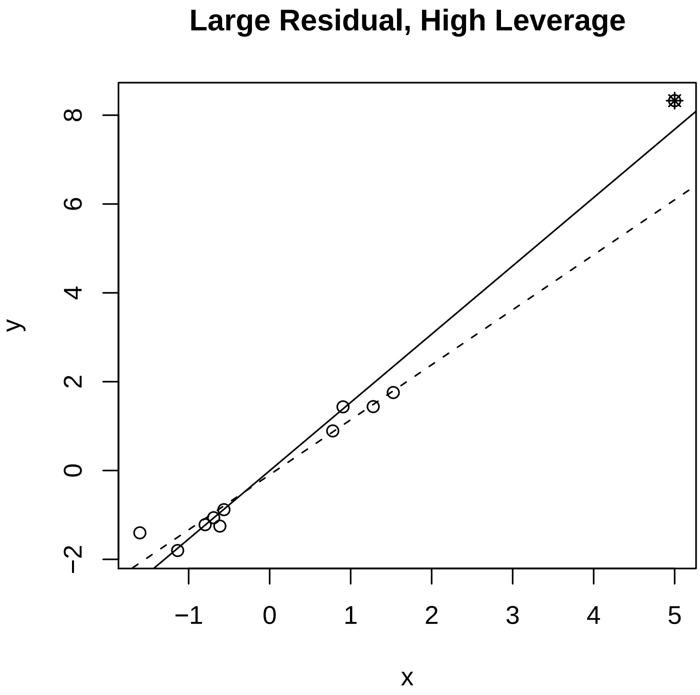
In practice, it is recommended that the studentized residuals be used for basic model checking as introduced in previous courses. In particular:

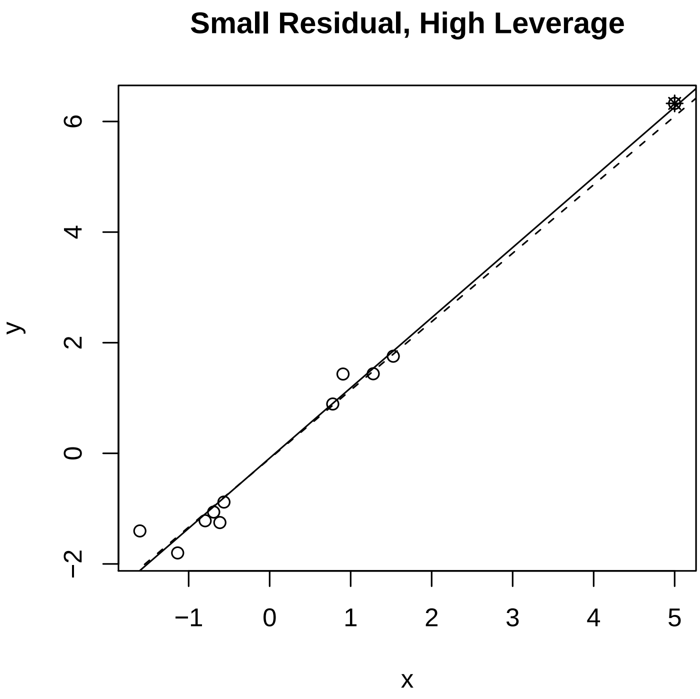
1. The $\hat{e}_i^*$ should be plotted against the fitted values to check for lack of fit and heteroscedasticity;
2. The $\hat{e}_i^*$ should be plotted against each of the predictor variables separately to check for lack of fit and heteroscedasticity;
3. If an additional variable such as "time" is recorded but not considered for inclusion in the model, it is good practice to examine a plot of the residuals versus time to check for any trends etc.
4. If it is established that there are no problems with lack of fit, heteroscedasticity and trends over time (when applicable), a normal quantile plot of the residuals can be used to assess whether the assumption of normality is reasonable.

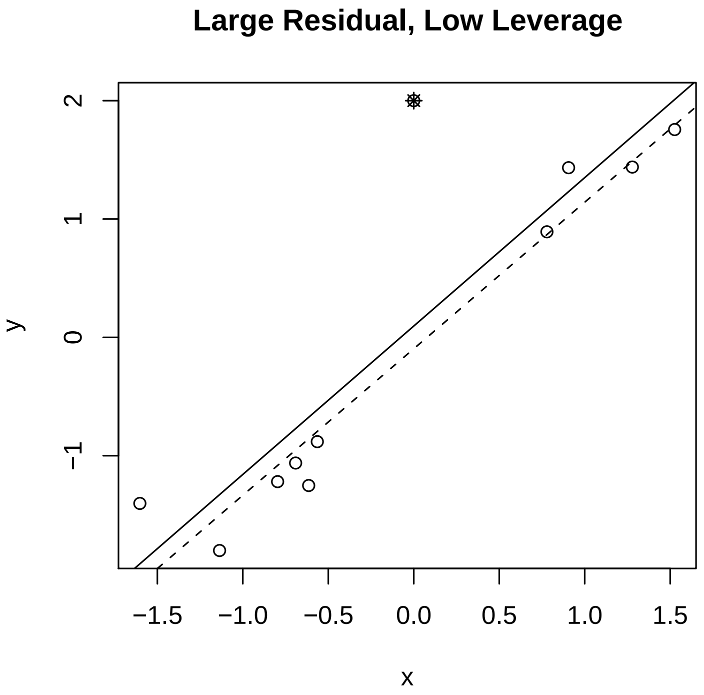
3.2 Influence diagnostics

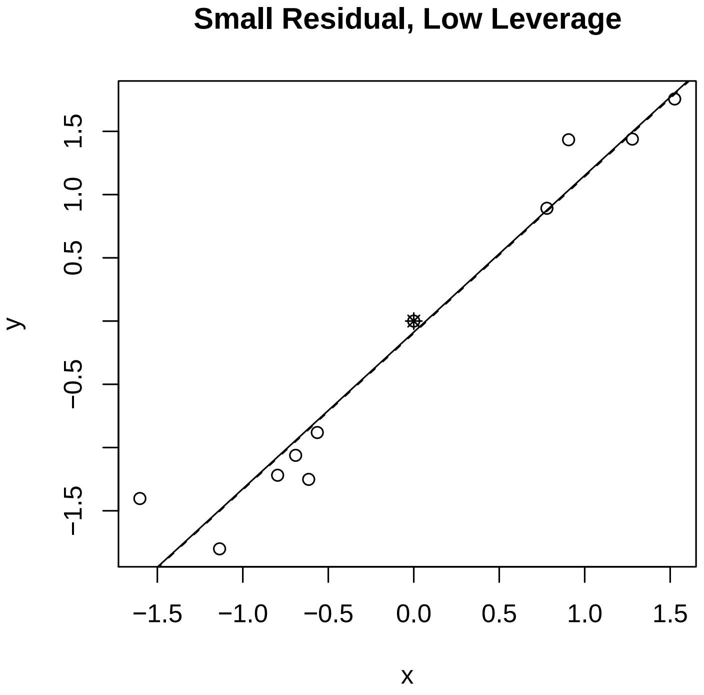
It sometimes happens that a data set contains a small number of outliers or points with large residuals. The occurrence of such points can be problematic in the sense that a single outlier may have an appreciable impact on the parameter estimates. If it cannot be verified that the data point is erroneous then there is no good basis for its omission although one may nevertheless be suspicious. Influence diagnostics are used to help identify points that have a disproportionately large impact on the estimates.

In general, a point with a small studentized residual is unlikely to have a large impact on the parameter estimates. A point with a large studentized residual, may or may not depending on its *leverage*. To illustrate, consider the following examples.


Large Residual, High Leverage
y
−2
0
2
4
6
8
−1
0
1
2
3
4
5
x


Small Residual, High Leverage
y
−2
0
2
4
6
−1
0
1
2
3
4
5
x





In each case, the solid line represents the least squares regression line when all points are included and the dashed line is the least squares regression line with the marked point excluded. In simple linear regression, the notion of *leverage* can be thought of as “the distance from x_i

to $\bar{x}$". In the first plot, where the marked point has a large residual and of high leverage, the slope is changed appreciably. In the remaining plots, the omission of the marked point has only minimal impact on the least squares regression line.

Definition 3.1 *The* leverage *of a data point is defined by*

$$h_{ii} = [X(X^TX)^{-1}X^T]_{ii} = \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i$$

where $\boldsymbol{x}_i^T$ is the ith row of X. □

The motivation for this definition will be discussed later. It can also be checked in the case of simple linear regression that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

where

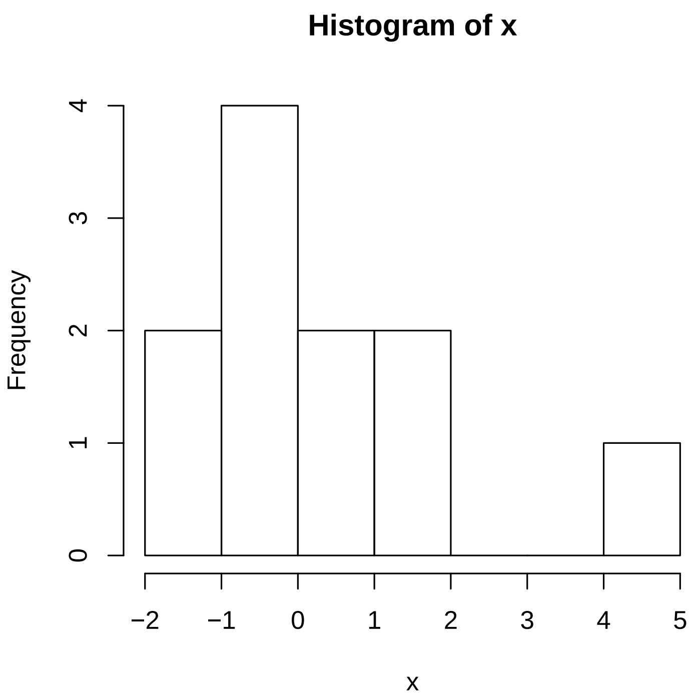
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

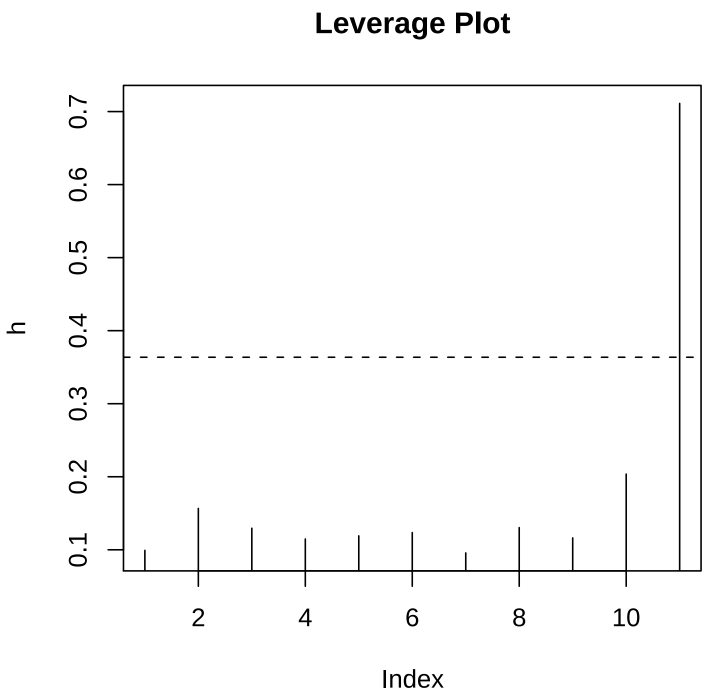
Having thus introduced the leverage, we need to define what is meant by "large". Consider

$$\begin{aligned}
\sum_{i=1}^{n} h_{ii} &= \mathrm{tr}(H) \\
&= \mathrm{tr}(X(X^TX)^{-1}X^T) \\
&= \mathrm{tr}((X^TX)^{-1}X^TX) \\
&= \mathrm{tr}(I_{p\times p}) = p.
\end{aligned}$$

Therefore, the average value of h_{ii} is p/n.

Leverages are sometimes displayed using an *index plot*. That is, a plot of h_{ii} vs i, to identify points of high leverage. Some authors suggest that points with $h_{ii} \geq 2p/n$ be highlighted. Shown below is the histogram of x-values and also a leverage plot of the data from the first scatter plot in the preceding example.


Histogram of x
Frequency
0
1
2
3
4
−2
−1
0
1
2
3
4
5
x


Leverage Plot
h
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2
4
6
8
10
Index

Points of high leverage have potentially a large impact on the estimates $\hat{\boldsymbol{\beta}}$. Cook's distance measures how much influence each data point actually has by calculating the effect of its removal upon $\hat{\boldsymbol{\beta}}$.

Let $\hat{\boldsymbol{\beta}}$ denote the least squares estimate of $\boldsymbol{\beta}$ based on the full data set and let $\hat{\boldsymbol{\beta}}^{(i)}$ be the least squares estimate when the ith data point is omitted. An obvious approach would be to use $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\|^2$ as a measure of the influence of the ith data point. However, such a measure is not satisfactory because it gives equal weight to all components of $\boldsymbol{\beta}$ even though their variances are generally different. Cook's distance uses a more appropriate distance measure.

Definition 3.2 *The* Cook's distance statistic *for the ith data point is defined by*

$$\begin{aligned} D_i^2 &= \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)^T \left[\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)}{p} \\ &= \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)^T (X^T X) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)}\right)}{p s_e^2}. \end{aligned}$$

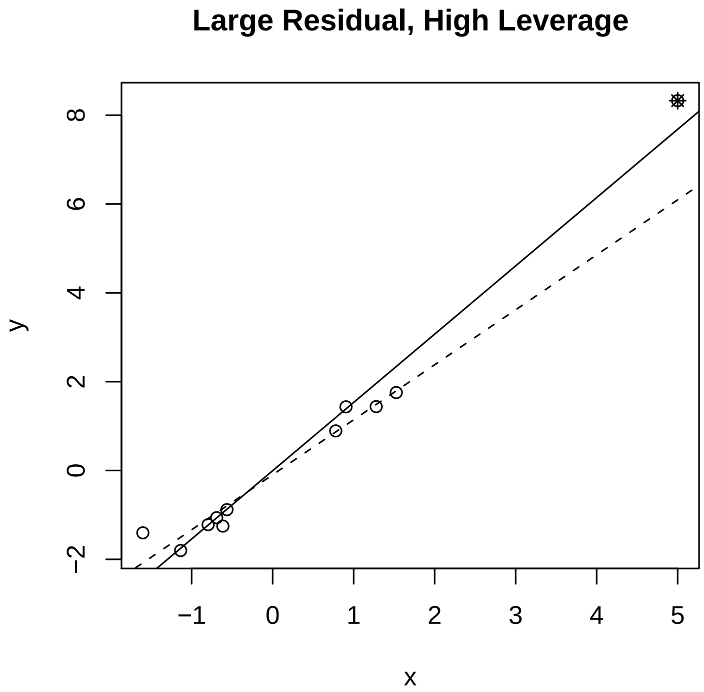
□

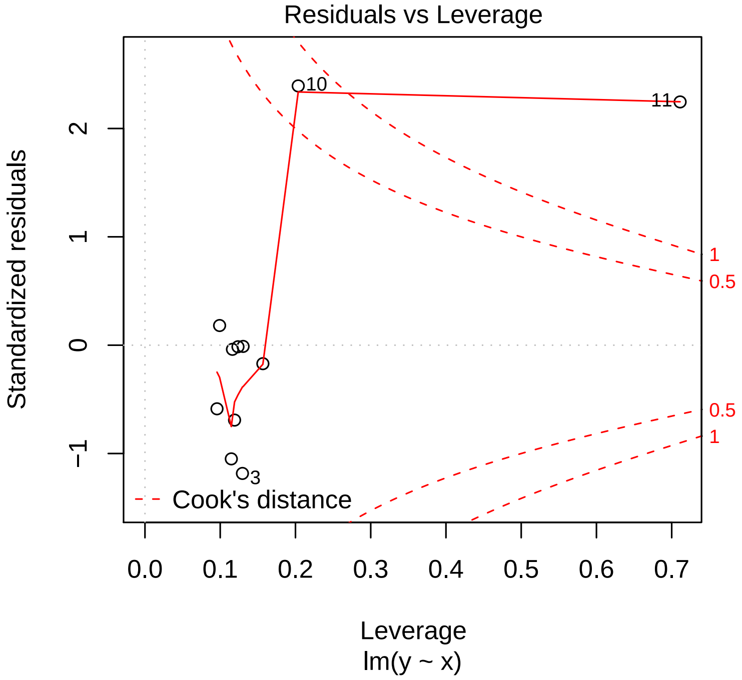
It can be shown that

$$D_i^2 = \frac{(\hat{e}_i')^2 h_{ii}}{p(1 - h_{ii})}.$$

Remark This formula for D_i^2 gives rise to the interpretation of the leverage h_{ii}. In particular, it follows that the total impact of the ith data point on the least squares estimate, $\hat{\boldsymbol{\beta}}$ will be large only if $\hat{e}_i'$ and $h_{ii}/(1 - h_{ii})$ are large. Since $h/(1-h)$ is increasing for $0 < h < 1$, a point can be influential only when it has a high leverage h_{ii}.

Some authors suggest that data points with $D_i^2 \geq 1$ are a cause for concern as this represents an average change to $\hat{\boldsymbol{\beta}}$ of 1 standard error. Cook's distance can be displayed using an index plot in order to identify influential points. An alternative, produced in R is to plot the standardized residual $\hat{e}_i'$ vs the leverage h_{ii} and indicate the Cook's distances by contour lines. To illustrate, a scatter plot and the corresponding residual vs leverage plot are shown below.


Large Residual, High Leverage
y
8
6
4
2
0
−2
−1
0
1
2
3
4
5
x

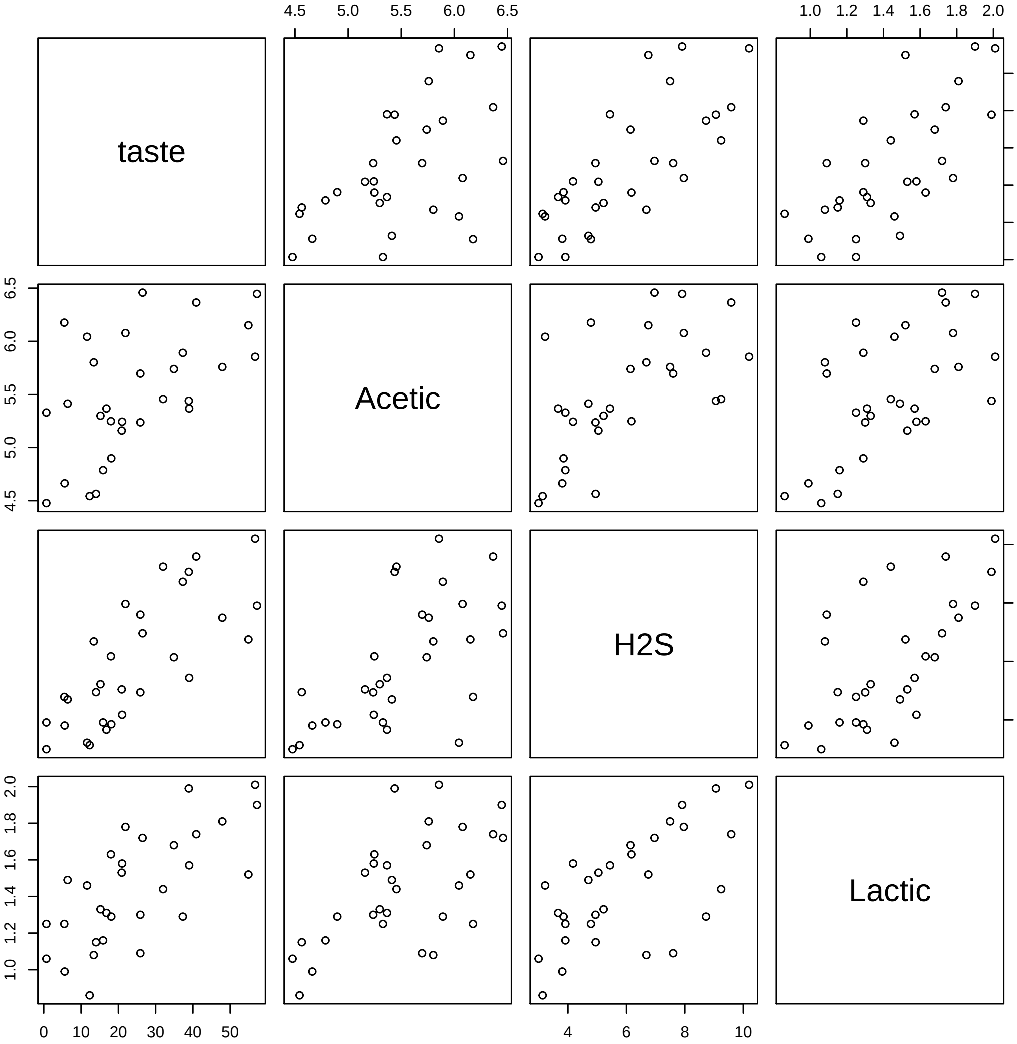

Residuals vs Leverage
Standardized residuals
2
1
0
−1
10
11
3
1
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Leverage
lm(y ~ x)

3.2.1 Example: Residuals in R

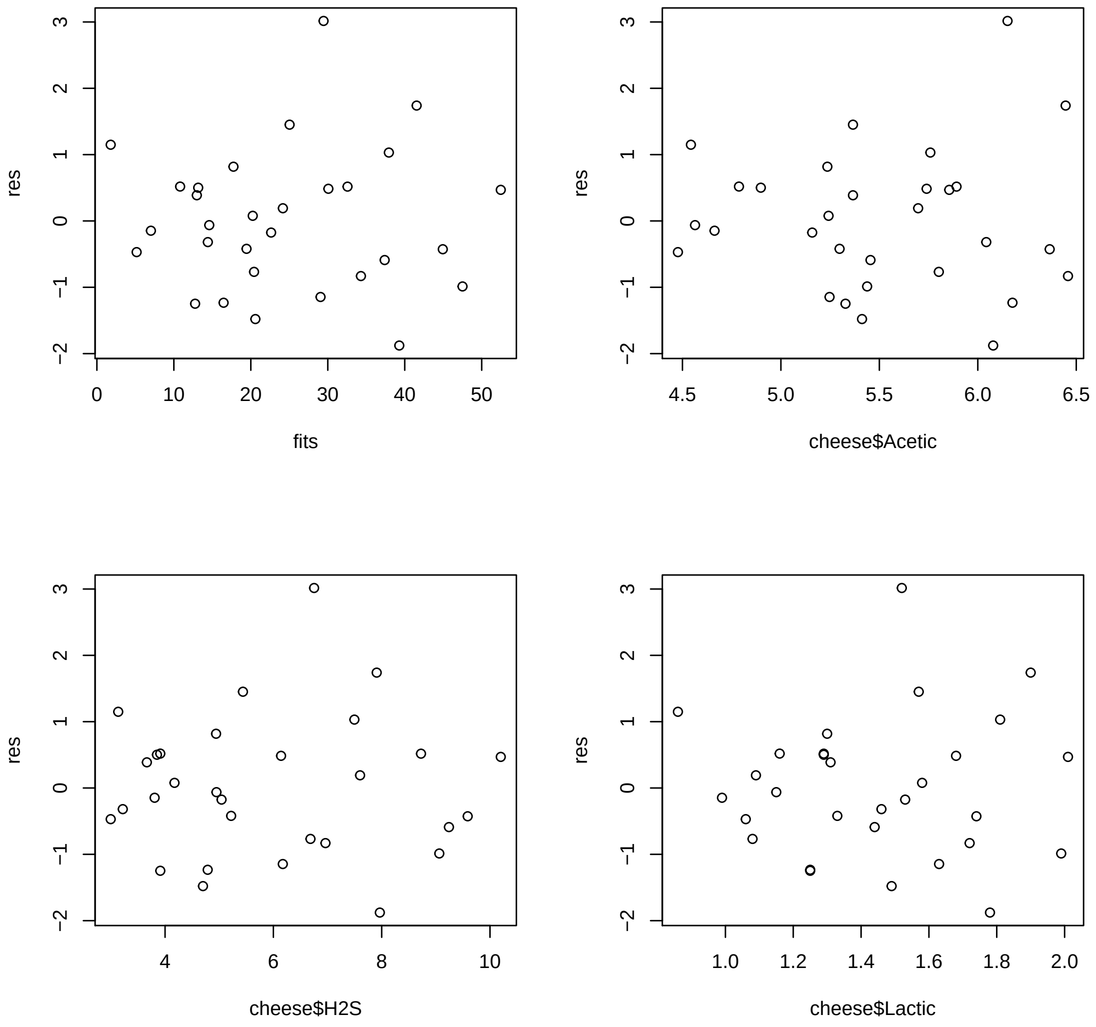
As cheese ages, various chemical processes take place that determine the taste of the final product. This dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample. The variables Acetic and H2S are the natural logarithm of the concentration of acetic acid and hydrogen sulphide respectively. The variable Lactic has not been transformed. The data are from Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics.

