MATH3823 Generalized Linear Models

Robert G Aykroyd

1/24/23

Table of contents

VV	eekly	schedule	3
O,	vervie		4
	Pref	ace	4
O	fficial	Module Description	5
	Mod	lule summary	5
		ectives	
	_	abus	
	-	versity Module Catalogue	
1	Intr	oduction	6
	1.1	Overview	6
	1.2	Motivating example	7
	1.3	Revision of least-squares estimation	
	1.4	Types of variables	
	1.5	Exercises	
2	Esse	entials of Normal Linear Models 1	2
	2.1	Overview	2
	2.2	Types of normal linear model	9
	2.3	Matrix representation of linear models	0
	2.4	Construction of the design matrix	
		2.4.1 Example: Simple linear regression	
		2.4.2 Example: One-way ANOVA	
	2.5	Model shorthand notation	
	2.6	Exercises 2	

Weekly schedule

! Important

Our regular class times are:

Tuesday 11-12, Roger Stevens, LT25 Thursday 2-3, Roger Stevens, LT23

i Week 1 (30 January - 3 February)

- Before next Lecture: Please read the Overview.
- Lecture on Tuesday: We will briefly cover all material in Chapter 1: Introduction
- Before next Lecture: Please re-read Chapter 1 carefully.
- Lecture on Thursday: Start Chapter 2: Essentials of Normal Linear Models with Section 2.1: Overview.
- Weekly feedback: Self-study the Exercises in Section 1.5 solutions to be posted during Week 1.

i Week 2 (6 - 10 February)

• Details will be added during Week 1.

i Coursework Practical Sessions (20 - 24 March)

• Details to follow in early March.

Overview

Preface

These lecture notes are produced for the University of Leeds module MATH3823 - Generalized Linear Models for the academic year 2022-23. Please note that this material also forms part of the module MATH5824 - Generalized Linear and Additive Models. They are based on those used previously for this module and I am grateful to previous module lecturers for their considerable effort: Lanpeng Ji, Amanda Minter, John Kent, Wally Gilks, and Stuart Barber. This is the first year, however, that they have been produced in accessible format and hence some errors might occur during this conversion process. For information, I am using Quarto (a successor to RMarkdown) from RStudio to produce both the html and PDF, and then GitHub to create the website which can be accessed at rgaykroyd.github.io/MATH3823/. Please note that the PDF versions will only be made available on the University of Leeds Minerva system. Although I am a long-term user of RStudio, I have not previously used Quarto/RMarkdown nor Github and hence please be patient if there are hitches along the way.

RG Aykroyd, Leeds, November 22, 2022



Warning

Statistical ethics and sensitive data

Please note that from time to time we will be using data sets from situations which some might perceive as sensitive. All such data sets will, however, be derived from real-world studies which appear in textbooks or in scientific journals. The daily work of many statisticians involves applying their professional skills in a wide variety of situations and as such it is important to include a range of commonly encountered examples in this module. Whenever possible, sensitive topics will be signposted in advance. If you feel that any examples may be personally upsetting then, if possible, please contact the module lecturer in advance. If you are significantly effected by any of these situations, then you can seek support from the Student Counselling and Wellbeing service.

Official Module Description

Module summary

Linear regression is a tremendously useful statistical technique but is very limited. Generalised linear models extend linear regression in many ways - allowing us to analyse more complex data sets. In this module we will see how to combine continuous and categorical predictors, analyse binomial response data and model count data.

Objectives

On completion of this module, students should be able to:

- a) carry out regression analysis with generalised linear models including the use of link functions;
- b) understand the use of deviance in model selection;
- c) appreciate the problems caused by overdispersion;
- d) fit and interpret the special cases of log linear models and logistic regression;
- e) use a statistical package with real data to fit these models to data and to write a report giving and interpreting the results.

Syllabus

Generalised linear model; probit model; logistic regression; log linear models.

University Module Catalogue

For any further details, please see MATH3823 Module Catalogue page

1 Introduction

1.1 Overview

In previous modules you have studied linear models with a normally distributed error term, such as simple linear regression, multiple linear regression and ANOVA for normally distributed observations. In this module we will study **generalized** linear models.

Outline of the module:

- 1. Revision of linear models with normal errors.
- 2. Introduction to generalized linear models, GLMs.
- 3. Logistic regression models.
- 4. Loglinear models, including contingency tables.

Important

This module will make extensive use of \mathbf{R} and hence it is very important that you are comfortable with its use. If you need some revision, then material is available on Minerva under $RStudio\ Support$.