The purpose of the analysis is to understand the influence of the different variables on the overall taste score. What follows is the generation of the standard diagnostic plots for the multiple regression of taste on Acetic, Lactic and H2S. In this case, the conclusion is that the regression assumptions appear to be reasonable and there are no influential points of concern.

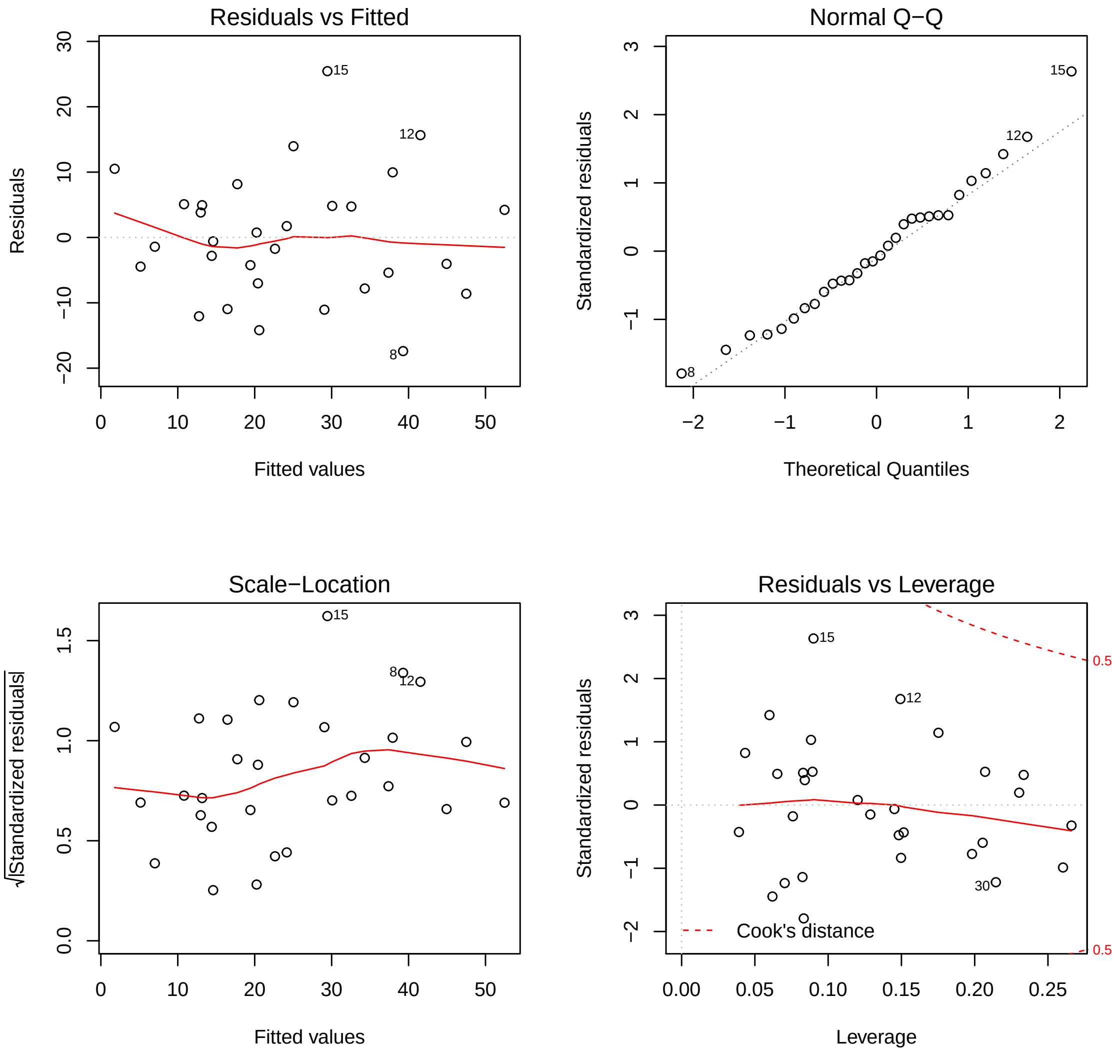
```
cheese<-read.table("cheese.txt",header=T)
cheese

##    taste Acetic    H2S Lactic
## 1   12.3  4.543  3.135   0.86
## 2   20.9  5.159  5.043   1.53
## 3   39.0  5.366  5.438   1.57
## 4   47.9  5.759  7.496   1.81
## 5    5.6  4.663  3.807   0.99
## 6   25.9  5.697  7.601   1.09
## 7   37.3  5.892  8.726   1.29
## 8   21.9  6.078  7.966   1.78
## 9   18.1  4.898  3.850   1.29
## 10  21.0  5.242  4.174   1.58
## 11  34.9  5.740  6.142   1.68
## 12  57.2  6.446  7.908   1.90
## 13   0.7  4.477  2.996   1.06
## 14  25.9  5.236  4.942   1.30
## 15  54.9  6.151  6.752   1.52
## 16  40.9  6.365  9.588   1.74
## 17  15.9  4.787  3.912   1.16
## 18   6.4  5.412  4.700   1.49
## 19  18.0  5.247  6.174   1.63
## 20  38.9  5.438  9.064   1.99
## 21  14.0  4.564  4.949   1.15
## 22  15.2  5.298  5.220   1.33
## 23  32.0  5.455  9.242   1.44
## 24  56.7  5.855 10.199   2.01
## 25  16.8  5.366  3.664   1.31
## 26  11.6  6.043  3.219   1.46
## 27  26.5  6.458  6.962   1.72
## 28   0.7  5.328  3.912   1.25
## 29  13.4  5.802  6.685   1.08
## 30   5.5  6.176  4.787   1.25

#Plot the data for  initial examination
pairs(cheese)
```


taste
Acetic
H2S
Lactic
4.5
5.0
5.5
6.0
6.5
1.0
1.2
1.4
1.6
1.8
2.0
0
10
20
30
40
50
4
6
8
10

```r
# Next, fit the full model and examine the residual plots
lm.full<-lm(taste~Acetic+H2S+Lactic,data=cheese)
# Extract the studentized residuals (Needs the MASS package loaded)
res<-studres(lm.full)
fits<-fitted(lm.full)
par(mfrow=c(2,2))
plot(fits,res)
plot(cheese$Acetic,res)
plot(cheese$H2S,res)
plot(cheese$Lactic,res)
```



```
# Basic residual plots look fine.
# examine the normal quantile, scale-location and leverage plots.
par(mfrow=c(2,2))
plot(lm.full)
# conclude that regression assumptions are suitable
# for the model taste~Acetic+H2S+Lactic.
```



4 Model Building

The mathematical treatment of the linear model begins with the assumption that the model

$$M : \boldsymbol{\eta} = X\boldsymbol{\beta}$$

is known to be true.

In practice, such information is not given in advance and an appropriate model must be deduced from the available data. In broad terms, the general principles that guide the selection of a suitable model are:

1. The linear modelling assumptions should be plausible;
2. The model should contain all important variables;
3. The model should not include unnecessary variables.

The implementation of these principles depends on the scientific context and the purpose of the analysis. To illustrate, we will consider various commonly occurring situations in which linear modelling techniques may be applied.

4.1 Common applications of linear models

4.1.1 Covariate adjustment

Consider data comprising a response Y and predictor x recorded over two groups. Suppose the data are of the form (y_{ij}, x_{ij}) for $j = 1, 2, \ldots, n_i$, $i = 1, 2$.

For example, in a certain experiment, the weight gain over 6 weeks, Y, was recorded for each of 10 pigs on two different diets. In addition, the initial weight, x at the start of the study was recorded for each pig.

In such examples, the analysis of covariance model

$$\eta_{ij} = \mu + \alpha_i + \beta x_{ij}$$

is often considered. This model can be applied in several different contexts with very different interpretations. In particular, we will consider the role of the analysis of covariance model:

- as a means of variance reduction in a randomized experiment;
- as a means of bias correction in an observational study or quasi-experiment;
- as a means of elucidation on a relationship.

Lecture 9, 2018

4.1.2 Variance reduction in randomised experiments

Consider a randomised experiment designed to compare a number of different treatments. For example, in a certain study on the effects of two drugs for the treatment of leprosy, 30 patients were randomly allocated to receive one drug A, drug D or the control (labelled F). The response variable Y is a score representing the abundance of the leprosy bacilli over six sites on the body. Since this was a randomized trial, unbiased estimates of treatment effects can be obtained from a one-way analysis of variance.

```
summary(lm(y~Treatment,data=leprosy))
```

```
##
## Call:
## lm(formula = y ~ Treatment, data = leprosy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -11.3   -4.9   -0.8    3.5   11.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.300      1.920   6.407 7.29e-07 ***
## Treatmentdrug A   -7.000      2.715  -2.578   0.0157 *
## Treatmentdrug D   -6.200      2.715  -2.284   0.0305 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.071 on 27 degrees of freedom
## Multiple R-squared:  0.2278,Adjusted R-squared:  0.1706
## F-statistic: 3.983 on 2 and 27 DF,  p-value: 0.03049
```

```
anova(lm(y~Treatment,data=leprosy))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Treatment  2  293.6 146.800  3.9831 0.03049 *
## Residuals 27  995.1  36.856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis shows that the two active treatments produce significantly lower levels of score than the control.

The abundance score for the leprosy bacilli before the treatments were administered, X, was also recorded. As a preliminary analysis, we consider X as the response variable.

```
summary(lm(x~Treatment,data=leprosy))

##
## Call:
## lm(formula = x ~ Treatment, data = leprosy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.30  -3.30  -1.10   2.75   9.70
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.900      1.482   8.705 2.55e-09 ***
## Treatmentdrug A   -3.600      2.096  -1.718   0.0973 .
## Treatmentdrug D   -2.900      2.096  -1.384   0.1778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.686 on 27 degrees of freedom
## Multiple R-squared:  0.1094,Adjusted R-squared:  0.04346
## F-statistic: 1.659 on 2 and 27 DF,  p-value: 0.2092

anova(lm(x~Treatment,data=leprosy))

## Analysis of Variance Table
##
## Response: x
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  2  72.87  36.433  1.6589 0.2092
## Residuals 27 593.00  21.963
```

This analysis shows that initial abundance scores were lower for the two active treatment groups than for the control, but the differences are not statistically significant.

To incorporate the pre-treatment information into the analysis we apply the analysis of covariance model.

```
summary(lm(y~x+Treatment,data=leprosy))

##
## Call:
## lm(formula = y ~ x + Treatment, data = leprosy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4115 -2.3891 -0.5711  1.7237  8.5885
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.4347     2.4714  -0.176   0.8617
## x                 0.9872     0.1645   6.001 2.45e-06 ***
## Treatmentdrug A  -3.4461     1.8868  -1.826   0.0793 .
## Treatmentdrug D  -3.3372     1.8539  -1.800   0.0835 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.006 on 26 degrees of freedom
## Multiple R-squared:  0.6763,Adjusted R-squared:  0.6389
## F-statistic:  18.1 on 3 and 26 DF,  p-value: 1.501e-06
```

```
anova(lm(y~x+Treatment,data=leprosy))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 802.94  802.94 50.0393 1.639e-07 ***
## Treatment  2  68.55   34.28  2.1361    0.1384
## Residuals 26 417.20   16.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(y~Treatment+x,data=leprosy))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Treatment  2  293.6  146.80  9.1486 0.0009812 ***
## x          1  577.9  577.90 36.0145 2.454e-06 ***
## Residuals 26  417.2   16.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remarks

- Perhaps the most obvious consequence of including the initial abundance scores in the model is that the treatment effects are no longer statistically significant. As shown in the previous analysis, this is because there were some pre-treatment differences between the groups and these have now been accounted for.
- The major point of this example, however, is the effect on the residual standard deviation.
  - In the initial (unadjusted) analysis, the residual standard deviation (called residual standard error in R) was 6.071.
  - In the adjusted analysis, the residual standard deviation is 4.006.
  - The standard errors of the treatment effects are similarly reduced.
  - This is because the covariate X explains a large part of the residual variance in the unadjusted analysis.
- As a technical remark, it should be noted that the order in which terms are included in the anova command is important.
- In this case it would also have been valid to account for the initial abundance X by taking the difference $Y - X$ as the response variable.

Conclusion

- In the context of a randomised experiment, the inclusion of a *covariate* in the model can lead to a substantial reduction in residual variance and a corresponding increase in precision.
- Because of the randomisation, both analyses should produce unbiased estimates of the treatment effects so large changes in the estimated effects are not expected although as the preceding example illustrates, they may occur.
- When the response and covariate are before and after measurements respectively, an alternative is to use the difference for the response variable.
- The analysis of covariance approach can be applied when the covariate and response are not of the same type or when there is more than one covariate.

Lecture 10, 2018

4.1.3 Observational studies and quasi-experiments

Multiple regression can also be used to try to adjust for bias that arises from *confounding* in observational studies.

Example FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. The data contains determinations of FEV on 654 children ages 6-22 who were seen in the Childhood Respiratory Disease Study in 1980 in East Boston, Massachusetts. The data are part of a larger study to follow the change in pulmonary function over time in children. Two of the *predictor* variables considered in the study were `Smoking Status` and `Age`.

Suppose now that we wish to investigate the effect of smoking on pulmonary function as measured by `FEV`. If we perform a simple regression of `log(FEV)` on smoking status, the following output is produced.

```
fev<-read.table("fev.txt",header=TRUE)
# head(fev)
summary(lm(log(FEV)~Smoker,data=fev))

##
## Call:
## lm(formula = log(FEV) ~ Smoker, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12285 -0.22803  0.01238  0.21777  0.86825
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.16048    0.04011  28.930  < 2e-16 ***
## SmokerNon   -0.27208    0.04227  -6.437 2.36e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3234 on 652 degrees of freedom
## Multiple R-squared:  0.05975,Adjusted R-squared:  0.05831
## F-statistic: 41.43 on 1 and 652 DF,  p-value: 2.364e-10
```

The output shows that smoking has a highly significant effect but paradoxically, the coefficient associated with non-smokers is negative suggesting that non-smokers tend to have lower FEV. In particular, $\exp(-0.27208) = 0.762$ so non-smokers' FEVs were roughly 23.8% lower than those of smokers on average.

However, this was an observational study so it is possible that the apparent beneficial effect of smoking can be explained by other differences. For example, in this study, we would expect smokers to be older than non-smokers. An obvious mechanism is that the apparent benefit of smoking is explained by the fact that the smokers tend to be older as a group and therefore have a higher average FEV than the non-smokers. To allow for this difference we can also include `Age` in the regression.

```
summary(lm(log(FEV)~Smoker+Age,data=fev))

## 
## Call:
## lm(formula = log(FEV) ~ Smoker + Age, data = fev)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.71124 -0.13458  0.00104  0.14909  0.60261 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.066988   0.048864  -1.371  0.17088    
## SmokerNon    0.089927   0.030118   2.986  0.00293 ** 
## Age          0.090768   0.003053  29.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2108 on 651 degrees of freedom
## Multiple R-squared:  0.6012,Adjusted R-squared:    0.6 
## F-statistic: 490.8 on 2 and 651 DF,  p-value: < 2.2e-16
```

In this analysis, there is still a significant effect due to smoking but now, the apparent effect is negative. In particular, $\exp(0.089927) = 1.094$ so, *for subjects of the same age*, the FEV for non-smokers is on average 9.4% higher than for smokers. On face value this appears to be a more credible analysis. However, because it is an observational study, we cannot be sure that there are still other important confounding factors that need to be considered. For example, `Sex`.

□

In this situation, the purpose of adjustment is to correct for bias rather than to improve precision. This type of analysis is often applied to non-randomized experiments as well as observational studies. Although such analyses can produce plausible results, any conclusions drawn from such studies are alway subject to question. In particular, we cannot be sure that the available variables allow us to adjust for all of the important confounding factors. Moreover, it is not obvious that the analysis of covariance model will always result in an appropriate adjustment. For example, in the preceding analysis we could have adjusted for `log(Age)` instead of `Age`. This would have produced a slightly different estimate for the effect of smoking.

4.1.4 Elucidation

The final common application of covariate adjustment is to elucidate upon a relationship between two variables. That is, to gain further insight into the nature of the relationship.

Example Continuing with the pulmonary function data, we consider the variables `Sex`, `Age` and `Height` as predictors for FEV. To investigate the effect of `Sex` on `FEV` allowing for age, consider the multiple regression model.

```
summary(lm(log(FEV)~Sex+Age,data=fev))
```

```
##
## Call:
## lm(formula = log(FEV) ~ Sex + Age, data = fev)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6777 -0.1281  0.0044  0.1492  0.5550
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.004991   0.029313   0.170    0.865
## SexMale     0.098128   0.016158   6.073 2.13e-09 ***
## Age         0.086599   0.002736  31.651  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2064 on 651 degrees of freedom
## Multiple R-squared:  0.6175,Adjusted R-squared:  0.6163
## F-statistic: 525.4 on 2 and 651 DF,  p-value: < 2.2e-16
```

The coefficient for Male is significantly positive and since $\exp(0.098128) = 1.1031$, it appears that, on average, FEV for males is approximately 10.3% higher than for females.

A possible explanation is that for a given Age, males tend to be taller than females and could therefore be expected to have higher FEV. To investigate, we consider the multiple regression model.

```
summary(lm(log(FEV)~Sex+Height+Age,data=fev))
```

```
##
## Call:
## lm(formula = log(FEV) ~ Sex + Height + Age, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63616 -0.08685  0.01134  0.09035  0.40188
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.939555   0.078845 -24.599  < 2e-16 ***
## SexMale      0.031436   0.011714   2.684  0.00747 **
## Height       0.042986   0.001682  25.561  < 2e-16 ***
## Age          0.021198   0.003207   6.610 8.03e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1459 on 650 degrees of freedom
## Multiple R-squared:  0.8092,Adjusted R-squared:  0.8083
## F-statistic:   919 on 3 and 650 DF,  p-value: < 2.2e-16
```

Examination of the regression output shows that the coefficient of Sex is still significantly positive but reduced in magnitude compared to the original analysis. In particular, since $\exp(0.03146) = 1.0319$ we conclude that for a given Age and Height, FEV for males is on average 3.19% higher than for females.

Finally, we can also use regression to investigate the effect of `Sex` on `Height`.

```
summary(lm(Height~Age+Sex,data=fev))

##
## Call:
## lm(formula = Height ~ Age + Sex, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6227  -2.4362   0.0274   2.4652  10.0917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.23674    0.48281   93.69  < 2e-16 ***
## Age          1.52144    0.04506   33.76  < 2e-16 ***
## SexMale      1.55149    0.26613    5.83 8.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.4 on 651 degrees of freedom
## Multiple R-squared:  0.6457,Adjusted R-squared:  0.6446
## F-statistic: 593.1 on 2 and 651 DF,  p-value: < 2.2e-16
```

Inspection of the regression output shows a significant effect due to `Sex`. In particular, it appears that, for a given age, males are on average 1.55 inches taller than females.

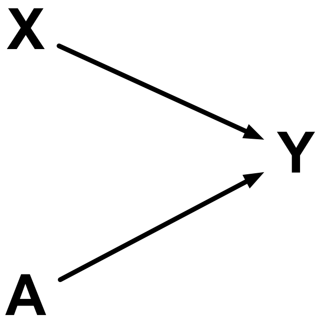
Combining the three analysis we can now make the following conclusions.

- Overall there is a strong relationship between `Sex` and `FEV`. In particular, `FEV` is substantially higher for males.
- This relationship can be explained in part by the variable `Height`. In particular,
  - Males tend to be taller than females;
  - Taller people tend to have higher `FEV`.
- However, this is not a complete explanation of the mechanism because after allowing for height, there is still a significant but reduced effect due to sex. We could conjecture that this is due to the fact that males also tend to have larger chests than females.

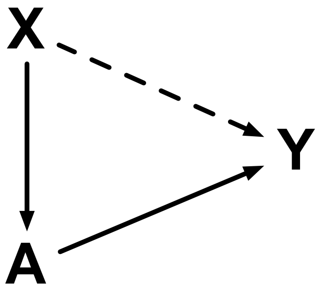
4.2 The role of predictor variables

In the preceding illustrations, we have seen predictor variables given different interpretation depending on the context. Consider a predictor of primary interest A, a covariate X and response Y. The following diagrams are intended to illustrate the logical relation of the variables in the different situations. In each case, the arrows can be interpreted very roughly as indicating the "direction of causation". The absence of an arrow indicates no association.

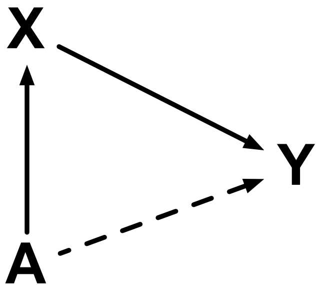
In the context of a randomized experiment, where the role of the covariate is variance reduction the situation is as follows. Note that there can be no association between X and A because of the randomization.



For an observational study or quasi-experiment, where the goal is to adjust for the confounder X, the arrows from X to A and from X to Y represent a pathway by which confounding may occur (Common Response). The arrow from A to Y represents the direct effect that cannot be explained in this way.



When the goal is elucidation, we see that A can influence Y through two pathways. Namely a direct effect and an indirect effect moderated through X.



In practice it is useful to understand the role of the variables being considered for inclusion in a regression. However some caution is needed in the interpretation of diagrams such as those shown above. First, the definition and establishment of causation is not at all straightforward and simply drawing an arrow does not imply causation. Second, there may also be situations where the two variables are associated but neither is the cause of the other. A more detailed discussion of these issues is beyond the scope of this course.

Lecture 11, 2018

4.3 Model selection algorithms

Consider the problem of choosing a suitable model given data

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$$

where $\boldsymbol{x}$ represents the vector of candidate predictor variables. Ideally, we would like to choose the smallest well-fitting model:

- Exclusion of important terms clearly leads to an incorrect model, which can lead to misleading conclusions;
- Including unnecessary terms diminishes the value of the model as a simplification of the data and also reduces the statistical accuracy of parameter estimates and predictions.

We consider three commonly used algorithms for model selection. It should be noted that the considerations discussed under covariate adjustment still apply in determining whether a particular term should be considered. In what follows, a *term* in a model may correspond either to a single numerical predictor or to a group of columns coding for a factor or interaction.

4.3.1 The forward selection algorithm

1. Begin with the null model.
2. For every term not currently included in the model, calculate a P-value for the inclusion of that term.
3. If the smallest P-value is less than the threshold p-in (usually chosen to be 0.05), add that term to the model.
4. Iterate (2), (3) until no further terms are significant.

4.3.2 The backwards elimination algorithm

1. Begin with the most complicated model to be considered.
2. For each term included in the model, calculate the P-value for its removal.
3. If the largest P-value is greater than the threshold p-out (usually chosen to be 0.05), remove that term from the model.
4. Iterate (2), (3) until the model contains only significant terms.

4.3.3 The stepwise selection procedure

1. Begin with the null model.
2. Perform one step of forward selection using a liberal value of p-in such as 0.2 or 0.15.
3. Perform one step of backward elimination with a value of p-out such as 0.05.
4. Iterate (2), (3) until no further changes occur or the algorithm cycles.

Remarks

- In all of these algorithms we can enforce the inclusion of certain variables.
- The marginality principle should also be observed at each step of the procedure.
- None of these methods is guaranteed to produce the "correct" model.
- The main weaknesses of forward selection are:
  - It typically begins with an incorrect model. Therefore the significance calculations will be wrong.
  - It can fail to detect situations where x_1 and x_2 are strong predictors for Y if considered jointly but not separately.
  - Terms included early in the process may subsequently become non-significant but will still be retained in the model.
- The main weaknesses of backward elimination are:
  - It assumes the initial model to be correct. If this is not the case, there is no scope to include additional terms.
  - It can be impractical if the number of candidate predictors is large.
- The stepwise procedure attempts to combine the good features of both methods but is by no means fool-proof.

4.4 Prediction

Suppose now that the purpose of the analysis is to develop a linear predictor $\hat{\eta}$ for Y. In this case, it is sensible to choose the model that minimises

$$E\left((Y - \hat{\eta})^2\right).$$

If we know the model

$$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

to be correct, then

$$\hat{\eta} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$$

is the best linear unbiased estimate for η.

At first sight it would appear that a reasonable strategy would be to seek the "correct" model using one of the previously discussed methods. However, the situation is slightly more complicated because we are using the data both to estimate the coefficients and also to estimate the model itself. (Note that the Gauss-Markov theorem assumes the correct model to be known *a priori*.) In this context it is important to note the following:

- Including unnecessary terms generally increases the estimation variance;
- Omitting necessary terms introduces bias into the estimation.

The choice of the optimal model for prediction is a trade-off between bias and variance.

Mallows(1973) suggests using the quantity

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p)$$

where p is the number of parameters,

$$RSS_p = \sum_{i=1}^{n} (y_i - \hat{\eta}_i)^2$$

and $\hat{\sigma}^2$ is an estimate of σ^2 based on the most complex model under consideration (assumed to be correct).

Now recall that for any correct model we will have

$$E(RSS_p) = (n - p)\sigma^2$$

and for an incorrect model

$$E(RSS_p) > (n - p)\sigma^2.$$

Therefore we would expect $C_p \approx p$ for any correct model and $C_p > p$ for an incorrect model. To find a suitable model, we choose the model with the smallest p for which $C_p \approx p$.

Lecture 12, 2018

4.5 Box-Cox transformations

When the assumptions of linearity, homoscedasticity and normality are found to be violated, it is sometimes possible to find a simple transformation of the data for which the regression assumptions are more reasonable.

For example, in previous courses, you may have tried transformations such as $\log y$, $\sqrt{y}$, $1/y$ when dealing with a positive variable Y.

A more systematic approach for positive Y, is to consider the family of power transformations, y^λ. A convenient formulation is to consider the family of Box-Cox transformations defined by

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0. \end{cases}$$

It can be shown that

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log y$$

so that $y^{(\lambda)}$ is a continuous function of λ.

The Box-Cox approach to transformation consists of the following steps.

- In the first instance, treat λ as an unknown parameter and obtain an estimate $\hat{\lambda}$ from the data;
- Perform the usual regression diagnostics on the transformed data.
  - If they are satisfactory, adopt $\hat{\lambda}$.
  - If they are not satisfactory, conclude that no suitable transformation is available.
- When a transformation is adopted, $\hat{\lambda}$ is treated as a known constant and analysis of the transformed data proceeds in the usual way.

Remark A more refined approach would be to treat $\hat{\lambda}$ as a parameter estimate throughout the entire analysis. However, this would make the analyses substantially more complicated and Box and Cox (1964) argued that the benefits would be very small.

4.5.1 The profile likelihood

The method of estimation for λ is a variant on maximum likelihood called profile likelihood and consists of the following steps:

1. Obtain the full log-likelihood function, $\ell(\lambda, \boldsymbol{\beta}, \sigma^2; \boldsymbol{y})$.
2. For fixed λ, let $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\sigma}^2_\lambda$ be the maximum likelihood estimates based on $\boldsymbol{y}^{(\lambda)}$.

3. The *profile likelihood* is defined by

$$\hat{\ell}(\lambda) = \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}^2_\lambda; \boldsymbol{y}).$$

4. Maximizing $\hat{\ell}(\lambda)$ gives the overall maximum likelihood estimate for λ.
5. The profile likelihood can also be used to obtain an approximate 95% confidence interval for λ. This allows us to choose a convenient nearby value for λ. For example, if $\hat{\lambda} = -0.483$ we might choose to use -0.5 etc.
6. In practice, when dealing with non-negative data, a constant a is sometimes added to all data values to avoid zeros.

Theorem 4.1 *The profile likelihood is given by*

$$\hat{\ell}(\lambda) = \textit{const} - \frac{n}{2}\log \mathrm{RSS}(z^{(\lambda)})$$

where RSS *is the residual sum of squares when*

$$z^{(\lambda)} = \frac{y^{(\lambda)}}{\dot{y}^{\lambda-1}}$$

is taken as response variable in the multiple regression

$$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

and $\dot{y}^{\lambda-1}$ *is the geometric mean of* $y_1^{\lambda-1}, \dots, y_n^{\lambda-1}$.

Proof: Let $\boldsymbol{x}_i$ be the ith row of X. The transformed regression model states

$$Y_i^{(\lambda)} \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2).$$

Applying the transformation rule for a continuous random variable yields

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i^{(\lambda)} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right) y_i^{\lambda-1}$$

The full log-likelihood is therefore,

$$\begin{aligned}\ell(\lambda, \boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) &= \log \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i^{(\lambda)} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right) y_i^{\lambda-1} \right\} \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^{(\lambda)} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + n\log\dot{y}^{\lambda-1}\end{aligned}$$

Maximising ℓ with respect to $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}}_\lambda = (X^TX)^{-1}X^T\boldsymbol{y}^{(\lambda)}$$

and then maximising with respect to σ^2 yields

$$\hat{\sigma^2_\lambda} = \frac{\mathrm{RSS}(\boldsymbol{y}^{(\lambda)})}{n}.$$

Substituting into ℓ, we obtain

$$\begin{aligned}
\hat{\ell}(\lambda) &= \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma^2_\lambda}; \boldsymbol{y}) \\
&= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \mathrm{RSS}(\boldsymbol{y}^{(\lambda)}) + \frac{n}{2}\log n - \frac{n}{2} + n\log \dot{y}^{\lambda-1} \\
&= \text{const} - \frac{n}{2}\log \frac{\mathrm{RSS}(\boldsymbol{y}^{(\lambda)})}{(\dot{y}^{\lambda-1})^2}
\end{aligned}$$

Finally, observe that $\mathrm{RSS}(c\boldsymbol{w}) = c^2\,\mathrm{RSS}(\boldsymbol{w})$ for any scalar x and vector $\boldsymbol{w}$.

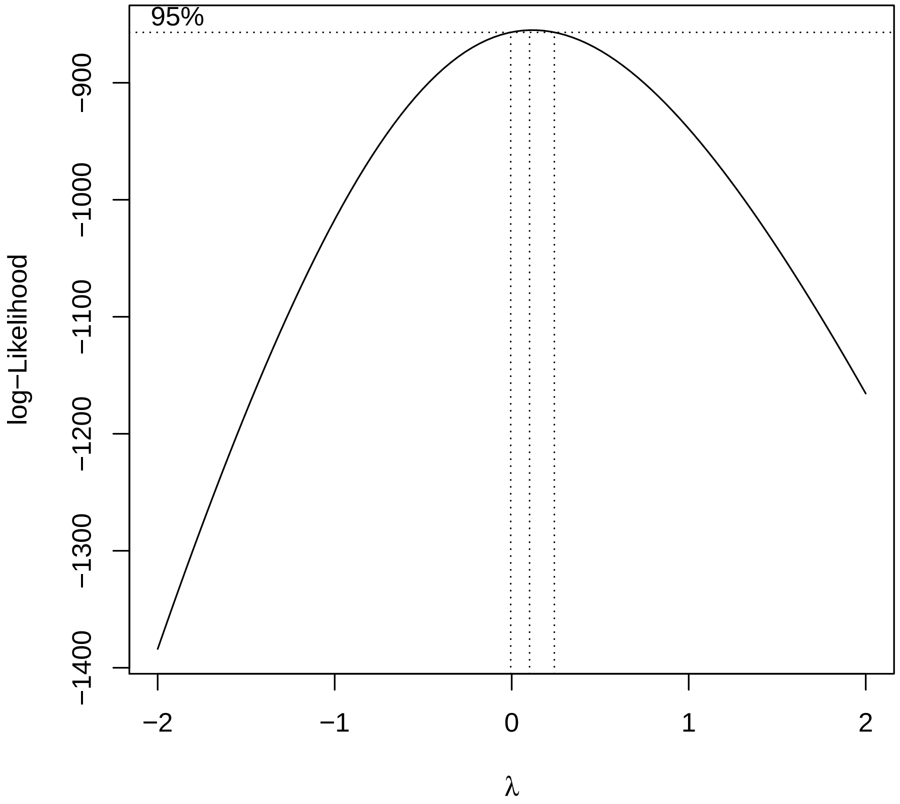
Hence

$$\hat{\ell}(\lambda) = \text{const} - \frac{n}{2}\log \mathrm{RSS}(\boldsymbol{z}^{(\lambda)}).$$

□

Example Consider the FEV example. In previous analyses we simply used `FEV` as the response variable. To investigate whether a transformation should be introduced, we can apply the Box-Cox analysis. This is performed in `R` as follows and produces the graphical output seen below.

```
# boxcox is requires "library(MASS)" to be entered first
boxcox(lm(FEV~Height+Sex+Smoker+Age,data=fev))
```



Since the 95% confidence interval does not contain 1, we deduce that a transformation is required in this case. Since the interval contains 0 a convenient choice would be to use `log(fev)` as the response. □

4.6 Generalised least squares

Thus far we have considered the linear model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$$

where $\boldsymbol{\mathcal{E}}$ is a random vector with

$$E(\boldsymbol{\mathcal{E}}) = \mathbf{0} \text{ and } \operatorname{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I.$$

In what follows we will relax the assumption $\operatorname{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I$ to $\operatorname{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 V$ where V is a known $n \times n$ positive definite, symmetric matrix.

Remarks The restrictions on V arise from its role as variance matrix.

1. Observe that
$$\sigma^2 v_{ij} = \operatorname{cov}(\mathcal{E}_i, \mathcal{E}_j) = \operatorname{cov}(\mathcal{E}_j, \mathcal{E}_i) = \sigma^2 v_{ji}.$$
Hence V must be a symmetric matrix.

2. The symmetric $n \times n$ matrix V is said to be *positive definite* if
$$\boldsymbol{a}^T V \boldsymbol{a} > 0 \text{ for all } \boldsymbol{a} \neq \mathbf{0}$$
and is said to be *non-negative definite* if
$$\boldsymbol{a}^T V \boldsymbol{a} \geq 0 \text{ for all } \boldsymbol{a}.$$
If $\boldsymbol{a}$ is a fixed vector then
$$0 \leq \operatorname{var}(\boldsymbol{a}^T \boldsymbol{\mathcal{E}}) = \sigma^2 \boldsymbol{a}^T V \boldsymbol{a}.$$
Hence in order to be a variance matrix V must be non-negative definite.

3. If V is non-negative definite but not positive definite, then there must exist $\boldsymbol{a} \neq \mathbf{0}$ for which $\operatorname{var}(\boldsymbol{a}^T \boldsymbol{\mathcal{E}}) = 0$. To eliminate any such redundancy, we assume that V is positive definite.

Suppose we wish to estimate $\boldsymbol{\beta}$. The ordinary least squares (OLS) estimate is

$$\hat{\boldsymbol{\beta}}_{OLS} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

It is easy to check that

$$E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta} \text{ and } \operatorname{Var}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2 (X^T X)^{-1} X^T V X (X^T X)^{-1}.$$

Hence $\hat{\boldsymbol{\beta}}_{OLS}$ is an unbiased linear estimator, but because the assumptions of the Gauss-Markov theorem do not hold, it is not the *best* linear unbiased estimator. To derive the BLUE, we transform the problem to one for which the Gauss-Markov theorem can be applied directly.

We will use the following facts about symmetric matrices.

1. Any square symmetric matrix V is expressible in the form

$$V = E\Lambda E^T$$

where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and E satisfies $E^T E = EE^T = I$. That is, E is orthogonal.

2. The matrix $V = E\Lambda E^T$ is non-negative definite if and only if $\lambda_i \geq 0$ and is positive definite if and only if $\lambda_i > 0$ for $i = 1, 2, \ldots, n$.

3. For any positive integer n, $V^n = E\Lambda^n E^T$ and, if V is positive definite, $V^{-1} = E\Lambda^{-1}E^T$.

4. For V positive definite the symmetric square root matrix and its inverse are defined by

$$V^{\frac{1}{2}} = E\Lambda^{\frac{1}{2}}E^T \text{ and } V^{-\frac{1}{2}} = E\Lambda^{-\frac{1}{2}}E^T$$

where

$$\Lambda^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}) \text{ and } \Lambda^{-\frac{1}{2}} = \operatorname{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \ldots, 1/\sqrt{\lambda_n})$$

It can be checked that $V^{\frac{1}{2}}$ and $V^{-\frac{1}{2}}$ are symmetric and

$$V^{\frac{1}{2}}V^{\frac{1}{2}} = V, \; V^{-\frac{1}{2}}V^{-\frac{1}{2}} = V^{-1}, \text{ and } V^{-\frac{1}{2}}VV^{-\frac{1}{2}} = I.$$

Consider now the regression model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \tag{3}$$

where

$$E(\boldsymbol{\mathcal{E}}) = \mathbf{0} \text{ and } \operatorname{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 V.$$

Pre-multiplying by $V^{-\frac{1}{2}}$ gives

$$\boldsymbol{Y}_* = X_*\boldsymbol{\beta} + \boldsymbol{\mathcal{E}_*} \tag{4}$$

where

$$\boldsymbol{Y}_* = V^{-\frac{1}{2}}\boldsymbol{Y}, \; X_* = V^{-\frac{1}{2}}X, \text{ and } \boldsymbol{\mathcal{E}}_* = V^{-\frac{1}{2}}\boldsymbol{\mathcal{E}}.$$

Note that $\boldsymbol{\beta}$ is the same in both (3) and (4).

Now applying the rules for linear transformation of random vectors, we find $E(\boldsymbol{\mathcal{E}}_*) = \mathbf{0}$ and

$$\begin{aligned} \operatorname{Var}(\boldsymbol{\mathcal{E}}_*) &= \operatorname{Var}(V^{-\frac{1}{2}}\boldsymbol{\mathcal{E}}) \\ &= V^{-\frac{1}{2}} \operatorname{Var}(\boldsymbol{\mathcal{E}}) \left\{V^{-\frac{1}{2}}\right\}^T \\ &= \sigma^2 V^{-\frac{1}{2}}VV^{-\frac{1}{2}} \\ &= \sigma^2 I. \end{aligned}$$

Hence the Gauss-Markov theorem applies in (4) and best linear unbiased estimator is therefore

$$\hat{\boldsymbol{\beta}} = (X_*^T X_*)^{-1}X_*^T\boldsymbol{y}_*$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X_*^T X_*)^{-1}.$$

Substituting for X_* and $\boldsymbol{y}_*$ produces the generalised least squares estimates

$$\hat{\boldsymbol{\beta}}_{GLS} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{y}$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{GLS}) = \sigma^2 (X^T V^{-1} X)^{-1}.$$

When V is a diagonal matrix, $V = \mathrm{diag}(v_1, v_2, \ldots, v_n)$ the generalised least squares estimator is called the weighted least squares estimator, with weights

$$w_i = 1/v_i.$$

Many computer packages allow for the direct specification of the weights w_i.

Example Suppose $Y_1, Y_2, \ldots, Y_n$ are independent with $E(Y_i) = \mu$ and $\mathrm{var}(Y_i) = \sigma_i^2$. The ordinary least squares estimate is just $\bar{Y}$ and it can be checked that $E(\bar{Y}) = \mu$ and $\mathrm{var}(\bar{Y}) = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2$.

The weighted least squares estimate is obtained by taking, $X = \mathbf{1}$, $\boldsymbol{\beta} = (\mu)$, and $V = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. It follows that the weighted least squares estimate is

$$\hat{\mu} = \sum_{i=1}^n a_i Y_i$$

where

$$a_i = \frac{1/\sigma_i^2}{\sum_{j=1}^n 1/\sigma_j^2}.$$

Intuitively, this estimator gives greater weight to observations with lower variance, compared to the OLS estimator which weights all observations equally. It can also be proved from first principles to be the best linear unbiased estimator for μ. (See Statistical Modelling and Inference II).

Preliminary reading for lecture 14

5 The geometry of least squares

5.1 Basic definitions and notation for vector spaces

Consider a linear model of the form $\boldsymbol{\eta} = X\boldsymbol{\beta}$. As was illustrated in the context of models involving factors, such a specification is not unique and can be given equivalently as $\boldsymbol{\eta} = X_*\boldsymbol{\beta}_*$ provided $X_* = XA$ where A is an invertible $p \times p$ matrix.

One formulation of the theory that avoids this ambiguity can be given in terms of the underlying linear subspaces.

Definition 5.1 *A vector space or linear space $\mathcal{M}$ is a set with the following properties.*

1. *There is an operation of vector addition with the following properties:*
   (a) $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M} \Rightarrow \boldsymbol{x}_1 + \boldsymbol{x}_2 \in \mathcal{M}$;

   (b) $\boldsymbol{x}_1 + (\boldsymbol{x}_2 + \boldsymbol{x}_3) = (\boldsymbol{x}_1 + \boldsymbol{x}_2) + \boldsymbol{x}_3$;

   (c) $\boldsymbol{x}_1 + \boldsymbol{x}_2 = \boldsymbol{x}_2 + \boldsymbol{x}_1$;

   (d) *There exists a neutral element* $\mathbf{0} \in \mathcal{M}$ *such that* $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$ *for all* $\boldsymbol{x} \in \mathcal{M}$;

   (e) *For every* $\boldsymbol{x} \in \mathcal{M}$ *there exists an inverse element denoted* $-\boldsymbol{x} \in \mathcal{M}$ *such that* $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$.

2. *There is an operation of scalar multiplication with the following properties:*
   (a) *For any* $\boldsymbol{x} \in \mathcal{M}$ *and any scalar* $\alpha \in \mathbb{R}$, $\alpha\boldsymbol{x} \in \mathcal{M}$;

   (b) $(\alpha_1 + \alpha_2)\boldsymbol{x} = \alpha_1\boldsymbol{x} + \alpha_2\boldsymbol{x}$

   (c) $(\alpha_1\alpha_2)\boldsymbol{x} = \alpha_1(\alpha_2\boldsymbol{x})$;

   (d) $\alpha(\boldsymbol{x}_1 + \boldsymbol{x}_2) = \alpha\boldsymbol{x}_1 + \alpha\boldsymbol{x}_2$.

□

Remarks

1. It can be checked that $\mathbb{R}^n$ is a linear space according to this definition.
2. Definition 5.1 is a very formal definition. In many situations the basic properties of addition and scalar multiplication are given. The key property is then,

   For every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}$ and $\alpha_1, \alpha_2 \in \mathbb{R}$ we have $\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 \in \mathcal{M}$.

Definition 5.2 *A subset $\mathcal{L} \subseteq \mathcal{M}$ is called a linear subspace (or vector subspace) if it is closed under scalar multiplication and addition. That is, if $\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 \in \mathcal{L}$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{L}$ and*

$\alpha_1, \alpha_2 \in \mathbb{R}$. □

Definition 5.3 *A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ is said to be **linearly independent** if*

$$\sum_{i=1}^{r} \alpha_i \boldsymbol{x}_i = \boldsymbol{0} \Rightarrow \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0.$$

□

Definition 5.4 *A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ is said to **span** the linear subspace $\mathcal{L}$ if every $\boldsymbol{v} \in \mathcal{L}$ is expressible as*

$$\boldsymbol{v} = \sum_{i=1}^{r} \alpha_i \boldsymbol{x}_i.$$

□

Definition 5.5 *A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ is called a **basis** for $\mathcal{L}$ if it spans $\mathcal{L}$ and is linearly independent.* □

Theorem 5.1 *If $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ and $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_s\}$ are bases for the linear subspace $\mathcal{L}$ then $r = s$.* □

Definition 5.6 *The **dimension** of a linear space $\mathcal{L}$ is the number of elements in any basis for $\mathcal{L}$.* □

In this framework, we define a linear model to be a model of the form

$$M : \boldsymbol{\eta} \in \mathcal{L}$$

where $\mathcal{L} \in \mathbb{R}^n$ is a linear subspace of dimension p. In terms of the matrix formulation, $\boldsymbol{\eta} = X\boldsymbol{\beta}$, it follows that

$$\mathcal{L} = \{\boldsymbol{\eta} : \ \boldsymbol{\eta} = X\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

That is $\mathcal{L}$ is the **column space** of X or in other words, the linear subspace spanned by the columns of X. We require the columns of X to be linearly independent to ensure that the columns of X form a basis for the **model space** $\mathcal{L}$. The ambiguity in the matrix formulation arises because there are infinitely many ways to choose a basis for a given vector space. When two model matrices X and X_* have the same column spaces the corresponding linear models are equivalent, although as previously discussed the parameters typically have different meanings and numerical values. In what follows, we outline the theory of estimation in terms of the underlying subspaces.

Definition 5.7 *If $\mathcal{L}_1$ and $\mathcal{L}_2$ are linear subspaces, then the vector space sum is defined by*

$$\mathcal{L}_1 + \mathcal{L}_2 = \{\boldsymbol{x}_1 + \boldsymbol{x}_2 : \ \boldsymbol{x}_1 \in \mathcal{L}_1, \ \boldsymbol{x}_2 \in \mathcal{L}_2\}.$$

It can be proved that $\mathcal{L}_1 + \mathcal{L}_2$ is a linear subspace.

If $\mathcal{L}_1 \cap \mathcal{L}_2 = \{\mathbf{0}\}$ then the vector space sum is called a ***direct sum*** *and we write it as $\mathcal{L}_1 \oplus \mathcal{L}_2$. For a direct sum, it can be proved that every $\boldsymbol{v} \in \mathcal{L}_1 \oplus \mathcal{L}_2$ is* ***uniquely*** *expressible in the form*

$$\boldsymbol{v} = \boldsymbol{x}_1 + \boldsymbol{x}_2 \text{ where } \boldsymbol{x}_1 \in \mathcal{L}_1,\ \boldsymbol{x}_2 \in \mathcal{L}_2.$$

□

Definition 5.8 *A linear space $\mathcal{M}$ is called an* ***inner product*** *space if there is a real-valued function $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ such that*

1. *$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \in \mathbb{R}$;*
2. *$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle$;*
3. $$\begin{array}{rcl} \langle \boldsymbol{x}, \boldsymbol{x} \rangle & \geq & 0 \\ & = & \text{if and only if } \boldsymbol{x} = \mathbf{0}; \end{array}$$
4. *$\langle \alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2, \boldsymbol{x_3} \rangle = \alpha_1 \langle \boldsymbol{x}_1, \boldsymbol{x}_3 \rangle + \alpha_2 \langle \boldsymbol{x}_2, \boldsymbol{x}_3 \rangle$.*

□

In this course, we are concerned exclusively with subspaces $\mathcal{L} \in \mathbb{R}^n$ and the usual "dot product" $\boldsymbol{x}_1 . \boldsymbol{x}_2 = \boldsymbol{x}_1^T \boldsymbol{x}_2$ is the inner product. More generally, it can also be shown that $\boldsymbol{x}_1^T V \boldsymbol{x}_2$ satisfies the axioms for an inner product if V is a positive definite, symmetric $n \times n$ matrix.

Definition 5.9 *Two vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ are said to be* ***orthogonal*** *if $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = 0$ and we write $\boldsymbol{x}_1 \perp \boldsymbol{x}_2$. If $\boldsymbol{x} \perp \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{L}$ then we write $\boldsymbol{x} \perp \mathcal{L}$.* □

Definition 5.10 *If $\mathcal{L} \subseteq \mathcal{M}$ is a linear subspace then the* ***orthogonal complement*** *is defined by*

$$\mathcal{L}^\perp = \{\boldsymbol{x} \in \mathcal{M} :\ \boldsymbol{x} \perp \mathcal{L}\}.$$

□

Theorem 5.2 *Suppose $\mathcal{L} \subseteq \mathcal{M}$ is a linear subspace. Then*

1. *$\mathcal{L}^\perp$ is a linear subspace;*
2. *$\mathcal{L} \cap \mathcal{L}^\perp = \{\mathbf{0}\}$;*
3. *$\dim(\mathcal{L}^\perp) + \dim(\mathcal{L}) = \dim(\mathcal{M})$;*
4. *$\mathcal{L} \oplus \mathcal{L}^\perp = \mathcal{M}$.*

□

Lecture 14, 2018

Definition 5.11 *If $\mathcal{L} \in \mathcal{M}$ is a linear subspace then the **orthogonal projection** on $\mathcal{L}$ is the linear mapping P defined by*

$$P\boldsymbol{y} = \boldsymbol{v}$$

where $\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{w}$ is the unique decomposition with $\boldsymbol{v} \in \mathcal{L}$ and $\boldsymbol{w} \in \mathcal{L}^\perp$. □

Theorem 5.3 *The linear mapping P is the orthogonal projection on its range if and only if it is symmetric and idempotent. That is, if and only if $P^2 = P^T = P$. Note that symmetry of the matrix P is a special case of the self-adjoint property $\langle P\boldsymbol{x}_1, \boldsymbol{x}_2\rangle = \langle \boldsymbol{x}_1, P\boldsymbol{x}_2\rangle$.* □

Proof: (of Theorem 5.3)

Suppose P is the orthogonal projection on $\mathcal{L}$ and observe if $\boldsymbol{v} \in \mathcal{L}$ then its unique decomposition is

$$\boldsymbol{v} = \boldsymbol{v} + \boldsymbol{0},\ \boldsymbol{v} \in \mathcal{L},\ \boldsymbol{0} \in \mathcal{L}^\perp$$

so that $P\boldsymbol{v} = \boldsymbol{v}$. Now consider any $\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{w}$ with $\boldsymbol{v} \in \mathcal{L}$ and $\boldsymbol{w} \in \mathcal{L}^\perp$. Then

$$P^2\boldsymbol{y} = P\boldsymbol{v} = \boldsymbol{v} = P\boldsymbol{y}.$$

That is $P^2\boldsymbol{y} = P\boldsymbol{y}$ for all $\boldsymbol{y}$ so $P^2 = P$. To establish that P is symmetric (self-adjoint), let

$$\boldsymbol{y}_1 = \boldsymbol{v}_1 + \boldsymbol{w}_1 \text{ and } \boldsymbol{y}_2 = \boldsymbol{v}_2 + \boldsymbol{w}_2$$

and observe

$$\langle P\boldsymbol{y}_1, \boldsymbol{y}_2\rangle = \langle \boldsymbol{v}_1, \boldsymbol{v}_2 + \boldsymbol{w}_2\rangle = \langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle + \langle \boldsymbol{v}_1, \boldsymbol{w}_2\rangle = \langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle.$$

Similarly

$$\langle \boldsymbol{y}_1, P\boldsymbol{y}_2\rangle = \langle \boldsymbol{v}_1 + \boldsymbol{w}_1, \boldsymbol{v}_2\rangle = \langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle + \langle \boldsymbol{w}_1, \boldsymbol{v}_2\rangle = \langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle.$$

so $\langle P\boldsymbol{y}_1, \boldsymbol{y}_2\rangle = \langle \boldsymbol{y}_1, P\boldsymbol{y}_2\rangle$ for all $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Hence, we must have $P = P^T$.