The purpose of a generalized linear model is to describe the dependence of a *response* variable y on a set of p explanatory variables $x = (x_1, x_2, \dots, x_p)$ where, conditionally on x, observation y has a distribution which is **not necessarily** normal.

Note that in these notes we may use lowercase letters, for example y or y_i , to denote both observed values or random variables, which is being considered should be clear from the context.

Important

This module will make extensive use of many basic ideas from statistics. If you need some revision, then see *Appendix A: Basic material* on Minerva under *Basic Prerequisite Material*.

1.2 Motivating example

Table 1.1 shows data¹ on the number of beetles killed by five hours of exposure to 8 different concentrations of gaseous carbon disulphide.

Table 1.1: Numbers of beetles killed by five hours of exposure to 8 different concentrations of gaseous carbon disulphide

Dose	No. of beetle	No. killed
x_i	m_i	y_{i}
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Figure 1.1a shows the same data with a linear regression line superimposed. Although this line goes close to the plotted points, we can see some fluctuations around it. More seriously, this is a stupid model: it would predict a mortality rate of greater than 100% at a dose of 1.9 units, and a negative mortality rate at 1.65 units!

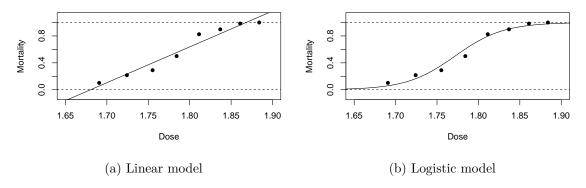


Figure 1.1: Beetle mortality rates with fitted dose- response curves.

A more sensible dose—response relationship for the beetle mortality data might be based on the *logistic* function (to be defined later), as plotted in Figure 1.1b. The resulting curve is a closer, more-sensible, fit. Later in this module we will see how this curve was fitted using maximum likelihood estimation for an appropriate generalized linear model.

¹Dobson and Barnett, 3rd edn, p.127

This is an example of a dose-response experiment which are widely used in medical and pharmaceutical situations.

Warning

Warning of potentially sensitive material. For further information on doseresponse experiments see, for example, www.britannica.com/science/dose-responserelationship.

1.3 Revision of least-squares estimation

Suppose that we have n paired data values $(x_1, y_1), \dots, (x_n, y_n)$ and that we believe these are related by a linear model

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

for all $i \in \{1, 2, \dots, n\}$, where $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed (iid) with $\mathbf{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. The aim will be to find values of the model parameters, α, β and σ^2 using the data. Specifically, we will estimate α and β using the values which minimize the residual sum of squares (RSS)

$$RSS(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$
 (1.1)

This measures how close the data points are around the regression line and hence the resulting estimates, $\hat{\alpha}$ and $\hat{\beta}$, will give us a fitted regression line which is *closest* to the data.

It can be shown that Equation 1.1 takes its minimum when the parameters are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \text{and} \quad \hat{\beta} = \frac{s_{xy}}{s_x^2}$$
 (1.2)

where \bar{x} and \bar{y} are the sample means,

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

is the sample covariance and

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

is the sample variance of the x values. It can be shown that these estimators are unbiased, that is $E[\hat{\alpha}] = \alpha$ and $E[\hat{\beta}] = \beta$ – see Section 1.5.

The fitted regression lines is then given by $\hat{y} = \hat{\alpha} + \hat{\beta}x$, the fitted values by $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, and the model residuals by $r_i = \hat{\epsilon}_i = y_i - \hat{y}_i$ for all $i \in \{1, \dots, n\}$.

To complete the model fitting, we also estimate the error variance, σ^2 , using

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n r_i^2. \tag{1.3}$$

Note that, by construction, $\bar{r} = 0$ and, further, it can be shown that $\hat{\sigma}^2$ is an unbiased estimator of σ^2 , that is $E[\hat{\sigma}^2] = \sigma^2$.

Returning to the above beetle data example, we have $\hat{\alpha} = -8.947843$, $\hat{\beta} = 5.324937$, and $\hat{\sigma}^2 = 0.0075151$.

We will interpret the output later, but in R, the fitting can be done with a single command with corresponding fitting output from a second command:

Call:

lm(formula = mortality ~ dose)

Residuals:

Min 1Q Median 3Q Max -0.10816 -0.06063 0.00263 0.05119 0.12818

Coefficients:

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08669 on 6 degrees of freedom Multiple R-squared: 0.9524, Adjusted R-squared: 0.9445

F-statistic: 120.2 on 1 and 6 DF, p-value: 3.422e-05

Important

You should have met R output like this in previous statistics modules, but if you need some revision then see Appendix-C: Background to Analysis of Variance on Minerva under Basic Pre-requisite Material.