Conversely, suppose $P = P^2 = P$ and let

$$\mathcal{L} = \operatorname{col}(P) = \{\boldsymbol{v} = P\boldsymbol{a}, \boldsymbol{a} \in \mathbb{R}^n\}.$$

Now let $\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{w}$ for $\boldsymbol{v} \in \mathcal{L}$ and $\boldsymbol{w} \in \mathcal{L}^\perp$ and observe

$$P\boldsymbol{y} = P\boldsymbol{v} + P\boldsymbol{w}.$$

We will show that $P\boldsymbol{v} = \boldsymbol{v}$ and $P\boldsymbol{w} = \boldsymbol{0}$.

Now, if $\boldsymbol{v} \in \mathcal{L}$ then $\boldsymbol{v} = P\boldsymbol{a}$ for some $\boldsymbol{a}$. Hence,

$$P\boldsymbol{v} = P(P\boldsymbol{a}) = P^2\boldsymbol{a} = P\boldsymbol{a}$$

as required.

To show that $P\boldsymbol{w} = \mathbf{0}$, observe for any $\boldsymbol{x}$ that $P\boldsymbol{x} \in \mathcal{L}$ and hence $\langle P\boldsymbol{x}, \boldsymbol{w}\rangle = 0$. But

$$\langle P\boldsymbol{x}, \boldsymbol{w}\rangle = \langle \boldsymbol{x}, P\boldsymbol{w}\rangle$$

so we have

$$\langle \boldsymbol{x}, P\boldsymbol{w}\rangle = 0$$

for all $\boldsymbol{x}$. Hence, $P\boldsymbol{w} = \mathbf{0}$ as required.

□

Remark It should be noted that if P is the orthogonal projection on $\mathcal{L}$ then $(I - P)$ is the orthogonal projection on $\mathcal{L}^{\perp}$.

Theorem 5.4 *If P is the orthogonal projection on $\mathcal{L}$ then for any $\boldsymbol{y} \in \mathbb{R}^n$*

$$\hat{\boldsymbol{\eta}} = Py = \underset{\boldsymbol{\eta}\in\mathcal{L}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{\eta}\|^2.$$

Proof: Observe first

$$\|\boldsymbol{y} - \boldsymbol{\eta}\|^2 = \|\boldsymbol{y} - P\boldsymbol{y}\|^2 + \|P\boldsymbol{y} - \boldsymbol{\eta}\|^2 + 2\langle \boldsymbol{y} - P\boldsymbol{y}, P\boldsymbol{y} - \boldsymbol{\eta}\rangle.$$

Next, note that $\boldsymbol{\eta} \in \mathcal{L} \Rightarrow P\boldsymbol{\eta} = \boldsymbol{\eta}$ so that

$$\langle \boldsymbol{y} - P\boldsymbol{y}, P\boldsymbol{y} - \boldsymbol{\eta}\rangle = \langle (I - P)\boldsymbol{y}, P(\boldsymbol{y} - \boldsymbol{\eta})\rangle = 0$$

since $(I - P)\boldsymbol{y} \in \mathcal{L}^{\perp}$ and $P(\boldsymbol{y} - \boldsymbol{\eta}) \in \mathcal{L}$. Hence,

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{\eta}\|^2 &= \|\boldsymbol{y} - P\boldsymbol{y}\|^2 + \|P\boldsymbol{y} - \boldsymbol{\eta}\|^2 \\ &\geq \|\boldsymbol{y} - P\boldsymbol{y}\|^2 \\ &= \text{ if and only if } P\boldsymbol{y} = \boldsymbol{\eta}. \end{aligned}$$

□

Consider now the matrix formulation

$$M : \boldsymbol{\eta} = X\boldsymbol{\beta}.$$

The corresponding model subspace is the column space,

$$\mathcal{L} = \operatorname{col}(X).$$

Taking $P = X(X^TX)^{-1}X^T$, we can check that $P^2 = P^T = P$ and the range of P is $\mathcal{L}$ so that P is the orthogonal projection on $\mathcal{L}$. Moreover, if X_1 and X_2 are model matrices such that

$$\operatorname{col}(X_1) = \operatorname{col}(X_2) = \mathcal{L}$$

then we must have

$$P = X_1(X_1^TX_1)^{-1}X_1^T = X_2(X_2^TX_2)^{-1}X_2^T.$$

As discussed previously, $\operatorname{col}(X_1) = \operatorname{col}(X_2) = \mathcal{L}$ occurs when the columns of X_1 and X_2 are bases for same subspace $\mathcal{L}$.

5.2 Estimation of σ^2

In the parametric development of the linear model, we introduced the *residual mean square* estimator for σ^2,

$$s_e^2 = \frac{1}{n-p}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n-p}\|\boldsymbol{y} - \hat{\boldsymbol{\eta}}\|^2.$$

Substituting $\hat{\boldsymbol{\eta}} = P\boldsymbol{y}$ we obtain

$$s_e^2 = \frac{1}{n-p}\|\boldsymbol{y}-P\boldsymbol{y}\|^2 = \frac{1}{n-p}\|(I-P)\boldsymbol{y}\|^2 = \frac{1}{n-p}\{(I-P)\boldsymbol{y}\}^T(I-P)\boldsymbol{y} = \frac{1}{n-p}\boldsymbol{y}^T(I-P)\boldsymbol{y}.$$

A proof of unbiasedness can also be given in this framework. Suppose $\boldsymbol{\eta} \in \mathcal{L}$ and observe first $(I-P)\boldsymbol{\eta} = \mathbf{0}$ so that

$$\boldsymbol{y}^T(I-P)\boldsymbol{y} = (\boldsymbol{y}-\boldsymbol{\eta})^T(I-P)(\boldsymbol{y}-\boldsymbol{\eta}) = \operatorname{tr}\left\{(\boldsymbol{y}-\boldsymbol{\eta})^T(I-P)(\boldsymbol{y}-\boldsymbol{\eta})\right\} = \operatorname{tr}\left\{(I-P)(\boldsymbol{y}-\boldsymbol{\eta})(\boldsymbol{y}-\boldsymbol{\eta})^T\right\}.$$

Now if $E(\boldsymbol{Y}) = \boldsymbol{\eta}$ and $\operatorname{Var}(\boldsymbol{Y}) = \sigma^2 I$, it follows as previously that

$$E\left(\operatorname{tr}\left\{(I-P)(\boldsymbol{Y}-\boldsymbol{\eta})(\boldsymbol{Y}-\boldsymbol{\eta})^T\right\}\right) = \sigma^2\operatorname{tr}(I-P).$$

Since $I-P$ is the orthogonal projection on $\mathcal{L}^\perp$, it follows that $(I-P)$ has eigenvalues 0 with multiplicity $p = \dim(\mathcal{L})$ and 1 with multiplicity $n-p = \dim\mathcal{L}^\perp$. The result is then proved using the fact that the trace of a real symmetric matrix is the sum of its eigenvalues.

5.3 Estimable functions

The geometric approach to linear models avoids reference to regression parameters $\boldsymbol{\beta}$ in the formulation of the model. However, such quantities can be formulated as *estimable functions*. As previously, an estimable function is defined to be a linear combination of the form

$$\boldsymbol{\lambda}^T\boldsymbol{\eta}$$

where $\boldsymbol{\lambda}$ is a fixed vector. The key distributional results parallel Theorem 2.2

Theorem 5.5 *Suppose $\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\mathcal{E}}$ where $\boldsymbol{\eta} \in \mathcal{L}$, $E(\boldsymbol{\mathcal{E}}) = \mathbf{0}$ and $\operatorname{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I$. If $\hat{\boldsymbol{\eta}} = P\boldsymbol{y}$ and $\boldsymbol{\lambda} \in \mathbb{R}^n$ is a fixed vector then:*

1. *$E(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) = \boldsymbol{\lambda}^T\boldsymbol{\eta}$;*
2. *$\operatorname{var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}) = \sigma^2\|P\boldsymbol{\lambda}\|^2$;*
3. *$E(s_e^2) = \sigma^2$.*

If, further, $\boldsymbol{\mathcal{E}} \sim N_n(\mathbf{0}, \sigma^2 I)$ then:

4. *$\boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}} \sim N(\boldsymbol{\lambda}^T\boldsymbol{\eta}, \sigma^2\|P\boldsymbol{\lambda}\|^2)$;*
5. *$\frac{(n-p)s_e^2}{\sigma^2} \sim \chi^2_{n-p}$;*
6. *$\hat{\boldsymbol{\eta}}$ and s_e^2 are independent.*

□

5.4 Hypothesis tests

Consider linear subspaces $\mathcal{L}_0 \subset \mathcal{L} \subset \mathbb{R}^n$ of dimensions $p_0 < p$ and consider the model

$$M : \ \eta \in \mathcal{L}$$

and hypothesis

$$H_0 : \ \eta \in \mathcal{L}_0.$$

The analysis of variance table is

Source	SS	DF	MS	F-ratio
H_0 vs M	$\boldsymbol{y}^T(P-P_0)\boldsymbol{y}$	$p-p_0$	$MS_H = \frac{\boldsymbol{y}^T(P-P_0)\boldsymbol{y}}{p-p_0}$	$F = \frac{MS_H}{MS_E}$
Residual Error	$\boldsymbol{y}^T(I-P)\boldsymbol{y}$	$n-p$	$MS_E = \frac{\boldsymbol{y}^T(I-P)\boldsymbol{y}}{n-p}$	
Total	$\boldsymbol{y}^T(I-P_0)\boldsymbol{y}$	$n-p_0$		

Remark In the matrix formulation, we considered hypothesis of the form $\beta_p = \beta_{p-1} = \dots \beta_{p_0+1} = 0$. This corresponds to the requirement $\mathcal{L}_0 \subset \mathcal{L}$.

Example Consider the one-way layout with r groups and n_k observations in group k for $k = 1, \dots, r$. The usual model and hypothesis are

$$M : \ \eta_{ij} = \mu + \alpha_i \text{ and } H_0 : \ \alpha_1 = \alpha_2 = \dots = \alpha_r = 0.$$

The underlying subspaces can be specified by their bases. In particular $\mathcal{L}$ is the r dimensional subspace with basis

$$\begin{array}{ll} \text{Obs} & \begin{matrix} \ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_r \end{matrix} \\ \begin{matrix} (1,1) \\ (1,2) \\ \vdots \\ (1,n_1) \\ (2,1) \\ (2,2) \\ \vdots \\ (2,n_2) \\ \vdots \\ (r,1) \\ (r,2) \\ \vdots \\ (r,n_r) \end{matrix} & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \end{array}$$

and $\mathcal{L}_0$ is the 1-dimensional space spanned by $\mathbf{1}$. It can be shown that the corresponding ANOVA table is (as usual)

Source	SS	DF
Between Groups	$\sum_{i=1}^{r} n_i(y_{i\bullet} - y_{\bullet\bullet})^2$	$r-1$
Within Groups	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - y_{i\bullet})^2$	$n-r$
Total	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - y_{\bullet\bullet})^2$	$n-1$

□

5.5 Expected mean squares

In what follows we will outline the derivation of the distribution of the sum of squares entries and in particular the expected values for the mean squared entries.

Consider now the residual sum of squares, $Q_E = \boldsymbol{y}^T(I-P)\boldsymbol{y}$. Assuming the model $\boldsymbol{\eta} \in \mathcal{L} \Leftrightarrow (I-P)\boldsymbol{\eta} = \mathbf{0}$ to be correct, we have

$$Q_E = (\boldsymbol{y} - \boldsymbol{\eta})^T(I-P)(\boldsymbol{y} - \boldsymbol{\eta}).$$

Next, let $A_{n\times(n-p)}$ be a matrix whose columns form an *orthonormal basis* for $\mathcal{L}^{\perp}$. It follows that

$$A^TA = I_{n-p}$$

and hence AA^T is symmetric and idempotent and therefore the orthogonal projection on its column space, $\mathcal{L}^{\perp}$. That is

$$(I-P) = AA^T.$$

Hence,

$$Q_E = (\boldsymbol{y} - \boldsymbol{\eta})^T AA^T(\boldsymbol{y} - \boldsymbol{\eta}) = \|A^T(\boldsymbol{y} - \boldsymbol{\eta})\|^2.$$

Now suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{\eta}, \sigma^2 I)$ and let $\boldsymbol{V} = A^T(\boldsymbol{Y} - \boldsymbol{\eta})$ and observe that

$$\boldsymbol{V} \sim N_{n-p}(\mathbf{0}, \sigma^2 A^TA) \equiv N_{n-p}(\mathbf{0}, \sigma^2 I).$$

Finally, observe

$$\frac{Q_E}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n-p} V_i^2 \sim \chi^2_{n-p}$$

and, in particular,

$$E(Q_E) = (n-p)\sigma^2 \text{ and } E(MS_E) = \sigma^2.$$

For the hypothesis sum of squares,

$$\begin{aligned} Q_H &= \boldsymbol{y}^T(P-P_0)\boldsymbol{y} \\ &= (\boldsymbol{y} - \boldsymbol{\eta} + \boldsymbol{\eta})^T(P-P_0)(\boldsymbol{y} - \boldsymbol{\eta} + \boldsymbol{\eta}) \\ &= (\boldsymbol{y} - \boldsymbol{\eta})^T(P-P_0)(\boldsymbol{y} - \boldsymbol{\eta}) + \boldsymbol{\eta}^T(P-P_0)\boldsymbol{\eta} + 2\boldsymbol{\eta}^T(P-P_0)(\boldsymbol{y} - \boldsymbol{\eta}). \end{aligned}$$

Note that $H_0 : \boldsymbol{\eta} \in \mathcal{L}_0$ corresponds to the condition $(P-P_0)\boldsymbol{\eta} = \mathbf{0}$ but we do not assume H_0 to be true in the following calculations. Now suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{\eta}, \sigma^2 I)$ and observe,

- By the similar arguments to before $E\left((\boldsymbol{Y}-\boldsymbol{\eta})^T(P-P_0)(\boldsymbol{Y}-\boldsymbol{\eta})\right)=(p-p_0)\sigma^2$;
- $E\left(\boldsymbol{\eta}^T(P-P_0)(\boldsymbol{Y}-\boldsymbol{\eta})\right)=\boldsymbol{\eta}^T(P-P_0)E(\boldsymbol{Y}-\boldsymbol{\eta})=\mathbf{0}$.

Hence, we have

$$E(Q_H)=(p-p_0)\sigma^2+\boldsymbol{\eta}^T(P-P_0)\boldsymbol{\eta}$$

and

$$E(MS_H)=\sigma^2+\frac{\boldsymbol{\eta}^T(P-P_0)\boldsymbol{\eta}}{p-p_o}.$$

The preceding calculations can be summarised by including a column of expected mean squared entries in the ANOVA table:

Source	SS	DF	E(MS)
$H_0\ vs\ M$	$\boldsymbol{y}^T(P-P_0)\boldsymbol{y}$	$p-p_0$	$\sigma^2+\frac{\boldsymbol{\eta}^T(P-P_0)\boldsymbol{\eta}}{p-p_0}$
Residual Error	$\boldsymbol{y}^T(I-P)\boldsymbol{y}$	$n-p$	σ^2
Total	$\boldsymbol{y}^T(I-P_0)\boldsymbol{y}$	$n-p_0$	

Remarks

- It is possible to show for the hypothesis sum of squares that

  $$\frac{Q_H}{\sigma^2}\sim\chi^2_{p-p_0,\delta^2}$$

  independently of the residual sum of squares, where the non-centrality parameter is given by

  $$\delta^2=\frac{\boldsymbol{\eta}^T(P-P_0)\boldsymbol{\eta}}{\sigma^2}.$$

  The detailed proof of this result is beyond the scope of this course.
- When H_0 is true, $\delta^2=0$ and it follows immediately that the F-ratio $F=\frac{MS_H}{MS_E}\sim F_{p-p_0,n-p}$.
- In cases where there are explicit expressions for the least squares estimates $\hat{\boldsymbol{\eta}}=P\boldsymbol{y}$ and $\hat{\boldsymbol{\eta}}_0=P_0\boldsymbol{y}$ respectively, we can calculate the noncentrality term explicitly using

  $$\boldsymbol{\eta}^T(P-P_0)\boldsymbol{\eta}=\sum(P\boldsymbol{\eta}-P_0\boldsymbol{\eta})^2.$$

  In other words, we can substitute the $\boldsymbol{\eta}$ for y in the expressions for the sums of squares.

Example Consider the one-way layout with r groups and n_k observations in group k for $k=1,\ldots,r$. As shown previously, the ANOVA table is

Source	SS	DF
Between Groups	$\sum_{i=1}^r n_i(y_{i\bullet}-y_{\bullet\bullet})^2$	$r-1$
Within Groups	$\sum_{i=1}^r\sum_{j=1}^{n_i}(y_{ij}-y_{i\bullet})^2$	$n-r$
Total	$\sum_{i=1}^r\sum_{j=1}^{n_i}(y_{ij}-y_{\bullet\bullet})^2$	$n-1$

The mean-square entry for the within-group sum of squares is just σ^2 and for the between-group sum of squares we have,

$$\mathsf{MS} = \sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\eta_{i\bullet} - \eta_{\bullet\bullet})^2.$$

Taking $\eta_{ij} = \mu + \alpha_i$ we find

$$\eta_{i\bullet} = \mu + \alpha_i \text{ and } \eta_{\bullet\bullet} = \mu + \bar{\alpha}$$

where

$$\bar{\alpha} = \frac{\sum_{i=1}^{r} n_i\alpha_i}{\sum_{i=1}^{r} n_i}.$$

These calculations are often presented in the extended ANOVA table:

Source	SS	DF	E(MS)
Between Groups	$\sum_{i=1}^{r} n_i(y_{i\bullet} - y_{\bullet\bullet})^2$	$r-1$	$\sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\alpha_i - \bar{\alpha})^2$
Within Groups	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - y_{i\bullet})^2$	$n-r$	σ^2
Total	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - y_{\bullet\bullet})^2$	$n-1$	

□

Lecture 16, 2018

In many applications we are concerned with several hypotheses. For example, consider the additive model for the two-way layout,

$$\eta_{ij} = \mu + \alpha_i + \beta_j$$

and the hypotheses

$$H_1 : \beta_1 = \beta_2 = \ldots \beta_s = 0; \text{ (no column effects)}$$

$$H_2 : \alpha_1 = \alpha_2 = \ldots \alpha_r = 0; \quad \text{(no row effects)}$$

$$H_0 : \alpha_1 = \ldots = \alpha_r = \beta_1 = \ldots = \beta_s = 0 \text{ (no row or column effects)}$$

In general, such hypotheses are generated by two subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$ and we consider the following hypotheses:

Model	Subspace	Dimension
M	$\mathcal{L}_1 + \mathcal{L}_2$	$r_1 + r_2 - r_0$
H_1	$\mathcal{L}_1$	r_1
H_2	$\mathcal{L}_2$	r_2
H_0	$\mathcal{L}_1 \cap \mathcal{L}_2$	r_0

In this situation, two different ANOVA tables can be constructed giving, in general, different sequential decompositions of the total sum of squares.

In particular, we may begin by performing a test of H_1 vs M and, if this hypothesis is accepted, we proceed with a second test of H_0 vs H_1. The corresponding ANOVA decomposition is shown below.

Source	SS	DF
$H_0 \ vs \ H_1$	$\boldsymbol{y}^T(P_1 - P_0)\boldsymbol{y}$	$r_1 - r_0$
$H_1 \ vs \ M$	$\boldsymbol{y}^T(P - P_1)\boldsymbol{y}$	$r_2 - r_0$
Residual Error	$\boldsymbol{y}^T(I - P)\boldsymbol{y}$	$n - r_1 - r_2 + r_0$
Total	$\boldsymbol{y}^T(I - P_0)\boldsymbol{y}$	$n - r_0$

Alternatively we might proceed by performing a test of H_2 vs M and, if this hypothesis is accepted, we proceed with a second test of H_0 vs H_2. The corresponding ANOVA decomposition is shown below.

Source	SS	DF
$H_0 \ vs \ H_2$	$\boldsymbol{y}^T(P_2 - P_0)\boldsymbol{y}$	$r_2 - r_0$
$H_2 \ vs \ M$	$\boldsymbol{y}^T(P - P_2)\boldsymbol{y}$	$r_1 - r_0$
Residual Error	$\boldsymbol{y}^T(I - P)\boldsymbol{y}$	$n - r_1 - r_2 + r_0$
Total	$\boldsymbol{y}^T(I - P_0)\boldsymbol{y}$	$n - r_0$

The sums of squares shown in the preceding ANOVA tables are called sequential sums of squares and the order in which the are considered is important.

When the two hypotheses arise on an equal footing, the following sequence of tests can be performed to determine the extent to which M can be simplified.

- Test H_1 vs M.
- If H_1 is *accepted* test H_0 vs H_1.
- If H_1 is *rejected*, test H_2 vs M.

However, it is equally valid to consider the sequence:

- Test H_2 vs M.
- If H_2 is *accepted* test H_0 vs H_2.
- If H_2 is *rejected*, test H_1 vs M.

In general the two sequential approaches are not guaranteed to produce the same conclusion.

Example Employees in a certain company were classified according to `Sex` (Female/Male) and `Education` (degree/no degree) and the annual salary (in $US) was recorded. Shown below is a transcript from an `R` analysis.

```
salary_dat<-read.csv("salary.csv",header=T)
salary_dat

##    Salary    Sex Degree
## 1      24 Female    Yes
## 2      26 Female    Yes
## 3      25 Female    Yes
## 4      24 Female    Yes
## 5      27 Female    Yes
## 6      24 Female    Yes
## 7      27 Female    Yes
## 8      23 Female    Yes
## 9      15 Female     No
## 10     17 Female     No
## 11     20 Female     No
## 12     16 Female     No
## 13     25   Male    Yes
## 14     29   Male    Yes
## 15     27   Male    Yes
## 16     19   Male     No
## 17     18   Male     No
## 18     21   Male     No
## 19     20   Male     No
## 20     21   Male     No
## 21     22   Male     No
## 22     19   Male     No
```

```
summary(aov(Salary~Sex+Degree,data=salary_dat))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Sex          1   0.30    0.30    0.11    0.743
## Degree       1 272.39  272.39  101.13 4.81e-09 ***
## Residuals   19  51.17    2.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Salary~Degree+Sex,data=salary_dat))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Degree       1 242.23  242.23   89.93 1.23e-08 ***
## Sex          1  30.46   30.46   11.31  0.00327 **
## Residuals   19  51.17    2.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the first analysis aov(Salary~Sex+Degree) the Degree sum of squares is used to test for whether the Degree term can be removed from the model Salary~Sex+Degree. The Sex term would be used to test the null hypothesis Salary~1 against Salary~Sex. However, this test is not applicable, since the first test clearly establishes that Degree cannot be removed from the model, and hence there is no interest in trying to simplify the incorrect model.

The interpretation of the second analysis of variance table is analogous and in this case both analyses lead to the conclusion that the model Salary~Sex+Degree cannot be simplified.

However, comparison of the two analyses shows the sums of squares attributed to Sex to be numerically very different. The fact that both terms are simply labelled "Sex" is potentially confusing. In particular, the Sex term in the first analysis represents the effect of Sex *without adjustment for* Degree and the term in the second analysis represents the effect of Sex *after adjustment for* Degree. To complete the example, consider the multiple regression output for the two regression models aov(Salary~Sex+Degree) and aov(Salary~Sex).

```
summary(lm(Salary~Degree+Sex,data=salary_dat))

##
## Call:
## lm(formula = Salary ~ Degree + Sex, data = salary_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3916 -0.8531 -0.3427  1.1678  2.7063
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.2937     0.6896  25.077 5.03e-16 ***
## DegreeYes     7.5594     0.7517  10.056 4.81e-09 ***
## SexMale       2.5385     0.7548   3.363  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.641 on 19 degrees of freedom
## Multiple R-squared:  0.842,Adjusted R-squared:  0.8254
## F-statistic: 50.62 on 2 and 19 DF,  p-value: 2.441e-08
```

```
summary(lm(Salary~Sex,data=salary_dat))

##
## Call:
## lm(formula = Salary ~ Sex, data = salary_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3333 -2.9083  0.2833  2.8417  6.9000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.3333     1.1611  19.234 2.27e-14 ***
## SexMale      -0.2333     1.7222  -0.135    0.894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.022 on 20 degrees of freedom
## Multiple R-squared:  0.000917,Adjusted R-squared:  -0.04904
## F-statistic: 0.01836 on 1 and 20 DF,  p-value: 0.8936
```

In the multiple regression analysis there is a signficant effect due to `Sex`. The estimated coefficient 2.5385 means that, on average, salaries for males were $2,538.50 higher than for females *with the same qualification*. On the other hand in the simple regression analysis, the estimated coefficient -0.2333 is not significant but suggests that overall males salaries are lower than for females. The two analyses can be reconciled as follows:

- For any given level of education, salaries for males are higher than for females.
- Salaries for degree qualified employees are higher than for non degree qualified employees.
- Most female employees had degrees but most male employees did not.

Lecture 17, 2018

5.5.1 Orthogonality

The previous example illustrates the fact that the sequential sums of squares in an ANOVA table can depend on the order in which the terms are included in the model. However, if the $\mathcal{L}_1$ and $\mathcal{L}_2$ are such that

$$P = P_1 + P_2 - P_0$$

then the corresponding sums of squares will be identical. In this case we say that the hypotheses H_1 and H_2 are orthogonal.

Theorem 5.6 *The hypotheses H_1 and H_2 are orthogonal if and only if*

$$\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \perp \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}.$$

Proof: Observe first that we can always write

$$\mathbb{R}^n = \{\mathcal{L}_1 + \mathcal{L}_2\}^{\perp} \oplus \{\mathcal{L}_1 + \mathcal{L}_2\}.$$

Similarly, we have

$$\mathcal{L}_1 = \{\mathcal{L}_1 \cap \mathcal{L}_2\} \oplus \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \text{ and } \mathcal{L}_2 = \{\mathcal{L}_1 \cap \mathcal{L}_2\} \oplus \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}$$

and hence

$$\mathcal{L}_1 + \mathcal{L}_2 = \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \oplus \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \oplus \{\mathcal{L}_1 \cap \mathcal{L}_2\}$$

Thus we have

$$\mathbb{R}^n = \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \oplus \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \oplus \{\mathcal{L}_1 \cap \mathcal{L}_2\} \oplus \{\mathcal{L}_1 + \mathcal{L}_2\}^{\perp}$$

so any $\boldsymbol{y} \in \mathbb{R}^n$ is uniquely expressible as

$$\boldsymbol{y} = \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_0 + \boldsymbol{w}$$

where

$$\begin{aligned} \boldsymbol{v_1} &\in \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \\ \boldsymbol{v_2} &\in \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \\ \boldsymbol{v}_0 &\in \{\mathcal{L}_1 \cap \mathcal{L}_2\} \\ \boldsymbol{w} &\in \{\mathcal{L}_1 + \mathcal{L}_2\}^{\perp}. \end{aligned}$$

Now, suppose

$$\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \perp \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}.$$

We will prove that $P = P_1 + P_2 - P_0$. By definition, $P\boldsymbol{y} = \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_0$ so it is sufficient to show

$$\begin{aligned} P_0\boldsymbol{y} &= \boldsymbol{v}_0 \\ P_1\boldsymbol{y} &= \boldsymbol{v}_1 + \boldsymbol{v}_0 \\ P_2\boldsymbol{y} &= \boldsymbol{v}_2 + \boldsymbol{v}_0 \end{aligned}$$

To show, $P_0\boldsymbol{y} = \boldsymbol{v}_0$, observe

$$\boldsymbol{v}_1 \in \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp,\ \boldsymbol{v}_2 \in \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp,\ \boldsymbol{w} \in \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp,\ \Rightarrow \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{w} \in \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp$$

and hence

$$P_0\boldsymbol{y} = \boldsymbol{v}_0.$$

To show $P_1\boldsymbol{y} = \boldsymbol{v}_1 + \boldsymbol{v}_0$ observe

$$\boldsymbol{v}_0 \in \mathcal{L}_1,\ \boldsymbol{v}_1 \in \mathcal{L}_1,\ \Rightarrow \boldsymbol{v}_0 + \boldsymbol{v}_1 \in \mathcal{L}_1$$

and

$$\boldsymbol{w} \in \mathcal{L}_1^\perp.$$

To obtain the result it is therefore sufficient to show that $\boldsymbol{v}_2 \in \mathcal{L}_1^\perp$. Observe first that

$$\boldsymbol{v}_2 \in \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp.$$

Clearly

$$\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp \perp \{\mathcal{L}_1 \cap \mathcal{L}_2\}$$

and, by assumption,

$$\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp \perp \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp$$

so that

$$\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp \perp \{\mathcal{L}_1 \cap \mathcal{L}_2\} + \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp = \mathcal{L}_1$$

as required. The proof that

$$P_2\boldsymbol{y} = \boldsymbol{v}_2 + \boldsymbol{v}_0$$

is similar.

Conversely, suppose

$$P = P_1 + P_2 - P_0$$

and consider

$$\boldsymbol{v}_1 \in \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp.$$

Using the facts that

$$P\boldsymbol{v}_1 = P_1\boldsymbol{v}_1 = \boldsymbol{v}_1$$

and $P_0\boldsymbol{v}_1 = \mathbf{0}$ it follows that $P_2\boldsymbol{v}_1 = \mathbf{0}$ and hence

$$\boldsymbol{v}_1 \in \mathcal{L}_2^\perp.$$

Hence

$$\boldsymbol{v}_1 \perp \mathcal{L}_2 \Rightarrow \boldsymbol{v}_1 \perp \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp \Rightarrow \mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp \perp \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^\perp$$

as required. □

Example Consider the two-way layout with one observation per cell and the no-interaction model,

$$\eta_{ij} = \mu + \alpha_i + \beta_j.$$

This model can be specified as

$$M: \ \boldsymbol{\eta} \in \mathcal{L}_1 + \mathcal{L}_2$$

where

$$\mathcal{L}_1 = \{\boldsymbol{\eta} : \eta_{ij} = \alpha_i\} \text{ and } \mathcal{L}_2 = \{\boldsymbol{\eta} : \eta_{ij} = \beta_j\}$$

It is easy to see that $\mathcal{L}_1 \cap \mathcal{L}_2 = \mathcal{S}(\mathbf{1})$. That is, the linear space spanned by the single vector $\mathbf{1}$. Moreover, bases for $\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}$ and $\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}$ are given respectively by the columns of

$$L_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\[1ex] 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\[1ex] \vdots & \vdots & & \vdots \\[1ex] 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\[1ex] -1 & -1 & \dots & -1 \\ -1 & -1 & \dots & -1 \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \dots & -1 \\ -1 & -1 & \dots & -1 \end{pmatrix} \text{ and } L_2 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \\[1ex] 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \\[1ex] \vdots & \vdots & & \vdots \\[1ex] 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \\[1ex] 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{pmatrix}$$

and it follows by inspection that

$$\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \perp \mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}.$$

Thus for a two-way layout with one observation per cell, the row effects and column effects are orthogonal. To provide a numerical illustration, consider the following example of a 12×3 layout.

According to Larsen and Marx (1986):

> In folklore, the full moon is often portrayed as something sinister, a kind of evil force possessing the power to control our behaviour. Over the centuries, many prominent writers and philosophers have shared this belief. Milton, in Paradise Lost, refers to
>
> *Demoniac frenzy, moping melancholy And moon-struck madness.*
>
> And Othello, after the murder of Desdemona, laments
>
> *It is the very error of the moon*
> *She comes more near the earth than she was want*
> *And makes men mad.*
>
> On a more scholarly level, Sir William Blackstone, the renowned eighteenth century English barrister, defined a "lunatic" as
>
> *one who hath ... lost the use of his reason and who hath lucid intervals, sometimes enjoying his senses and sometimes not, and that frequently depending upon changes of the moon.*
>
> The data give the admission rates to the emergency room of a Virginia mental health clinic before, during and after the 12 full moons from August 1971 to July 1972.

Shown below is a transcript of an R analysis of the data.

```
lunatic<-read.csv("lunatic.csv")
lunatic

##    Month   Moon Admission
## 1    Aug Before       6.4
## 2    Sep Before       7.1
## 3    Oct Before       6.5
## 4    Nov Before       8.6
## 5    Dec Before       8.1
## 6    Jan Before      10.4
## 7    Feb Before      11.5
## 8    Mar Before      13.8
## 9    Apr Before      15.4
## 10   May Before      15.7
## 11   Jun Before      11.7
## 12   Jul Before      15.8
## 13   Aug During       5.0
## 14   Sep During      13.0
## 15   Oct During      14.0
## 16   Nov During      12.0
## 17   Dec During       6.0
## 18   Jan During       9.0
## 19   Feb During      13.0
## 20   Mar During      16.0
## 21   Apr During      25.0
## 22   May During      14.0
## 23   Jun During      14.0
## 24   Jul During      20.0
## 25   Aug  After       5.8
## 26   Sep  After       9.2
## 27   Oct  After       7.9
## 28   Nov  After       7.7
## 29   Dec  After      11.0
## 30   Jan  After      12.9
## 31   Feb  After      13.5
## 32   Mar  After      13.1
## 33   Apr  After      15.8
## 34   May  After      13.3
## 35   Jun  After      12.8
## 36   Jul  After      14.5
```

```
summary(aov(Admission~Month+Moon,data=lunatic))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Month       11  455.6   41.42   7.129 5.08e-05 ***
## Moon         2   41.5   20.76   3.573   0.0453 *
## Residuals   22  127.8    5.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Admission~Moon+Month,data=lunatic))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Moon         2   41.5   20.76   3.573   0.0453 *
## Month       11  455.6   41.42   7.129 5.08e-05 ***
## Residuals   22  127.8    5.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# In this case there is a significant effect due to moon and a highly
#  significant effect due to month.
# Note that unlike in the salary example, the sums of squares do not depend
#  on the order in which they have been entered.
# Now look at the parameter estimates
summary(lm(Admission~Moon+Month,data=lunatic))

##
## Call:
## lm(formula = Admission ~ Moon + Month, data = lunatic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8528 -1.3590 -0.0236  1.1368  4.7806
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.2611     1.5031  12.149 3.14e-11 ***
## MoonBefore   -0.5417     0.9840  -0.550 0.587555
## MoonDuring    1.9583     0.9840   1.990 0.059149 .
## MonthAug    -13.0000     1.9681  -6.605 1.21e-06 ***
## MonthDec    -10.3667     1.9681  -5.267 2.76e-05 ***
## MonthFeb     -6.0667     1.9681  -3.083 0.005443 **
## MonthJan     -7.9667     1.9681  -4.048 0.000537 ***
## MonthJul     -1.9667     1.9681  -0.999 0.328522
## MonthJun     -5.9000     1.9681  -2.998 0.006627 **
## MonthMar     -4.4333     1.9681  -2.253 0.034593 *
## MonthMay     -4.4000     1.9681  -2.236 0.035841 *
## MonthNov     -9.3000     1.9681  -4.725 0.000103 ***
## MonthOct     -9.2667     1.9681  -4.708 0.000107 ***
## MonthSep     -8.9667     1.9681  -4.556 0.000155 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 22 degrees of freedom
## Multiple R-squared:  0.7955,Adjusted R-squared:  0.6746
## F-statistic: 6.581 on 13 and 22 DF,  p-value: 6.265e-05

summary(lm(Admission~Moon,data=lunatic))

##
## Call:
## lm(formula = Admission ~ Moon, data = lunatic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4167 -3.0021  0.5833  2.1771 11.5833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4583     1.2138   9.440 6.71e-11 ***
## MoonBefore   -0.5417     1.7165  -0.316    0.754
## MoonDuring    1.9583     1.7165   1.141    0.262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.205 on 33 degrees of freedom
## Multiple R-squared:  0.06643,Adjusted R-squared:  0.009851
## F-statistic: 1.174 on 2 and 33 DF,  p-value: 0.3217

# Note that the parameter estimates for Moon are identical in both models. However, the
#  standard errors are very different so there is no correspondence between the t-stats
```

5.6 Generalized least squares

The extension to generalised least squares can also be given a geometrical interpretation. In particular, the same theory applies but instead of using the standard inner product

$$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \boldsymbol{x}_1^T \boldsymbol{x}_2$$

we use

$$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_* = \boldsymbol{x}_1^T V^{-1} \boldsymbol{x}_2$$

The orthogonal_* projection

$$P = X(X^T V^{-1} X)^{-1} X^T V^{-1}$$

can be seen to satisfy idempotence $P^2 = P$ and the self adjoint property,

$$\langle P\boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_* = \langle \boldsymbol{x}_1, P\boldsymbol{x}_2 \rangle_*$$

Moreover, it can be checked that

$$P\boldsymbol{y} = \underset{\boldsymbol{\eta} \in \mathcal{L}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{\eta}\|_*^2$$

where $\mathcal{L}$ is the column space of X and

$$\|\boldsymbol{x}\|_*^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_* = \boldsymbol{x}^T V^{-1} \boldsymbol{x}.$$

Finally, it should be noted that the definition of

$$P\boldsymbol{x} = \boldsymbol{v}$$

from the decomposition

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{w}$$

with

$$\boldsymbol{v} \in \mathcal{L} \text{ and } \boldsymbol{w} \in \mathcal{L}^\perp$$

is the same. However, the space $\mathcal{L}^\perp$ is orthogonal with respect to the * inner product.

Lecture 18, 2018

6 Multistratum experiments

Multistratum experiments occur when there is more than one source of random variation in an experiment. A simple illustration is the split plot experiment.

Example Yates(1935) considered an experiment involving three varieties of oats (Marvellous, Victory, Golden Rain) and four levels of manure application (0, 0.2, 0.4, 0.6, cwt per acre). Note 1cwt=50.8kg.

The experiment was arranged in 6 blocks of three whole plots. Each whole plot was then split into four sub plots as shown below where each row corresponds to a whole plot.

		0.0	0.2	0.4	0.6
Block I	Victory	111	130	157	174
	Golden Rain	117	114	161	141
	Marvellous	105	140	118	156
Block II	Victory	61	91	97	100
	Golden Rain	70	108	126	149
	Marvellous	96	124	121	144
Block III	Victory	68	64	112	86
	Golden Rain	60	102	89	96
	Marvellous	89	129	132	124
Block IV	Victory	74	89	81	122
	Golden Rain	64	103	132	133
	Marvellous	70	89	104	117
Block V	Victory	62	90	100	116
	Golden Rain	80	82	94	126
	Marvellous	63	70	109	99
Block VI	Victory	53	74	118	113
	Golden Rain	89	82	86	104
	Marvellous	97	99	119	121

Now let Y_{ijk} be the yield for the subplot in block k with variety i and manure application level j. The model we consider is

$$y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \beta_k + d_{ik} + e_{ijk}$$

where d_{ik} are assumed to be independent $N(0, \sigma^2_{BM})$ realizations and e_{ijk} are assumed to be independent $N(0, \sigma^2)$ realizations.

The inclusion of the term d_{ik} is necessary in this case because we cannot assume that there are no `whole.plot*block` effects present. However, if we were to include a simple *fixed effect*, say δ_{ik} in the model, we would not be able to identify any variety effects because they would be completely confounded with the `whole.plot*block` effects.

The solution is therefore to introduce the `whole.plot*block` term as a *random effect* and the approriate analysis can be obtained in `R` as follows.

```
oats<-read.table("oats.txt",header=T)
head(oats)

##   Blocks    Variety Manure Yield
## 1      I    Victory 0.0cwt   111
## 2      I    Victory 0.2cwt   130
## 3      I    Victory 0.4cwt   157
## 4      I    Victory 0.6cwt   174
## 5      I GoldenRain 0.0cwt   117
## 6      I GoldenRain 0.2cwt   114

tail(oats)

##    Blocks    Variety Manure Yield
## 67     VI GoldenRain 0.4cwt    86
## 68     VI GoldenRain 0.6cwt   104
## 69     VI Marvellous 0.0cwt    97
## 70     VI Marvellous 0.2cwt    99
## 71     VI Marvellous 0.4cwt   119
## 72     VI Marvellous 0.6cwt   121

summary(aov(Yield~Manure*Variety+Error(Blocks/Variety),data=oats))

## 
## Error: Blocks
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  5  15875    3175               
## 
## Error: Blocks:Variety
##           Df Sum Sq Mean Sq F value Pr(>F)
## Variety    2   1786   893.2   1.485  0.272
## Residuals 10   6013   601.3               
## 
## Error: Within
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## Manure          3  20020    6673  37.686 2.46e-12 ***
## Manure:Variety  6    322      54   0.303    0.932    
## Residuals      45   7969     177                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on this analysis, we see that there is a significant effect due to `Manure` but no significant `Variety` effect or `Variety*Manure` interaction.

To see intuitively how the split plot ANOVA accounts for the different sources of variability, it is instructive to consider two other analyses.

First, suppose we wished to estimate only the variety effects. The presence of the random effect d_{ik} implies that there will be correlations between observations within the same whole-plot but, because the sub-plot treatments (Manure) are the same for each whole-plot, a valid approach would simply be to analyse the whole-plot averages as simple randomized block design.

```
whole_oat
```

```
##    blocks    variety  yield
## 3       I    Victory 143.00
## 1       I GoldenRain 133.25
## 2       I Marvellous 129.75
## 6      II    Victory  92.00
## 4      II GoldenRain  95.50
## 5      II Marvellous  85.25
## 9     III    Victory  91.50
## 7     III GoldenRain 108.00
## 8     III Marvellous  95.00
## 12     IV    Victory  82.50
## 10     IV GoldenRain  86.75
## 11     IV Marvellous 118.50
## 15      V    Victory  87.25
## 13      V GoldenRain 113.25
## 14      V Marvellous 121.25
## 18     VI    Victory  89.50
## 16     VI GoldenRain  90.25
## 17     VI Marvellous 109.00
```

```
summary(aov(yield~blocks+variety,data=whole_oat))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## blocks       5   3969   793.8   5.280 0.0124 *
## variety      2    447   223.3   1.485 0.2724
## Residuals   10   1503   150.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inspection of the ANOVA table shows the result be equivalent to the section of the split plot ANOVA table that uses the `Error:  Blocks:Variety` error term. In particular, the degrees of freedom are identical and the sums of squares are proportional so that the F-statistics coincide.

Now consider the `Manure` and `Manure:Variety` terms. Both of these terms would be estimable within a conventional fixed effects analysis and, as can be seen below, this coincides exactly with the section of the split plot ANOVA that uses the `Error:  Within` error term.

```
summary(aov(Yield~Manure*Variety+Blocks*Variety,data=oats))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Manure          3  20020    6673  37.686 2.46e-12 ***
## Variety         2   1786     893   5.044  0.01056 *
## Blocks          5  15875    3175  17.930 9.53e-10 ***
## Manure:Variety  6    322      54   0.303  0.93220
## Variety:Blocks 10   6013     601   3.396  0.00225 **
## Residuals      45   7969     177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In particular the degrees of freedom and sums of squares are identical. In this light, it is natural to wonder whether any sensible meaning can be given to the `Variety` main effect that now has a "significant" F-statistic in this analysis. The answer is no, for the following reasons:

- First, in the context of the fixed effects analysis, it is not valid to test for the absence of a `Variety` effect because there is a significant `Variety:Block` interaction in the model.
- Second, it cannot be interpreted as a valid test in the context of the split plot model with a random `Variety:Block` term because the denominator does not contain the relevant error sum of squares. In particular, the residual sum of squares for the fixed effects ANOVA does not include the component due to `Variety:Block`.

To formalise these considerations it is useful to consider the expected mean squared entries for the ANOVA table.

Lecture 19, 2018

6.1 Expected mean squares for the split plot experiment

Consider first the fixed effects model

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \beta_k + \delta_{ik} + \mathcal{E}_{ijk}$$

where

$$\sum_i \alpha_i = \sum_j \tau_j = \sum_k \beta_k = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = \sum_i \delta_{ik} = \sum_k \delta_{ik} = 0$$

and $\mathcal{E}_{ijk} \sim N(0, \sigma^2)$ independently for $i = 1, \ldots, r;\ j = 1, \ldots, s; k = 1, \ldots, b$. Consider also the nested sequence of hypotheses

$$\begin{array}{rl}
M: & \eta_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \beta_k + \delta_{ik} \\
H_1: & \eta_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \delta_{ik} \\
H_2: & \eta_{ijk} = \mu + \alpha_i + \beta_k + \delta_{ik} \\
H_3: & \eta_{ijk} = \mu + \alpha_i + \beta_k \\
H_4: & \eta_{ijk} = \mu + \beta_k \\
H_5: & \eta_{ijk} = \mu
\end{array}$$

The least-squares estimates are given by

$$\begin{array}{ll}
M & \hat{\eta}_{ijk} = y_{ij\bullet} + y_{i\bullet k} - y_{i\bullet\bullet} \\
H_1 & \hat{\eta}^{(1)}_{ijk} = y_{i\bullet k} + y_{\bullet j\bullet} - y_{\bullet\bullet\bullet} \\
H_2 & \hat{\eta}^{(2)}_{ijk} = y_{i\bullet k} \\
H_3 & \hat{\eta}^{(3)}_{ijk} = y_{i\bullet\bullet} + y_{\bullet\bullet k} - y_{\bullet\bullet\bullet} \\
H_4 & \hat{\eta}^{(4)}_{ijk} = y_{\bullet\bullet k} \\
H_5 & \hat{\eta}^{(5)}_{ijk} = y_{\bullet\bullet\bullet}
\end{array}$$

Each of these expressions can be verified by observing that the orthogonal projection of $\boldsymbol{y}$ on $\mathcal{L}$ is defined uniquely by the conditions

$$\hat{\boldsymbol{\eta}} \in \mathcal{L} \text{ and } \boldsymbol{y} - \hat{\boldsymbol{\eta}} \perp \mathcal{L}.$$

To confirm the least squares estimates under M observe first that

$$\hat{\eta}_{ijk} = y_{ij\bullet} + (y_{i\bullet k} - y_{i\bullet\bullet}) = \alpha^*_{ij} + \delta^*_{ik} \Rightarrow \hat{\boldsymbol{\eta}} \in \mathcal{L}.$$

To verify $\boldsymbol{y} - \hat{\boldsymbol{\eta}} \perp \mathcal{L}$, observe

$$\begin{aligned}
\boldsymbol{\eta}^T(\boldsymbol{y} - \hat{\boldsymbol{\eta}}) &= \sum_{ijk}(\mu + \alpha_i + \tau_j + \gamma_{ij} + \beta_k + \delta_{ik})(y_{ijk} - y_{ij\bullet} - y_{i\bullet k} + y_{i\bullet\bullet}) \\
&= \sum_{ij}(\mu + \alpha_i + \tau_j + \gamma_{ij}) \sum_k (y_{ijk} - y_{ij\bullet} - y_{i\bullet k} + y_{i\bullet\bullet}) \\
&\quad + \sum_{ik}(\beta_k + \delta_{ik}) \sum_j (y_{ijk} - y_{ij\bullet} - y_{i\bullet k} + y_{i\bullet\bullet}) \\
&= \sum_{ij}(\mu + \alpha_i + \tau_j + \gamma_{ij}) \sum_k (by_{ij\bullet} - by_{ij\bullet} - by_{i\bullet\bullet} + by_{i\bullet\bullet}) \\
&\quad + \sum_{ik}(\beta_k + \delta_{ik})(sy_{i\bullet k} - sy_{i\bullet\bullet} - sy_{i\bullet k} + sy_{i\bullet\bullet}) \\
&= 0
\end{aligned}$$

for $\boldsymbol{\eta} \in \mathcal{L}$. The proofs for the other estimates are similar.

Next observe that the dimensions of the subspaces underlying H_5, H_4, H_3, H_2, H_1, M are respectively 1, b, $r + b - 1$, rb, $rb + s - 1$ and $rb + rs - r$.

The sums of squares and degrees of freedom entries for the ANOVA table can now be written down.

Source	SS	DF
H_5 vs H_4	$rs\sum_{k=1}^{b}(y_{\bullet\bullet k} - y_{\bullet\bullet\bullet})^2$	$b-1$
H_4 vs H_3	$sb\sum_{i=1}^{r}(y_{i\bullet\bullet} - y_{\bullet\bullet\bullet})^2$	$r-1$
H_3 vs H_2	$s\sum_{i=1}^{r}\sum_{k=1}^{b}(y_{i\bullet k} - y_{i\bullet\bullet} - y_{\bullet\bullet k} + y_{\bullet\bullet\bullet})^2$	$(r-1)(b-1)$
H_2 vs H_1	$rb\sum_{j=1}^{s}(y_{\bullet j\bullet} - y_{\bullet\bullet\bullet})^2$	$s-1$
H_1 vs M	$b\sum_{i=1}^{r}\sum_{j=1}^{s}(y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + y_{\bullet\bullet\bullet})^2$	$(r-1)(s-1)$
Residual	$\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{b}(y_{ijk} - y_{ij\bullet} - y_{i\bullet k} + y_{i\bullet\bullet})^2$	$r(s-1)(b-1)$
Residual	$\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{b}(y_{ijk} - y_{\bullet\bullet\bullet})^2$	$rsb-1$

Lecture 20, 2018

To derive the expected mean squares, we use the fact that the residual mean square has expectation σ^2 and, in general, the expected mean square corresponding to the test of $H_{\nu+1}$ vs H_ν is

$$\sigma^2 + \frac{\boldsymbol{\eta}^T(P_\nu - P_{\nu+1})\boldsymbol{\eta}}{p_\nu - p_{\nu+1}}.$$

Since the corresponding sum of squares entry is given by

$$\boldsymbol{y}^T(P_\nu - P_{\nu+1})\boldsymbol{y}$$

the expected mean square can be evaluated by substituting

$$\eta_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \beta_k + \delta_{ik}$$

for y_{ijk} in the previously derived sum of squares entries.

For example, to obtain the expected mean square for H_1 vs M, observe

$$\begin{aligned}
\boldsymbol{\eta}^T(P_1 - P)\boldsymbol{\eta} &= b\sum_i\sum_j(\eta_{ij\bullet} - \eta_{i\bullet\bullet} - \eta_{\bullet j\bullet} + \eta_{\bullet\bullet\bullet})^2 \\
&= b\sum_i\sum_j(\alpha_i + \tau_j + \gamma_{ij} + \beta_\bullet + \delta_{i\bullet} - \alpha_i - \tau_\bullet - \gamma_{i\bullet} - \beta_\bullet - \delta_{i\bullet} \\
&\quad -\alpha_\bullet - \tau_j - \gamma_{\bullet j} - \beta_\bullet - \delta_{\bullet\bullet} + \alpha_\bullet + \tau_\bullet + \gamma_{\bullet\bullet} + \beta\bullet + \delta_{\bullet\bullet})^2 \\
&= b\sum_i\sum_j(\gamma_{ij} - \gamma_{i\bullet} - \gamma_{j\bullet} + \gamma_{\bullet\bullet})^2 \\
&= b\sum_i\sum_j\gamma_{ij}^2
\end{aligned}$$

since

$$\gamma_{i\bullet} = \gamma_{j\bullet} = \gamma_{\bullet\bullet} = 0.$$

Note that in this expression all of the other effects, and in particular δ_{ik}, cancel algebraically without need to apply the constraints. The other expected mean square entries can be calculated similarly and the results are summarised as follows.

Source	DF	MS
H_5 vs H_4	$b-1$	$\sigma^2 + \frac{rs}{b-1}\sum_{k=1}^{b}\beta_k^2$
H_4 vs H_3	$r-1$	$\sigma^2 + \frac{sb}{r-1}\sum_{i=1}^{r}\alpha_i^2$
H_3 vs H_2	$(r-1)(b-1)$	$\sigma^2 + \frac{s}{(r-1)(b-1)}\sum_{i=1}^{r}\sum_{k=1}^{b}\delta_{ik}^2$
H_2 vs H_1	$s-1$	$\sigma^2 + \frac{rb}{s-1}\sum_{j=1}^{s}\tau_j^2$
H_1 vs M	$(r-1)(s-1)$	$\sigma^2 + \frac{b}{(r-1)(s-1)}\sum_{i=1}^{r}\sum_{j=1}^{s}\gamma_{ij}^2$
Residual	$(s-1)(b-1)$	σ^2

Finally, consider the mixed model for the split plot experiment

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \beta_k + D_{ik} + \mathcal{E}_{ijk}$$

where

$$\sum_i \alpha_i = \sum_j \tau_j = \sum_k \beta_k = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

and $\mathcal{E}_{ijk} \sim N(0, \sigma^2)$ and $D_{ik} \sim N(0, \sigma_D^2)$ independently.

Examination of the sum of squares entries shows that D_{ik} cancels algebraically from all expressions except H_3 vs H_2, H_4 vs H_3 and H_5 vs H_4. To derive the expected means squares in those cases, consider the averaged data

$$Y_{i\bullet k} = \mu + \alpha_i + \beta_k + D_{ik} + \mathcal{E}_{i\bullet k} = \mu + \alpha_i + \beta_k + F_{ik}$$

where $F_{ik} \sim N(0, \sigma^2/s + \sigma_D^2)$ independently. Next, observe that the corresponding sums of squares are proportional to those for the additive model in $r \times b$ two-way layout, $\{y_{i\bullet k}\}$,

Source	SS	DF
H_5 vs H_4	$rs\sum_{k=1}^{b}(y_{\bullet\bullet k} - y_{\bullet\bullet\bullet})^2$	$b-1$
H_4 vs H_3	$sb\sum_{i=1}^{r}(y_{i\bullet\bullet} - y_{\bullet\bullet\bullet})^2$	$r-1$
H_3 vs H_2	$s\sum_{i=1}^{r}\sum_{k=1}^{b}(y_{i\bullet k} - y_{i\bullet\bullet} - y_{\bullet\bullet k} + y_{\bullet\bullet\bullet})^2$	$(r-1)(b-1)$

and, hence expected mean squared entries are respectively

$$\sigma^2 + s\sigma_D^2,\ \sigma^2 + s\sigma_D^2 + \frac{sb}{r-1}\sum_{i=1}^{r}\alpha_i^2 \text{ and } \sigma^2 + s\sigma_D^2 + \frac{rs}{b-1}\sum_{k=1}^{b}\beta_i^2$$

Hence the full table of expected mean squares for the split plot experiment is

Source	DF	MS
H_5 vs H_4	$b-1$	$\sigma^2 + s\sigma_D^2 + \frac{rs}{b-1}\sum_{k=1}^{b}\beta_k^2$
H_4 vs H_3	$r-1$	$\sigma^2 + s\sigma_D^2 + \frac{sb}{r-1}\sum_{i=1}^{r}\alpha_i^2$
H_3 vs H_2	$(r-1)(b-1)$	$\sigma^2 + s\sigma_D^2$
H_2 vs H_1	$s-1$	$\sigma^2 + \frac{rb}{s-1}\sum_{j=1}^{s}\tau_j^2$
H_1 vs M	$(r-1)(s-1)$	$\sigma^2 + \frac{b}{(r-1)(s-1)}\sum_{i=1}^{r}\sum_{j=1}^{s}\gamma_{ij}^2$
Residual	$(s-1)(b-1)$	σ^2

In this light the split-plot analysis of variance can be seen to be justified by the principle that the expected mean squared entries on the numerator and denominator of an F-statistic should be the same when the null hypothesis is true.

6.2 Algebraic calculation of least squares estimates for balanced factorial experiments

6.2.1 Index notation

In the preceding discussion, we introduced expressions for the least squares estimates for a sequence of factorial models. We proved that the given expressions were indeed the least squares estimates but this does not show how the expressions were obtained. In what follows, we show how such expressions can be obtained in the general case.

Consider a factorial experiment defined by factors $B_1, B_2, \ldots, B_p$ with $r_1, r_2, \ldots, r_n$ levels respectively. Let

$$\boldsymbol{y} = [y(i_1 i_2 \ldots i_p)]$$

be an $r_1 \times r_2 \times \ldots r_p$ data array such that $y(i_1 i_2 \ldots i_p)$ is the observed response when $B_1 = i_1, B_2 = i_2, \ldots, B_p = i_p$. Let

$$\boldsymbol{\eta} = [\eta(i_1 i_2 \ldots i_p)]$$

be the corresponding array of expected values.

Let $\boldsymbol{i} = (i_1, i_2, \ldots, i_p)$ and for any subset $S = \{s_1, s_2, \ldots, s_q\} \subset \{1, 2, \ldots, p\}$ let

$$\boldsymbol{i}_S = (i_{s_1}, i_{s_2}, \ldots, i_{s_q}). \tag{5}$$

We consider models of the form

$$\eta(i_1 i_2 \ldots i_p) = \sum_{S \in \mathcal{G}} \theta_S(\boldsymbol{i}_S)$$

where $\mathcal{G}$ is a collection of subsets of $\{1, 2, \ldots, p\}$.

Although this notation appears complicated, it is convenient for representing ANOVA type models in a general setting.

Example Consider a three-way experiment and the model

$$\eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta_k + \xi_{ik}.$$

In the general notation we would write this as

$$\eta(i_1, i_2, i_3) = \theta_\phi + \theta_{\{1\}}(i_1) + \theta_{\{2\}}(i_2) + \theta_{\{1,2\}}(i_1, i_2) + \theta_{\{3\}}(i_3) + \theta_{\{1,3\}}(i_1, i_3)$$

with

$$\begin{aligned}
i_1 &= i \\
i_2 &= j \\
i_3 &= k \\
\theta_\phi &= \mu \\
\theta_{\{1\}}(i_1) &= \alpha_i \\
\theta_{\{2\}}(i_2) &= \beta_j \\
\theta_{\{1,2\}}(i_1, i_2) &= \gamma_{ij} \\
\theta_{\{3\}}(i_3) &= \delta_k \\
\theta_{\{1,3\}}(i_1, i_3) &= \xi_{ik}
\end{aligned}$$

6.2.2 Averaging operators

Now, for any subset $S \subseteq \{1, 2, \ldots, p\}$ let A_S be the averaging operator that replaces the array $[y(i_1, i_2, \ldots, i_p]$ by the array with values averaged over all indices not contained in S.

For example, if $p = 4$ and $S = \{1, 2\}$, we have

$$A_{\{1,2\}}[y(i_1, i_2, i_3, i_4)] = [y(i_1, i_2, \bullet, \bullet)]$$

where

$$y(i_1, i_2, \bullet, \bullet) = \frac{1}{r_3 r_4} \sum_{i_3=1}^{r_3} \sum_{i_4=1}^{r_4} y(i_1, i_2, i_3, i_4).$$

Note that we still think of $A_{\{1,2\}}[y(i_1, i_2, i_3, i_4)] = [y(i_1, i_2, \bullet, \bullet)]$ as a $r_1 \times r_2 \times r_3 \times r4$ array whose elements are the averages as defined above and hence do not depend on i_3 and i_4.

Given this definition it is relatively simple to check the following two properties of these operators.

$$A_S A_R = A_{S \cap R} = A_R A_S. \tag{6}$$

For example, if $R = \{1, 2, 3\}$ and $S = \{2, 3, 4\}$ then

$$A_{\{1,2,3\}} A_{\{2,3,4\}}[y(i_1, i_2, i_3, i_4)] = A_{\{1,2,3\}}[y(\bullet, i_2, i_3, i_4)] = [y(\bullet, i_2, i_3, \bullet)] = A_{\{2,3\}}[y(i_1, i_2, i_3, i_4)].$$

As a simple corollary, we have

$$A_R^2 = A_R.$$

6.2.3 Construction of the least squares estimates

Consider the model (5). The least squares estimates may be constructed as follows.

1. Take $\mathcal{G}'$ to be the collection of subsets of $\mathcal{G}$ that are maximal with respect to subset inclusion. That is, start with the subsets in $\mathcal{G}$ and remove any $R \in \mathcal{G}$ if there is also $S \in \mathcal{G}$ with $R \subset S$.
2. The least squares estimate is then given by

$$\hat{\boldsymbol{\eta}} = \boldsymbol{y} - \prod_{S \in \mathcal{G}'} (I - A_S)\boldsymbol{y}. \tag{7}$$

Example (continued)

1. The model may be represented in set notation as

$$\mathcal{G} = \{\phi, \{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 3\}\}.$$

   Removing the redundant subsets gives

$$\mathcal{G}' = \{\{1, 2\}, \{1, 3\}\}.$$

2. The least squares estimate is then given by

$$\begin{aligned}\boldsymbol{\eta} &= \boldsymbol{y} - (I - A_{\{1,2\}})(I - A_{\{1,3\}})\boldsymbol{y} \\ &= \boldsymbol{y} - (I - A_{\{1,2\}} - A_{\{1,3\}} + A_{\{1,2\}}A_{\{1,3\}})\boldsymbol{y} \\ &= \boldsymbol{y} - (I - A_{\{1,2\}} - A_{\{1,3\}} + A_{\{1\}})\boldsymbol{y}\end{aligned}$$

so that

$$\begin{aligned}[\hat{\eta}(i_1,i_2,i_3)] &= [y(i_1,i_2,i_3)] - (I - A_{\{1,2\}} - A_{\{1,3\}} + A_{\{1\}})[y(i_1,i_2,i_3)] \\ &= y(i_1,i_2,i_3) - y(i_1,i_2,i_3) + y(i_1,i_2,\bullet) + y(i_1,\bullet,i_3) - y(i_1,\bullet,\bullet) \\ &= y(i_1,i_2,\bullet) + y(i_1,\bullet,i_3) - y(i_1,\bullet,\bullet).\end{aligned}$$

In the more familiar subscript notation, this may be written as

$$\hat{\eta}_{ijk} = y_{ij\bullet} + y_{i\bullet k} - y_{i\bullet\bullet}.$$

Proof that (7) is the least squares estimate

Observe first that since, (5) is a linear model we need only check that

$$\hat{\boldsymbol{\eta}} \in \mathcal{M} \text{ and } \boldsymbol{y} - \hat{\boldsymbol{\eta}} \perp \mathcal{M}$$

where $\mathcal{M}$ is the linear space of all arrays that satisfy (5).

To see that $\hat{\boldsymbol{\eta}} \in \mathcal{M}$, observe that in (7) $\hat{\boldsymbol{\eta}}$ is of the form

$$\hat{\boldsymbol{\eta}} = \sum A_R \boldsymbol{y}$$

where each R is a subset of $\{1, 2, \ldots, p\}$ such that $R \subseteq S$ for some $S \in \mathcal{G}'$. Each such $A_R\boldsymbol{y}$ is an element of $\mathcal{M}$ and, since $\mathcal{M}$ is a linear space it follows that

$$\hat{\boldsymbol{\eta}} = \sum A_R \boldsymbol{y} \in \mathcal{M}.$$

To show that $\boldsymbol{y} - \hat{\boldsymbol{\eta}} \perp \mathcal{M}$ we need to show that

$$\sum_{\boldsymbol{i}} a(\boldsymbol{i})(y(\boldsymbol{i} - \hat{\eta}(\boldsymbol{i})) = 0$$

for any array $[a(i_1, i_2, \ldots, i_p)] \in \mathcal{M}$. But any array $[a(i_1, i_2, \ldots, i_p)] \in \mathcal{M}$ is expressible in the form

$$a(i_1, i_2, \ldots, i_p) = \sum_{S \in \mathcal{G}'} \theta_S(\boldsymbol{i}_S)$$

so it is sufficient to show

$$\sum_{\boldsymbol{i}} \theta_S(\boldsymbol{i}_S)(y(\boldsymbol{i} - \hat{\eta}(\boldsymbol{i})) = 0$$

for any function $\theta_S(\boldsymbol{i}_S)$ with $S \in \mathcal{G}'$. Now for any array, $\boldsymbol{x} = [x(i_1, i_2, \ldots, i_p)]$ it is easy to see that

$$\sum_{\boldsymbol{i}} \theta_S(\boldsymbol{i}_S) x(\boldsymbol{i}) = \sum_{\boldsymbol{i}} \theta_S(\boldsymbol{i}_S)[A_S x](\boldsymbol{i})$$

so to complete the proof, we show that

$$A_S(\boldsymbol{y} - \hat{\boldsymbol{\eta}}) = \mathbf{0}$$

each $S \in \mathcal{G}'$. But from (7),

$$(\boldsymbol{y} - \hat{\boldsymbol{\eta}}) = \prod_{R \in \mathcal{G}'} (I - A_R)\boldsymbol{y} = (I - A_S) \prod_{R \in \mathcal{G}', R \neq S} (I - A_R)\boldsymbol{y}$$

since the averaging operators commute (6). Hence,

$$A_S(\boldsymbol{y} - \hat{\boldsymbol{\eta}}) = A_S(I - A_S) \prod_{R \in \mathcal{G}', R \neq S} (I - A_R)\boldsymbol{y}$$

and since

$$A_S(I - A_S) = A_S - A_S^2 = A_S - A_S = 0$$

by (6), the result is proved. □

Remark We have considered averaging operators A in the construction of the least squares estimates for balanced factorial models. It can also be seen that the averaging operators are in fact orthogonal projections and the balance in the designs implies that the corresponding sub-models are orthogonal in the sense of section 5.5.1. We will not discuss the details.

Lecture 21, 2018

7 Logistic regression

Consider data of the form

$$(n_1, y_1, \boldsymbol{x}_1),\ (n_2, y_2, \boldsymbol{x}_2),\ \ldots, (n_m, y_m, \boldsymbol{x}_m)$$

and the model

$$Y_i \sim B(n_i, \pi_i)$$

independently for $i = 1, 2, \ldots, m$.

Example In a series of experiments, beetles were exposed to gaseous carbon disulphide at various concentrations for a period of 5 hours and the numbers of deaths recorded. The following data were recorded.

Concentration (x)	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Number Exposed (n)	59	60	62	62	63	59	62	60
Deaths (y)	6	13	18	28	52	53	61	60

The purpose of the analysis is to relate the probability of death to the concentration of carbon disulphide. □

In general, the purpose of the analysis is to relate the success probability π to the predictor variable $\boldsymbol{x}_i$. In particular, we seek a model of the from

$$\pi_i = \pi(\boldsymbol{x}_i).$$

In analogy to linear regression, a simple suggestion would be to consider the linear model

$$\pi_i = \boldsymbol{\beta}^T \boldsymbol{x}_i.$$

However, such a model can be seen to be unsuitable because probabilities are constrained to satisfy $0 \leq \pi_i \leq 1$ whereas linear functions are not.

To overcome this difficulty the logistic regression model is defined by

$$\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i.$$

where

$$\eta_i = \log \frac{\pi_i}{1 - \pi_i}.$$

It is easy to check that for $0 < \pi < 1$ the *logit*

$$\eta = \log \frac{\pi}{1 - \pi}$$

satisfies $-\infty < \eta < \infty$ and moreover the transformation is invertible. In particular,

$$\pi = \frac{\exp(\eta)}{1+\exp(\eta)}$$

so the model may be written

$$\pi_i = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}.$$

For the logistic regression model, the parameters and approximate standard errors are obtained using the method of maximum likelihood.

Recall from Mathematical Statistics III that, for a scalar parameter, θ, the maximum likelihood estimate $\hat{\theta}$ is obtained by maximizing the log-likelihood function $\ell(\theta; \boldsymbol{y})$ and, for large n,

$$\mathrm{var}(\hat{\theta}) \approx \frac{1}{\mathcal{I}(\theta)}$$

where

$$\mathcal{I}(\theta) = E\left(-\frac{\partial^2 \ell}{\partial \theta^2}\right).$$

In practice, an approximate variance is obtained from $1/\mathcal{I}(\hat{\theta})$.

The method can also be generalised to the case of a vector parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. In particular, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is the vector that minimizes the log-likelihood function, $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ and for large n,

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) \approx [\mathcal{I}(\boldsymbol{\theta})]^{-1}$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the $p \times p$ matrix

$$E\left(-\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right).$$

Consider now the logistic regression model and the binomial likelihood,

$$p(\boldsymbol{y}; \boldsymbol{\beta}) = \prod_{i=1}^{m} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}.$$

The log-likelihood is thus

$$\begin{aligned} \ell(\boldsymbol{\beta}; \boldsymbol{y}) &= \sum_{i=1}^{m} y_i \log \pi_i + \sum_{i=1}^{m} (n_i - y_i) \log(1-\pi_i) \\ &\quad + \log\left\{\prod_{i=1}^{m} \binom{n_i}{y_i}\right\} \\ &= \sum_{i=1}^{m} \left\{ y_i \log \frac{\pi_i}{1-\pi_i} + n_i \log(1-\pi_i) \right\} \\ &\quad + \log\left\{\prod_{i=1}^{m} \binom{n_i}{y_i}\right\} \\ &= \sum_{i=1}^{m} \{ y_i \eta_i - n_i \log(1+e^{\eta_i}) \} + \text{Const.} \end{aligned}$$

To find the maximum likelihood estimates, we need to solve the equations

$$\frac{\partial \ell}{\partial \beta_j} = 0 \text{ for } j = 1, 2, \ldots, p.$$

Now,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

and noting that

$$\frac{\partial}{\partial \eta_i} \log(1 + e^{\eta_i}) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \pi_i$$

we find

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{m} (y_i - n_i \pi_i) \frac{\partial \eta_i}{\partial \beta_j}.$$

Using vector notation, and noting that

$$\boldsymbol{\eta} = X\boldsymbol{\beta} \Rightarrow \left[\frac{\partial \eta_i}{\partial \beta_j}\right] = X$$

we find the score vector, expressed as column vector, is

$$S(\boldsymbol{\beta}) = \left[\frac{\partial \ell}{\partial \beta_j}\right] = X^T(\boldsymbol{y} - \boldsymbol{\mu})$$

where $\boldsymbol{\mu}$ is the vector of expected frequencies

$$\mu_i = n_i \pi_i \text{ with } \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{8}$$

To obtain the Fisher information matrix, we use the fact that

$$\text{Var}(S(\boldsymbol{\beta})) = \mathcal{I}(\boldsymbol{\beta})$$

where

$$S(\boldsymbol{\beta}) = X^T(\boldsymbol{Y} - \boldsymbol{\mu})$$

is the score, expressed as a column vector.

Since X is a fixed matrix, we obtain

$$\begin{aligned} \mathcal{I}(\boldsymbol{\beta}) &= \text{Var}(X^T(\boldsymbol{Y} - \boldsymbol{\mu})) \\ &= X^T \,\text{Var}(\boldsymbol{Y}) X \\ &= X^T D X \end{aligned}$$

where

$$D = \text{diag}(n_1\pi_1(1-\pi_1), n_2\pi_2(1-\pi_2), \ldots, n_m\pi_m(1-\pi_m)) = \text{Var}(\boldsymbol{Y}) \tag{9}$$

since $Y_1, Y_2, \ldots, Y_m$ are assumed independent $Y_i \sim B(n_i, \pi_i)$.

The maximum likelihood estimates are found by solving the system of equations

$$\frac{\partial \ell}{\partial \beta_j} = 0$$

or, equivalently

$$S(\boldsymbol{\beta}) = \mathbf{0}.$$

This system cannot be solved analytically, and an iterative scheme such as the Newton-Raphson algorithm or the Fisher scoring algorithm is usually used.

Recall that the Newton-Raphson algorithm consists of choosing an initial estimate $\boldsymbol{\beta}^{(0)}$ and then iterating

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[S'(\boldsymbol{\beta}^{(t)})\right]^{-1} S(\boldsymbol{\beta}^{(t)})$$

where

$$S'(\boldsymbol{\beta})$$

is the $p \times p$ matrix

$$\left[\frac{\partial S_j}{\partial \beta_k}\right] = \left[\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right].$$

The Fisher scoring algorithm is obtained by replacing the matrix $S'(\boldsymbol{\beta})$ by its expected value, namely $-\mathcal{I}(\boldsymbol{\beta})$. Thus the iterative step of the Fisher scoring algorithm in the general case is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [\mathcal{I}(\boldsymbol{\beta}^{(t)})]^{-1} S(\boldsymbol{\beta}^{(t)}).$$

For the logistic regression model, this can be simplified to

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (X^T D^{(t)} X)^{-1} X^T (\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$$

where $D^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ are as defined in (8) and (9) evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$. The initial approximation, $\boldsymbol{\beta}^{(0)}$ is usually taken to be

$$\boldsymbol{\beta}_0 = (X^T X)^{-1} X^T \boldsymbol{v}$$

where

$$v_i = \log \frac{y_i + 0.5}{n_i - y_i + 0.5}.$$

7.1 Inference for regression coefficients

When the logistic regression model is fit, large sample standard errors for the parameters are obtained by taking the square-roots of the diagonal elements of the inverse information matrix. For large m or large n_i, the approximate distribution of $\hat{\beta}_j$ is

$$N\left(\beta_j, \sqrt{i^{jj}(\hat{\boldsymbol{\beta}})}\right)$$

where i^{jj} is the jth diagonal element of $\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}$.

This allows for the construction of approximate confidence intervals and z-tests for individual parameters in the usual way.

Example The linear logistic model can be fit to the beetle data in `R` as follows.

```
beetle<-read.csv("beetle.csv",header=TRUE)
attach(beetle)
beetle

##   Concentration Exposed Dead
## 1        1.6907      59    6
## 2        1.7242      60   13
## 3        1.7552      62   18
## 4        1.7842      62   28
## 5        1.8113      63   52
## 6        1.8369      59   53
## 7        1.8610      62   61
## 8        1.8830      60   60

y<-cbind(Dead,Exposed-Dead)
summary(glm(y~Concentration,family=binomial))

##
## Call:
## glm(formula = y ~ Concentration, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.206  -0.226   1.012   1.359   1.678
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -60.556      5.151  -11.76   <2e-16 ***
## Concentration   34.146      2.893   11.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 287.499  on 7  degrees of freedom
## Residual deviance:  14.514  on 6  degrees of freedom
## AIC: 44.804
##
## Number of Fisher Scoring iterations: 4
```

The table of regression coefficients has the same layout and interpretation as in linear regression. In particular, the z-statistics and approximate P-values are provided for each coefficient and approximate $100(1-\alpha)\%$ confidence intervals can be obtained from

$$\hat{\beta}_j \pm z(\alpha/2)\sqrt{i^{jj}}.$$

So, for example, a 95% confidence interval for β_1 is

$$34.146 \pm 1.96 \times 2.893.$$

$\square$

7.1.1 Hypotheses concerning several parameters

Consider the logistic regression model

$$M:\ \boldsymbol{\eta} = X\boldsymbol{\beta} \text{ where } \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$$

and the hypothesis

$$H_0:\ \beta_p = \beta_{p-1} = \ldots = \beta_{p_0+1} = 0.$$

The log likelihood ratio test statistic is defined by

$$G^2 = 2(\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0))$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are the MLEs under M and H_0 respectively. If H_0 is true then the asymptotic distribution of G^2 is $\chi^2_{p-p_0}$ and we reject H_0 for

$$G^2 \geq \chi^2_{p-p_0}(\alpha).$$

The deviance statistic for any model M is defined by

$$D(M) = 2(\hat{\ell} - \ell(\hat{\boldsymbol{\beta}}))$$

where $\hat{\ell}$ is the value of the likelihood maximized without restriction and is produced in `R` as the `Residual Deviance`.

For grouped data such that $n_i\pi_i(1-\pi_i)$ are all large, the asymptotic distribution of the residual deviance is χ^2_{m-p}. In this case, the residual deviance can be used to test the overall fit of the model.

Example In the beetle data, the residual deviance is 14.514 with 6 degrees of freedom. The P-value is thus

```
1-pchisq(14.514,df=6)

## [1] 0.02439289
```

which suggests that linear logistic model does not adequately describe these data.

If we fit the quadratic logistic model, we find

```
summary(glm(y~Concentration+I(Concentration^2),family=binomial))

## 
## Call:
## glm(formula = y ~ Concentration + I(Concentration^2), family = binomial)
## 
## Deviance Residuals: 
##        1         2         3         4         5         6         7         8  
## -0.53225   0.93101   0.00522  -0.97086   1.09464  -0.76726   0.13629   0.70724  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)   
## (Intercept)          486.06     179.80   2.703  0.00686 **
## Concentration       -582.53     203.63  -2.861  0.00423 **
## I(Concentration^2)   173.82      57.63   3.016  0.00256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 287.4994  on 7  degrees of freedom
## Residual deviance:   4.3984  on 5  degrees of freedom
## AIC: 36.688
## 
## Number of Fisher Scoring iterations: 4

1-pchisq(4.3984,df=5)

## [1] 0.4935911
```

Note that the quadratic term is significant and also the residual deviance is now non-significant, so we conclude that the quadratic regression model is adequate. □

It should be noted that this interpretation of the residual deviance as an absolute test of fit applies only to grouped data with $n_i\pi_i(1-\pi_i)$ large. In other situations, the residual deviance can still be calculated but no suitable approximating χ^2 distribution is available so significance cannot be calculated.

The residual deviance can also be used to calculate the likelihood ratio test statistic,

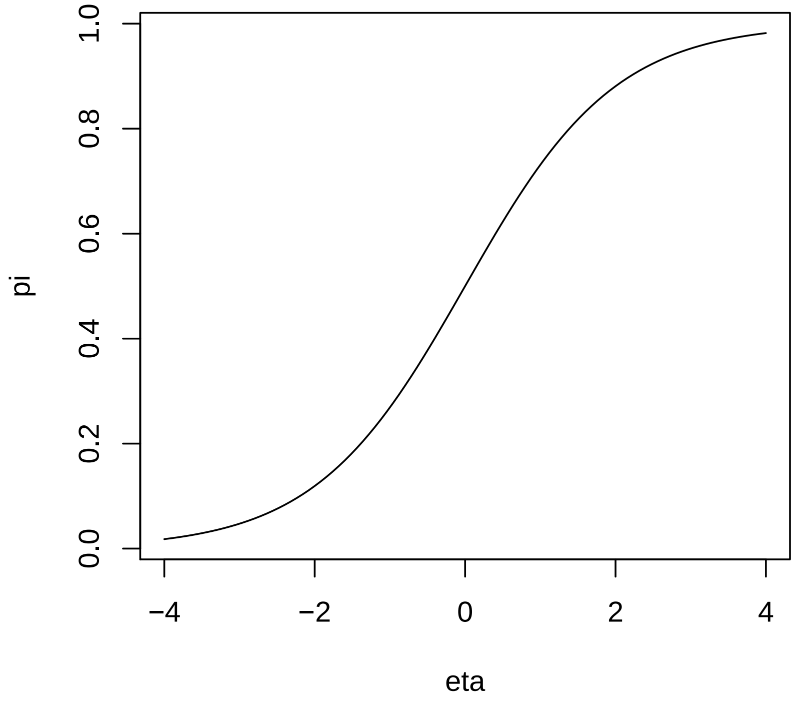
$$G^2 = D(H_0) - D(M).$$

However, it must be kept in mind that a valid likelihood ratio test can be performed only when H_0 is a sub-model of M.

7.2 Interpretation of regression coefficients

A useful aspect of the logistic regression model is that the same model formula conventions can be applied as for linear regressions and in particular factorial, interaction and polynomial terms can be introduced in the same way.

Consider now a term $\beta_j x_j$ in the context of a logistic regression. Its interpretation is that, provided all other variables are held fixed, a one unit change in x_j produces a change of β_j in the log odds η. However, the change to the probability π depends on the value of π.



An important case is when π is small so that

$$1 - \pi \approx 1 \Rightarrow \log \frac{\pi}{1-\pi} \approx \log \pi.$$

In this situation, increasing x_j by one unit corresponds approximately to a change of β_j to $\log \pi$ or in other words, π is (approximately) multiplied by e^{β_j}. So, for example, if π is small and $\beta_j = \ln 2$ then a one unit increase in x_j corresponds to a doubling of the success probability.

Lecture 23, 2018

7.3 Common odds ratios for several 2×2 tables.

Sometimes it happens that an experiment or observational study is replicated across many *strata*.

Example In an observational study conducted at the University of California, Berkeley, applications for admission to graduate study were classified according to the sex of the applicant and whether or not the applicant was admitted. The following data were recorded for applicants to 6 departments.

	Male		Female	
Department	Admitted	Rejected	Admitted	Rejected
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

The purpose of the analysis is to identify any systematic effect of sex upon the probability of admission. In this context, `Sex` is the "treatment" variable, `Admission` is the response variable, and the departments can be thought of as defining *strata*. Within each stratum, the data can be represented as 2×2 contingency table.

Department A

	Admitted	Rejected	Total
Female	89	19	108
Male	512	313	825
Total	601	332	943

Department B

	Admitted	Rejected	Total
Female	17	8	25
Male	353	207	560
Total	370	215	585

Department C

	Admitted	Rejected	Total
Female	202	391	593
Male	120	205	325
Total	322	596	918

Department D

	Admitted	Rejected	Total
Female	131	244	375
Male	138	279	417
Total	269	523	792

Department E

	Admitted	Rejected	Total
Female	94	299	393
Male	53	138	191
Total	147	437	584

Department F

	Admitted	Rejected	Total
Female	24	317	341
Male	22	351	373
Total	46	668	714

Let π_{ij} denote the probability of admission for sex j in department i.

Department i

	Admitted	Rejected	Total
Female	π_{i1}	$1-\pi_{i1}$	1
Male	π_{i2}	$1-\pi_{i2}$	1

The log odds-ratio,

$$\beta_i = \log \frac{\pi_{i2}/(1-\pi_{i2})}{\pi_{i1}/(1-\pi_{i1})}$$

can be used as a measure of the association between `Sex` and `Admission`. In particular

$$\begin{aligned} \beta > 0 &\Rightarrow \text{ Admission more likely for males;} \\ \beta < 0 &\Rightarrow \text{ Admission more likely for females;} \\ \beta = 0 &\Rightarrow \text{ Admission equally likely for males and females.} \end{aligned}$$

□

Consider now the logistic regression model

$$\operatorname{logit} \pi_{ij} = \mu + \alpha_i + \beta_j. \tag{10}$$

It follows that

$$\log \frac{\pi_{i2}/(1-\pi_{i2})}{\pi_{i1}/(1-\pi_{i1})} = \operatorname{logit} \pi_{i2} - \operatorname{logit} \pi_{i1} = \beta_2 - \beta_1.$$

That is, the log odds ratio is the same for all strata (i). For this reason (10) is sometimes called the "common odds ratio" model. It postulates that the association between the treatment and response, as measured by the odds ratio, is constant across strata, although the success probabilities need not be constant.

The common odds ratio can be analysed in `R` as follows.

Example (Continued)

```
berkeley<-read.csv("berkeley.csv",header=TRUE)
```

```
berkeley

##    Accepted Rejected    sex dept
## 7        89       19 Female    A
## 1       512      313   Male    A
## 8        17        8 Female    B
## 2       353      207   Male    B
## 9       202      391 Female    C
## 3       120      205   Male    C
## 10      131      244 Female    D
## 4       138      279   Male    D
## 11       94      299 Female    E
## 5        53      138   Male    E
## 12       24      317 Female    F
## 6        22      351   Male    F

attach(berkeley)
y<-cbind(Accepted,Rejected)
# Fit the saturated model.
summary(glm(y~dept/sex,family=binomial))

##
## Call:
## glm(formula = y ~ dept/sex, family = binomial)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.54420    0.25272   6.110 9.94e-10 ***
## deptB          -0.79043    0.49769  -1.588    0.112
## deptC          -2.20464    0.26716  -8.252  < 2e-16 ***
## deptD          -2.16617    0.27495  -7.878 3.32e-15 ***
## deptE          -2.70135    0.27902  -9.682  < 2e-16 ***
## deptF          -4.12505    0.32968 -12.512  < 2e-16 ***
## deptA:sexMale -1.05208    0.26271  -4.005 6.21e-05 ***
## deptB:sexMale -0.22002    0.43759  -0.503    0.615
## deptC:sexMale  0.12492    0.14394   0.868    0.385
## deptD:sexMale -0.08199    0.15021  -0.546    0.585
## deptE:sexMale  0.20019    0.20024   1.000    0.317
## deptF:sexMale -0.18890    0.30516  -0.619    0.536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.7706e+02  on 11  degrees of freedom
## Residual deviance: 3.1863e-13  on  0  degrees of freedom
## AIC: 92.94
##
## Number of Fisher Scoring iterations: 3

# Note that -1.05208=log(512*19/(313*89)) etc
# Now fit the common odds ratio model
```

```
summary(glm(y~dept+sex,family=binomial))

##
## Call:
## glm(formula = y ~ dept + sex, family = binomial)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8        9       10
##  3.7189  -1.2487   0.2706  -0.0560  -0.9243   1.2533  -0.0858   0.0826  -0.8509   1.2205
##      11       12
##  0.2052  -0.2076
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68192    0.09911   6.880 5.97e-12 ***
## deptB       -0.04340    0.10984  -0.395    0.693
## deptC       -1.26260    0.10663 -11.841  < 2e-16 ***
## deptD       -1.29461    0.10582 -12.234  < 2e-16 ***
## deptE       -1.73931    0.12611 -13.792  < 2e-16 ***
## deptF       -3.30648    0.16998 -19.452  < 2e-16 ***
## sexMale     -0.09987    0.08085  -1.235    0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.056  on 11  degrees of freedom
## Residual deviance:  20.204  on  5  degrees of freedom
## AIC: 103.14
##
## Number of Fisher Scoring iterations: 4

# The overall log odds ratio estimate is -0.09987 with SE 0.08085.
# There is some evidence against the common odds ratio model as the
# Residual Deviance is significant.
1-pchisq(20.204,df=5)

## [1] 0.001144215

# Although there is no interest in department effects, they are
# highly significant and must not be removed
summary(glm(y~sex,family=binomial))

##
## Call:
## glm(formula = y ~ sex, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -16.7915   -4.7613   -0.4365    5.1025   11.2022
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.83049    0.05077 -16.357   <2e-16 ***
## sexMale      0.61035    0.06389   9.553   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 877.06  on 11  degrees of freedom
## Residual deviance: 783.61  on 10  degrees of freedom
## AIC: 856.55
##
## Number of Fisher Scoring iterations: 4
```

```
# Note that in this completely incorrect model, the estimated
# log odds-ratio is
# 0.61035 with se 0.06389
# This is an example of Simpson's paradox.
```

□

To summarise:

- The common odds ratio model is often used to describe the effect of a single binary factor upon a binary response over several strata.
- A test of significance can be used to determine whether the common odds ratio model is suitable.
- When the common odds ratio model is suitable it is desirable to calculate a single estimate that combines the information from the separate strata.
  - The maximum likelihood estimate of the log odds ratio obtained from the common odds ratio model is the correct way to estimate this parameter.
  - It is not valid to estimate the common odds ratio from the marginal 2×2 table obtained by combining the strata.
  - Omitting the stratum effects from the common odds ratio model is equivalent to analysing the marginal 2×2 table obtained by combining the strata and will not produce a valid estimate if stratum effects are present.

Finally, it must be noted that the preceding analysis is applicable only in situations where the stratum row and column totals are large. When the stratum totals are all small, it is still reasonable to consider the common odds ratio model but the method of maximum likelihood should not be used as it is known to produce biased and possibly inconsistent estimates in such settings. Alternative approaches such as conditional logistic regression or the Mantel-Haenszel test can be used but are beyond the scope of this course.

7.4 Prospective and retrospective studies

One of the key assumptions for linear regression models is that each observation of the response variable, y, can be thought of as an observation from the corresponding conditional distribution of $Y|\boldsymbol{x}$. For example, in the beetle experiment, it is natural to think of the number of beetles that died at a particular concentration as a binomial response. Experiments or observational studies of this type are often called prospective studies:

- In the beetle example, we expose a number of beetles to a certain concentration of carbon disulphide and observe the numbers that die.
- In an epidemiological study, we might identify individuals with given levels of a certain risk factor (e.g. smoking status) and then observe the number that develop a given disease (e.g. emphysema) within a give time.

An alternative design that is frequently used in epidemiological studies is called a retrospective experiment. In this case, individuals are selected according to their response status and then the exposure to the risk factors of interest is determined. For example, a retrospective study to determine the effect of smoking on emphysema could:

- Select a group of emphysema patients (sometimes called cases);
- Select a comparable group of subjects without emphysema (sometimes called controls);
- Determine the relevant smoking history for all of the subjects.

At first sight, it would appear that retrospective studies require a very different analysis to ordinary prospective studies. While this is generally the case, it can be seen that for certain retrospective studies, the logistic regression model can be validly applied as if it were a prospective study.

Consider a population in which the binary response Y satisfies the logistic regression model,

$$P(Y=1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}.$$

Now let S be a sampling indicator defined by

$$S = \begin{cases} 1 & \text{if the subject is sampled} \\ 0 & \text{if the subject is not sampled} \end{cases}$$

The assumptions for retrospective sampling can be stated as

$$P(S=1|Y=1,\boldsymbol{x}) = \rho_1 \text{ and } P(S=1|Y=0,\boldsymbol{x}) = \rho_0.$$

This assumption implies that cases are sampled at rate ρ_1 and controls at rate ρ_0. There is no requirement that $\rho_1 = \rho_0$ but it is critical that the probability of being sampled does not depend on the predictor $\boldsymbol{x}$.

Now consider the success probability

$$P(Y=1|S=1,\boldsymbol{x})$$

for a subject sampled retrospectively. To justify the claim that logistic regression can be used, we must show that $P(Y=1|S=1,\boldsymbol{x})$ satisfies a logistic regression model with the same parameters.

$$
\begin{aligned}
P(Y=1|S=1,\boldsymbol{x}) &= \frac{P(Y=1\cap S=1|\boldsymbol{x})}{P(S=1|\boldsymbol{x})} \\
&= \frac{P(Y=1|\boldsymbol{x})P(S=1|Y=1,\boldsymbol{x})}{P(Y=1|\boldsymbol{x})P(S=1|Y=1,\boldsymbol{x})+P(Y=0|\boldsymbol{x})P(S=1|Y=0,\boldsymbol{x})} \\
&= \frac{\rho_1 P(Y=1|\boldsymbol{x})}{\rho_1 P(Y=1|\boldsymbol{x})+\rho_0 P(Y=0|\boldsymbol{x})} \\
&= \frac{\rho_1 \exp(\beta_0+\boldsymbol{x}^T\boldsymbol{\beta})}{\rho_1 \exp(\beta_0+\boldsymbol{x}^T\boldsymbol{\beta})+\rho_0} \\
&= \frac{(\rho_1/\rho_0)\exp(\beta_0+\boldsymbol{x}^T\boldsymbol{\beta})}{1+(\rho_1/\rho_0)\exp(\beta_0+\boldsymbol{x}^T\boldsymbol{\beta})} \\
&= \frac{\exp(\beta_0+\log(\rho_1/\rho_0)+\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\beta_0+\log(\rho_1/\rho_0)+\boldsymbol{x}^T\boldsymbol{\beta})} \\
&= \frac{\exp(\beta_0^*+\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\beta_0^*+\boldsymbol{x}^T\boldsymbol{\beta})}
\end{aligned}
$$

where $\beta_0^* = \beta_0 + \log(\rho_1/\rho_0)$. Hence, the probability of the positive outcome $Y = 1$ in a retrospective sample satisfies the same logistic regression model as for a prospective sample except that the intercept parameter β_0 is not estimable.

Remarks

- It should be intuitively plausible the intercept term cannot be estimated in a retrospective design. If it could, we would be able to estimate the probability of the positive response $Y = 1$ which is clearly not possible with retrospective sampling.
- Retrospective designs are especially useful for rare diseases, such as cancer, where the rate may be as low as 1/1000 and where the incubation period may be very long. If a prospective design was used in this situation, we would need to recruit and follow roughly 20,000 subjects in order to observe 20 cases of the disease.
- There are other generalised linear models available for binary response data but only logistic regression has the property that prospective and retrospective designs can be analysed interchangeably.
- We have considered only the simplest retrospective design. In general, more complicated designs such as matched case-control studies are used. We will not discuss the details here.

Lecture 24, 2018

7.5 Model fit

As with any statistical analysis, it is important to check the model assumptions when logistic regression is used. In this case, there are two approaches that can be used. Namely, a formal approach in which the model in question is embedded within a more general model and a hypothesis test is conducted.

For example, in the Berkeley Admissions data, we can test the common odds ratio model,

$$\operatorname{logit} \pi_{ij} = \mu + \alpha_i + \beta_j$$

against the saturated model,

$$\operatorname{logit} \pi_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

using a likelihood ratio test.

```
# We have previously seen that the likelihood ratio test statistic for a test of
# the given model against the saturated model can be obtained as the residual deviance.
#
# The likelihood ratio test statistic can also be obtained using the anova function to
# compare the glm object for the model m1 against a more general model m2.
#
# attach(berkeley) has been entered already
y<-cbind(Accepted,Rejected)
m1<-glm(y~dept+sex,family=binomial)
m2<-glm(y~dept*sex,family=binomial)
anova(m1,m2)

## Analysis of Deviance Table
##
## Model 1: y ~ dept + sex
## Model 2: y ~ dept * sex
##   Resid. Df Resid. Dev Df Deviance
## 1         5     20.204
## 2         0      0.000  5   20.204

# The LRT statistic is 20.204 with 5 degrees of freedom. The P-value is
p.value<-1-pchisq(20.204,df=5)
p.value

## [1] 0.001144215

# Hence, as was previously seen, we reject the common odds model in this case.
```

Remarks

- In this example, we have embedded the model of interest (i.e. the common odds ratio model) within the saturated model. This constitutes an *absolute* test of fit because, by definition, the saturated model is correct. However, this approach can be used only when the data are *grouped* and the totals are all large.

- When the data are not grouped, or the marginal totals are small, a limited check can be conducted by testing the model in question against a more general but not necessarily saturated model. For example, in the beetle data, we assessed the fit of the linear logistic model by testing it against a quadratic model. In that case we found a significant quadratic term and concluded that the linear model was not adequate.

  In general, this approach cannot be considered as a *absolute* test of fit because it may happen that the more general model is incorrect.

- When using the anova function in R to compare two models, it is essential that one of the models is a sub-model of the other. For example, the common odds ratio model is sub-model of the saturated model and the linear logistic model is a sub-model of the quadratic model. If the two models are not nested in this sense than a valid likelihood ratio test cannot be performed.

The second, less formal, approach that can sometimes be used is to plot suitably defined residuals from the model fit. For logistic regression, two commonly used types of residuals are:

The Pearson residuals

$$r_i^{(p)} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}};$$

The Deviance residuals

$$r_i^{(d)} = \operatorname{sgn}(y_i - n_i\hat{\pi}_i)\sqrt{2y_i \log\frac{y_i}{n_i\hat{\pi}_i} + 2(n_i - y_i)\log\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}}$$

The Pearson residuals may be interpreted as the contributions to the Pearson χ^2 statistic and the deviance residuals are the corresponding contributions to the residual deviance. Both types of residuals can be used in roughly the same way as residuals from ordinary linear regression. In particular, these residuals may be plotted against the fitted values and also against the individual predictor variables. In considering these residuals, some care must be taken as they are not standardized.

```r
residuals_pearson<-residuals(m1,type="pearson")
residuals_deviance<-residuals(m1,type="deviance")
# Compare the values of the two types of residuals
cbind(residuals_pearson,residuals_deviance)
```

```
##    residuals_pearson residuals_deviance
## 1         3.51866744         3.71892028
## 2        -1.25380765        -1.24867404
## 3         0.26895159         0.27060804
## 4        -0.05602052        -0.05600850
## 5        -0.92077831        -0.92433979
## 6         1.26287232         1.25333751
## 7        -0.08573167        -0.08577122
## 8         0.08260773         0.08256736
## 9        -0.84403319        -0.85093316
## 10        1.24151319         1.22051370
## 11        0.20648081         0.20517793
## 12       -0.20620096        -0.20756402
```

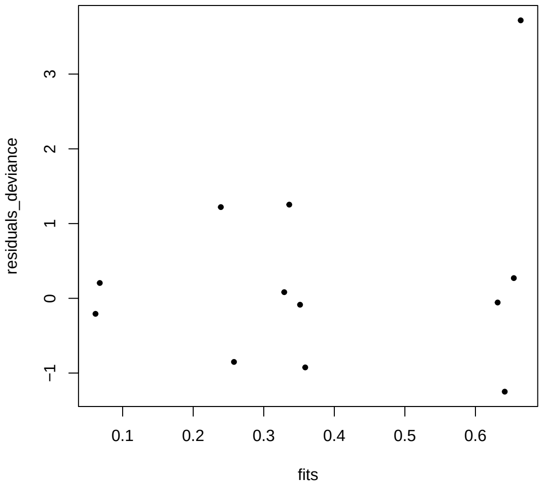
```
# Note also that the sum of the squared deviance residuals is the residual deviance
sum(residuals_deviance^2)
```

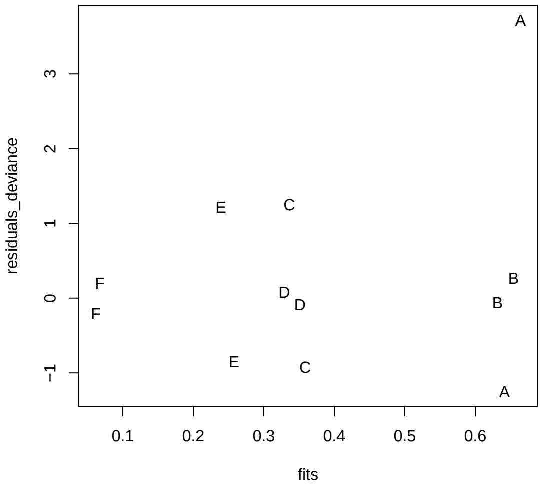
```
## [1] 20.20428
```

```
deviance(m1)
```

```
## [1] 20.20428
```

```
fits<-fitted(m1)
# Plot the Residuals vs Fitted Values
plot(fits,residuals_deviance,pch=20)
# Repeat the Residuals vs Fitted plot, but use the department name as plotting symbol
plot(fits,residuals_deviance,type="n")
text(fits,residuals_deviance,dept)
# This plot shows a large residual associated with Department A that coincides with
# the high admission rate for females previously noted in that department.
```





Remarks

- The preceding residuals are most useful when dealing with grouped data. For ungrouped data, where every y-value is either 1 or 0, the residuals can take only two possible values for any given $\hat{\pi}$. This will produce artifacts that can make the residual plot difficult to interpret.
- Standardized residuals can also be calculated using approximate variances, but the details are beyond the scope of this course.
- Approximate influence diagnostics are also available but are beyond the scope of this course.

7.6 Over-dispersion

It should also be noted that the assumption of binomial variability in the binary outcomes may also be violated in various ways. For example, the binomial distribution dictates that

$$E(Y) = n\pi = \mu \text{ and } \operatorname{var}(Y) = n\pi(1-\pi) = \mu(n-\mu)/n.$$

However, it is not the only possible discrete distribution on $0, 1, \ldots, n$ and it can happen that Y is such that $E(Y) = \mu$ but $\operatorname{var}(Y) \neq \mu(n-\mu)/n$.

- When $\operatorname{var}(Y) > \mu(n-\mu)/n$ the distribution is said to be over-dispersed (relative to the binomial).
- When $\operatorname{var}(Y) < \mu(n-\mu)/n$ the distribution is said to be under-dispersed (relative to the binomial).

Example In an experiment to measure the mortality of cancer cells under radiation conducted in the Department of Radiology, University of Cape Town, 400 cells were placed on a dish and the dishes were irradiated in batches of 3. After the cells were irradiated, the surviving cells were counted. Since cells also die naturally, dishes with cells were put into the radiation chamber without being irradiated, to establish the natural mortality. The following data show the numbers of surviving cells at zero dose from 9 batches.

```
cells<-read.table("cells.txt",header=TRUE)
cells

##    Batch Survived
## 1      1      178
## 2      1      193
## 3      1      217
## 4      2      109
## 5      2      112
## 6      2      115
## 7      3       66
## 8      3       75
## 9      3       80
## 10     4      118
## 11     4      125
## 12     4      137
## 13     5      123
## 14     5      146
## 15     5      170
## 16     6      115
## 17     6      130
## 18     6      133
## 19     7      200
## 20     7      189
## 21     7      173
## 22     8       88
## 23     8       76
## 24     8       90
## 25     9      121
## 26     9      124
## 27     9      136
```

```
batch<-factor(cells$Batch)
y<-cbind(cells$Survived,400-cells$Survived)
summary(glm(y~batch,family="binomial"))
```

```
##
## Call:
## glm(formula = y ~ batch, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4534  -0.7924   0.0000   0.6989   2.4324
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.04001    0.05775  -0.693    0.488
## batch2      -0.90446    0.08642 -10.466  < 2e-16 ***
## batch3      -1.44836    0.09424 -15.369  < 2e-16 ***
## batch4      -0.72913    0.08477  -8.601  < 2e-16 ***
## batch5      -0.51013    0.08323  -6.129 8.82e-10 ***
## batch6      -0.73684    0.08483  -8.686  < 2e-16 ***
## batch7      -0.08683    0.08174  -1.062    0.288
## batch8      -1.27490    0.09126 -13.970  < 2e-16 ***
## batch9      -0.72528    0.08474  -8.559  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 495.631  on 26  degrees of freedom
## Residual deviance:  32.794  on 18  degrees of freedom
## AIC: 219.77
##
## Number of Fisher Scoring iterations: 3
```

```
# Since these are grouped data, we can use the residual deviance
# to test the fit of the model
1-pchisq(32.794,df=18)
```

```
## [1] 0.01767412
```

Remarks

- The salient feature of the data is that there are clearly big differences between batches. This is obvious in the raw data and also supported by the value of the null deviance that could be used for a formal test.

- The data also contain three separate binomial counts within each batch. Since there should be no differences between the dishes within each batch, it would be natural to assume

  $$Y_{ij} \sim B(400, \pi_i) \text{ independently for } i = 1, 2, \ldots, 9 \text{ (batches); } j = 1, 2, 3 \text{ (replicates).}$$

  This model can be tested using the residual deviance and is rejected.
- The conclusion is that within batches the Y_{ij} are displaying greater variance than would be expected if they were binomial observations. That is, the data are over-dispersed relative to the binomial distribution.

□

In general, under-dispersion can also occur but is uncommon in practice. When over-dispersion is present, simple logistic regression cannot be usefully applied and more complicated analyses are required. Diagnosing over-dispersion is not always straightforward. In particular, if the data are not grouped or there are no replicate observations. A range of methods are available for dealing with over-dispersed data but are beyond the scope of this course.

Additional Material 1, 2018

8 Log linear models for count data

The linear regression model can also be generalised to allow for situations where the response is a count.

Example Ten experimental electronic devices were allowed to operate in two different modes of operation for varying periods. For each operating period, Mode 1 is the time spent operating in one mode and Mode 2 is the time spent operating in the other. The response variable is the number of failures observed in the total period of operation. The data are shown below. In this case the response variable `Failures` represents a count rather than a continuous measurement or a binomial proportion.

Mode1	Mode2	Failures
33.3	25.3	15
52.2	14.4	9
64.7	32.5	14
137.0	20.5	24
125.9	97.6	27
116.3	53.6	27
131.7	56.6	23
85.0	87.3	18
91.9	47.8	22

□

8.1 Poisson regression models

Consider data $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$. If the response variable Y is a count, it is natural to consider a Poisson model,

$$Y_i \sim Po(\mu_i) \text{ independently, for } i = 1, 2, \ldots, n.$$

The regression problem is then to relate the Poisson mean, $-mu_i$, to the predictor $\boldsymbol{x}_i$. That is, we seek a suitable model

$$M: \; \mu_i = \mu(\boldsymbol{x}_i).$$

To ensure that the Poisson mean μ_i is positive, we define

$$\eta_i = \log \mu_i$$

and consider log linear regression models

$$M: \; \eta_i = \boldsymbol{x_i}^T \boldsymbol{\beta} \text{ for } i = 1, 2, \ldots, n,$$

or, in matrix notation,

$$\boldsymbol{\eta} = X\boldsymbol{\beta}.$$

Analysis can be performed by maximum likelihood and the derivations are similar to those for the logistic regression model.

Starting with the log likelihood

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \log \left(\prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right) \text{ with } \log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_i,$$

it follows that the score vector is

$$\mathcal{S}(\boldsymbol{\beta}) = \left[\frac{\partial \ell}{\partial \beta_j} \right] = X^T (\boldsymbol{y} - \boldsymbol{\mu})$$

and the Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = X^T D_\mu X$$

where $D_\mu = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_n)$. The maximum likelihood equations $\mathcal{S}(\boldsymbol{\beta}) = \mathbf{0}$ do not in general have an explicit solution and the Fisher scoring algorithm is often used. Starting at an initial estimate, usually

$$\hat{\boldsymbol{\beta}}_0 = (X^T X)^{-1} X^T \log(\boldsymbol{y} + 0.5)$$

the iterative step

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (X^T D_{\mu^{(t)}} X)^{-1} X^T (\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$$

is applied until convergence.

8.2 Inference for Poisson regression models

8.2.1 Inference for regression coefficients

When the Poisson log linear regression model is fit, large sample standard errors for the parameters are obtained by taking the square-roots of the diagonal elements of the inverse information matrix. For large n or large μ_i, the approximate distribution of $\hat{\beta}_j$ is

$$N\left(\beta_j, \sqrt{i^{jj}(\hat{\boldsymbol{\beta}})} \right)$$

where i^{jj} is the jth diagonal element of $\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}$.

This allows for the construction of approximate confidence intervals and z-tests for individual parameters in the usual way.

8.2.2 Hypotheses concerning several parameters

Consider the log linear regression model

$$M: \ \boldsymbol{\eta} = X\boldsymbol{\beta} \text{ where } \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$$

and the hypothesis

$$H_0:\ \beta_p = \beta_{p-1} = \ldots = \beta_{p_0+1} = 0.$$

The log likelihood ratio test statistic is defined by

$$G^2 = 2(\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0))$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are the MLEs under M and H_0 respectively. If H_0 is true then the asymptotic distribution of G^2 is $\chi^2_{p-p_0}$ and we reject H_0 for

$$G^2 \geq \chi^2_{p-p_0}(\alpha).$$

The deviance statistic for any model M is defined by

$$D(M) = 2(\hat{\ell} - \ell(\hat{\boldsymbol{\beta}}))$$

where $\hat{\ell}$ is the value of the likelihod maximized without restriction and is produced in `R` as the `Residual Deviance`.

When the μ_i are all large, so that the normal approximation to the Poisson distribution applies, asymptotic distribution of the residual deviance is χ^2_{m-p}. In this case, the residual deviance can be used to test the overall fit of the model.

Example (Continued) The basic log linear regression model can be fit to the component failure data in `R` using the `glm` function.

```
failure<-read.table("failure.txt",header=TRUE)
failure

##   Mode1 Mode2 Failures
## 1  33.3  25.3       15
## 2  52.2  14.4        9
## 3  64.7  32.5       14
## 4 137.0  20.5       24
## 5 125.9  97.6       27
## 6 116.3  53.6       27
## 7 131.7  56.6       23
## 8  85.0  87.3       18
## 9  91.9  47.8       22

summary(glm(Failures~Mode1+Mode2,family=poisson,data=failure))

##
## Call:
## glm(formula = Failures ~ Mode1 + Mode2, family = poisson, data = failure)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.21984  -0.44735  -0.05893   0.68351   0.87510
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.175168   0.255456   8.515  < 2e-16 ***
## Mode1       0.007015   0.002429   2.888  0.00387 **
## Mode2       0.002549   0.002835   0.899  0.36852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16.9964  on 8  degrees of freedom
## Residual deviance:  4.0033  on 6  degrees of freedom
## AIC: 53.06
##
## Number of Fisher Scoring iterations: 4
```

In this case the residual deviance suggests that the multiple regression model provide an adequate fit. The naive interpretation is that the time spent in Mode 1 operation may be the important factor in failure. To explore this further, the predictor variables can be recoded to the total operating time and the proportion of the total spent in Mode 1.

```
Total<-failure$Mode1+failure$Mode2
Prop<-failure$Mode1/Total
summary(glm(Failures~Total+Prop,family=poisson,data=failure))
```

```
##
## Call:
## glm(formula = Failures ~ Total + Prop, family = poisson, data = failure)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3367  -0.3906  -0.2340   0.6864   0.9259
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.910202   0.574514   3.325 0.000885 ***
## Total       0.005272   0.001549   3.403 0.000666 ***
## Prop        0.447455   0.693969   0.645 0.519072
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16.9964  on 8  degrees of freedom
## Residual deviance:  4.6578  on 6  degrees of freedom
## AIC: 53.715
##
## Number of Fisher Scoring iterations: 4
```

This analysis suggests that the total time in operation is the important factor rather than the time spent in Mode 1. The reason that Mode 1 but not Mode 2 appears significant in the initial analysis is due to collinearity.

```
summary(glm(Failures~Mode2,family=poisson,data=failure))
```

```
##
## Call:
## glm(formula = Failures ~ Mode2, family = poisson, data = failure)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9451  -0.9648   0.1867   0.5382   1.6664
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.699895   0.158116  17.075   <2e-16 ***
## Mode2       0.005736   0.002636   2.176   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16.996  on 8  degrees of freedom
## Residual deviance: 12.356  on 7  degrees of freedom
## AIC: 59.413
##
## Number of Fisher Scoring iterations: 4
```

```
1-pchisq(12.4,7)
```

```
## [1] 0.08814848
```

□

The apparent dependence of the failure rate on the total operating time suggests the rate may be proportional to the operating time. In general, models such as

$$\mu_i = t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$$

can be considered in such cases where t_i is the total operating time (or more generally total exposure to risk) and $\boldsymbol{x}_i$ is the vector of other predictor variables. Taking logs yields

$$\eta_i = \log \mu_i = \log t_i + \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

This model corresponds to including $\log t_i$ as a predictor in the log linear regression and *forcing its coefficient to be 1*. Such models are implemented in the `lm` and `glm` functions by specifying an `offset`

Example (Continued) Based on the residual deviance, the simple model of proportionality appears to be suitable.

```
summary(glm(Failures~Prop,family=poisson,offset=log(Total),data=failure))
```

```
##
## Call:
## glm(formula = Failures ~ Prop, family = poisson, data = failure,
##     offset = log(Total))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9325  -0.5393  -0.1202   0.5311   2.2604
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2957     0.4488  -5.115 3.14e-07 ***
## Prop          0.5044     0.6662   0.757    0.449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 8.1774  on 8  degrees of freedom
## Residual deviance: 7.6079  on 7  degrees of freedom
## AIC: 54.665
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(Failures~1,family=poisson,offset=log(Total),data=failure))
```

```
##
## Call:
## glm(formula = Failures ~ 1, family = poisson, data = failure,
##     offset = log(Total))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.32400  -0.68909   0.09131   0.52366   2.11140
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.96222    0.07474  -26.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8.1774  on 8  degrees of freedom
## Residual deviance: 8.1774  on 8  degrees of freedom
## AIC: 53.234
##
## Number of Fisher Scoring iterations: 4
```

8.3 Residuals

Deviance residuals and Pearson residuals are also defined for Poisson regression models.

The Pearson residuals

$$r_i^{(p)} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}};$$

The Deviance residuals

$$r_i^{(d)} = \operatorname{sgn}(y_i - \hat{\mu}_i)\sqrt{2\left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right)}.$$

Either of these residuals may be plotted against the fitted values and also the individual predictors to provide an informal assessment of model fit. As in the binomial case, it must be kept in mind that these residuals are not standardized

Example An insurance company recorded the number of policies held and the number of accident claims in 64 categories defined by

- `District`: 1-4, where 4 represents major cities;
- `Engine`: Engine capacity of the car, < 1 litre, $1 - 1.5$ litre, $1.5 - 2$ litre, > 2 litre;
- `Age`: < 25, $25 - 29$, $30 - 35$, > 35.

Since it is reasonable to expect the number of claims to be proportional to the number of policies within each class, the number of policies is used as and offset.

```
insurance<-read.csv("insurance.csv",header=TRUE)
# The most complicated model we can consider is no three factor interaction
m1<-glm(Claims ~ District*Engine*Age-District:Engine:Age
        ,offset=log(Policies)
        ,data = insurance
        ,family = poisson)
# summary(m1) produces large amount of output; extracting elements from m1 instead
summary(m1)$call
```

```
## glm(formula = Claims ~ District * Engine * Age - District:Engine:Age,
##     family = poisson, data = insurance, offset = log(Policies))
```

```
head(summary(m1)$coefficients)
```

```
##                  Estimate Std. Error      z value     Pr(>|z|)
## (Intercept)  -1.67022621  0.1392118 -11.99773801 3.651396e-33
## DistrictD2    0.15185663  0.1734587   0.87546281 3.813221e-01
## DistrictD3    0.16355121  0.2355240   0.69441434 4.874224e-01
## DistrictD4   -0.05989171  0.3102918  -0.19301736 8.469454e-01
## Engine1-1.5l  0.05087680  0.1632101   0.31172573 7.552490e-01
## Engine1.5-2l -0.01339125  0.1982379  -0.06755142 9.461427e-01
```

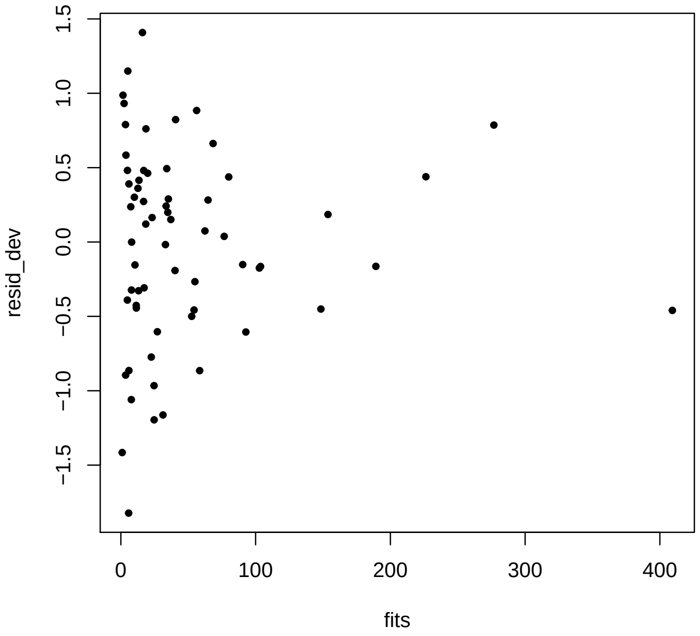
```
tail(summary(m1)$coefficients)
```

```
##                         Estimate Std. Error   z value    Pr(>|z|)
## Engine1-1.5l:Age30-35 0.2096221  0.2181218 0.9610322 0.336535984
## Engine1.5-2l:Age30-35 0.6685406  0.2476032 2.7000486 0.006932935
## Engine2+l:Age30-35    0.3411600  0.3818123 0.8935281 0.371574454
## Engine1-1.5l:Age35+   0.1713639  0.1690966 1.0134083 0.310865175
## Engine1.5-2l:Age35+   0.4246343  0.2041806 2.0796997 0.037553086
## Engine2+l:Age35+      0.3357971  0.3382998 0.9926023 0.320903812
```

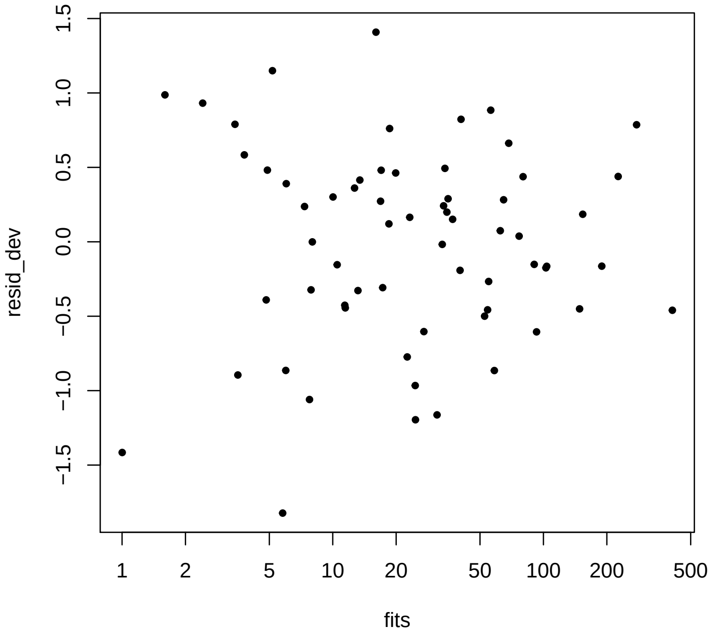
```
cat("Null deviance:",summary(m1)$null.deviance,"on",summary(m1)$df.null,"degrees of freedom"
,"\nResidual deviance:",summary(m1)$deviance,"on",summary(m1)$df.residual,"degrees of freedom"
,"\nAIC:",summary(m1)$aic,"\n")
```

```
## Null deviance: 236.259 on 63 degrees of freedom
## Residual deviance: 27.28967 on 27 degrees of freedom
## AIC: 418.6112
```

```
# Residual deviance looks fine
# Extract and plot deviance residuals for a visual check
resid_dev<-residuals(m1,type="deviance")
fits<-fitted(m1)
plot(fits,resid_dev,pch=20)
```



```
# Also plot against log fits as scale is distorted by one very large value
plot(fits,resid_dev, pch=20, log="x")
# No obvious pattern in residual plot
```



```
# Finally use a hypothesis test to check that offset term is correct.
m2<-update(m1,.~.+log(Policies))
summary(m2)$call
```

```
## glm(formula = Claims ~ District + Engine + Age + log(Policies) +
##     District:Engine + District:Age + Engine:Age, family = poisson,
##     data = insurance, offset = log(Policies))
```

```
summary(m2)$coefficients[1:11,]
```

```
##                  Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)   -3.38199144  4.3375831 -0.7796949 0.43557048
## DistrictD2     0.42549221  0.7139473  0.5959715 0.55119428
## DistrictD3     0.73209670  1.4585489  0.5019350 0.61571326
## DistrictD4     0.67061941  1.8753799  0.3575912 0.72064925
## Engine1-1.5l  -0.06783394  0.3421849 -0.1982377 0.84285913
## Engine1.5-2l   0.11380767  0.3781548  0.3009552 0.76344865
## Engine2+l      0.98443221  1.8070166  0.5447831 0.58590267
## Age25-29      -0.48936759  0.3117312 -1.5698382 0.11645277
## Age30-35      -0.72195289  0.3051045 -2.3662483 0.01796939
## Age35+        -1.41024476  1.7515493 -0.8051414 0.42073811
## log(Policies)  0.32382918  0.8201993  0.3948177 0.69297746
```

```
tail(summary(m2)$coefficients)
```

```
##                          Estimate Std. Error     z value  Pr(>|z|)
## Engine1-1.5l:Age30-35  0.01886557  0.5295518  0.03562555 0.9715809
## Engine1.5-2l:Age30-35  0.44251709  0.6230117  0.71028697 0.4775262
## Engine2+l:Age30-35    -0.05670816  1.0733112 -0.05283478 0.9578635
## Engine1-1.5l:Age35+    0.03843199  0.3767011  0.10202252 0.9187388
## Engine1.5-2l:Age35+    0.30063400  0.3744333  0.80290401 0.4220302
## Engine2+l:Age35+       0.05816735  0.7765242  0.07490732 0.9402884
```

```
cat("Null deviance:",summary(m2)$null.deviance,"on",summary(m2)$df.null,"degrees of freedom"
,"\nResidual deviance:",summary(m2)$deviance,"on",summary(m2)$df.residual,"degrees of freedom"
,"\nAIC:",summary(m2)$aic,"\n")

## Null deviance: 236.259 on 63 degrees of freedom
## Residual deviance: 27.13342 on 26 degrees of freedom
## AIC: 420.4549

# log(Policies) term is not significant so offset is OK.
```

```
m2<-update(m1,.~.-Engine:Age)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Engine + Age + District:Engine + District:Age
## Model 2: Claims ~ District * Engine * Age - District:Engine:Age
##   Resid. Df Resid. Dev Df Deviance
## 1        36     37.685
## 2        27     27.290  9   10.395
```

```
m2<-update(m1,.~.-District:Age)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Engine + Age + District:Engine + Engine:Age
## Model 2: Claims ~ District * Engine * Age - District:Engine:Age
##   Resid. Df Resid. Dev Df Deviance
## 1        36     33.527
## 2        27     27.290  9   6.2374
```

```
m2<-update(m1,.~.-District:Engine)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Engine + Age + District:Age + Engine:Age
## Model 2: Claims ~ District * Engine * Age - District:Engine:Age
##   Resid. Df Resid. Dev Df Deviance
## 1        36     34.457
## 2        27     27.290  9   7.1668
```

```
# District:Age is least significant so remove
# No need to calculate p-value because G^2<df
m1<-update(m1,.~.-District:Age)
m2<-update(m1,.~.-Engine:Age)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Engine + Age + District:Engine
## Model 2: Claims ~ District + Engine + Age + District:Engine + Engine:Age
##   Resid. Df Resid. Dev Df Deviance
## 1        45     44.132
## 2        36     33.527  9   10.604
```

```
m2<-update(m1,.~.-District:Engine)
anova(m2,m1)

## Analysis of Deviance Table
```

```
##
## Model 1: Claims ~ District + Engine + Age + Engine:Age
## Model 2: Claims ~ District + Engine + Age + District:Engine + Engine:Age
##   Resid. Df Resid. Dev Df Deviance
## 1        45     40.907
## 2        36     33.527  9   7.3803
```

```
# Now remove District:Engine
m1<-update(m1,.~.-District:Engine)
m2<-update(m1,.~.-Engine:Age)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Engine + Age
## Model 2: Claims ~ District + Engine + Age + Engine:Age
##   Resid. Df Resid. Dev Df Deviance
## 1        54     51.420
## 2        45     40.907  9   10.513
```

```
# Can also remove Engine:Age
m1<-update(m1,.~.-Engine:Age)
# Now check for simplification by removal of main effects
m2<-update(m1,.~.-Engine)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Age
## Model 2: Claims ~ District + Engine + Age
##   Resid. Df Resid. Dev Df Deviance
## 1        57     140.09
## 2        54      51.42  3   88.667
```

```
m2<-update(m1,.~.-Age)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ District + Engine
## Model 2: Claims ~ District + Engine + Age
##   Resid. Df Resid. Dev Df Deviance
## 1        57     136.29
## 2        54      51.42  3    84.87
```

```
m2<-update(m1,.~.-District)
anova(m2,m1)

## Analysis of Deviance Table
##
## Model 1: Claims ~ Engine + Age
## Model 2: Claims ~ District + Engine + Age
##   Resid. Df Resid. Dev Df Deviance
## 1        57     65.291
## 2        54     51.420  3   13.871
```

```
# No further simplification possible
# Examine model parameters
summary(m1)

##
```

```
## Call:
## glm(formula = Claims ~ District + Engine + Age, family = poisson,
##     data = insurance, offset = log(Policies))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.46558  -0.50802  -0.03198   0.55555   1.94026
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.82174    0.07679 -23.724  < 2e-16 ***
## DistrictD2    0.02587    0.04302   0.601 0.547597
## DistrictD3    0.03852    0.05051   0.763 0.445657
## DistrictD4    0.23421    0.06167   3.798 0.000146 ***
## Engine1-1.5l  0.16134    0.05053   3.193 0.001409 **
## Engine1.5-2l  0.39281    0.05500   7.142 9.18e-13 ***
## Engine2+l     0.56341    0.07232   7.791 6.65e-15 ***
## Age25-29     -0.19101    0.08286  -2.305 0.021149 *
## Age30-35     -0.34495    0.08137  -4.239 2.24e-05 ***
## Age35+       -0.53667    0.06996  -7.672 1.70e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 236.26  on 63  degrees of freedom
## Residual deviance:  51.42  on 54  degrees of freedom
## AIC: 388.74
##
## Number of Fisher Scoring iterations: 4
```

Conclusions

- The log additive (multiplicative) model

$$\mu_{ijk} = n_{ijk} e^{\alpha_i + \beta_j + \gamma_k}$$

  provides an adequate description of the data, where α_i is the effect of District i, β_j is the effect of Engine j, γ_k is the effect of Age group k and n_{ijk} is the number of policies.

- Based on the parameter estimates District 4 (major cities) carries signifcantly higher risk than the other three categories that appear to be roughly equal. Since $e^{0.23421} = 1.26$ we estimate that claims occur roughly 26% more frequently in major cities.

- Based on the parameter estimates, the frequency of claims also increases with engine size and decreases with age.