1.4 Types of variables

The way a variable enters a model will depends on its type. The most common five types of variable are:

1. Quantitative

- a. Continuous: for example, height; weight; duration. Real valued. Note that although recorded data is rounded it is still usually best regarded as continuous.
- b. Count (discrete): for example, number of children in a family; accidents at a road junction; number of items sold. Non-negative and integer-valued.

2. Qualitative

- a. Ordered categorical (ordinal): for example, severity of illness (Mild/ Moderate/Severe); degree classification (first/ upper-second/ lower-second/ third).
- b. Unordered categorical (nominal):
 - Dichotomous (binary): two categories: for example sex (M/F); agreement (Yes/No); coin toss (Head/Tail).
 - Polytomous (also known as polychotomous): more than two categories: for example blood group (A/ B/ O); eye colour (Brown/ Blue/ Green).

Note that although dichotomous is clearly a special case of polytomous, making the distinction is usually worthwhile as it often leads to a simplified modelling and testing approach.

1.5 Exercises

Important

Unless otherwise stated, data files will be available online at: rgaykroyd.github.io/MATH3823/Datasets/filename.ext, where filename.ext is the stated filename with extension.

1.1 Consider again the beetle data in Table 1.1. Perform the calculations by hand and then check the answers using R – a copy of the data is available in the file beetle.txt. Finally plot the fitted regression line on a scatter plot of the data. [Hint: See the code chunk used to produce Figure 1.1.]

1.2 Consider the following synthetic data:

	i = 1	i = 2	i = 3	i=4	i = 5	i = 6	i = 7	i = 8
$\overline{x_i}$	-1	0	1	2	2.5	3	4	6
y_{i}	-2.8	-1.1	7.2	8.0	8.9	9.2	14.8	24.7

Plot the data to check that a linear model is suitable and then fit a linear regression model. Do you think that the fitted model can be reliably used to predict the values of y when x = 5 and x = 10? Justify your answers.

1.3 Starting from Equation 1.1, derive the estimation equations given in Equation 1.2. Further, show that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β . [Hint: Check your MATH1712 lecture notes.]

What can be said about $\hat{\sigma}^2$ as an estimator of σ^2 ? [Hint: There is a careful theoretical proof, but here only an intuitive explanation is expected.]

1.4 The *Brownlee's Stack Loss Plant Data*² is already available in **R**, with background details on the help page, ?stackloss. [Hint: You already met this example in MATH1712.]

After plotting all pairs of variables, which of Air.Flow, Water.Temp and Acid.Conc do you think could be used to model stack.loss using a linear regression? Justify your answer.

Perform a simple linear regression with using stack.loss as the response variable and your chosen variable as the explanatory variable. Add the fitted regression line to a scatter plot of the data and comment.

1.5 In an experiment conducted by de Silva et al. in 2020³ data was obtained to investigate falling objects and gravity, as first consider by Galileo and Newton. A copy of the data is available in the file physics_from_data.csv.

Read the data file into R and perform a simple linear regression of the maximum Reynolds number as the response variable and, in turn, each of the other variables as the explanatory variable.

Plot the data and add the corresponding fitted linear models. Which variable do you think helps explain Reynolds number the best? Why do you think this?

Here are an infinite number of further numerical examples from **maths e.g.** (thanks to https://www.mathcentre.ac.uk/):

Finding the intersercept

Finding the slope - Part 1

Finding the slope - Part 2

²Brownlee, K. A. (1960, 2nd ed. 1965) Statistical Theory and Methodology in Science and Engineering. New York: Wiley. pp. 491–500.

³de Silva BM, Higdon DM, Brunton SL, Kutz JN. Discovery of Physics From Data: Universal Laws and Discrepancies. Front Artif Intell. 2020 Apr 28;3:25. doi: 10.3389/frai.2020.00025. PMID: 33733144; PMCID: PMC7861345.

2 Essentials of Normal Linear Models

2.1 Overview

In many fields of application, we might assume the response variable is normally distributed. For example: heights, weights, log prices, etc.

The data¹ in Table 2.1 record the birth weights of 12 girls and 12 boys and their gestational ages (time from conception to birth).

A key question is, can we predict the birth weight of a baby born at a given gestational age using these data. For this we will need to make assumptions about the relationship between birth weight and gestational age, and any associated natural variation – that is we require a model.

First we should explore the data. Figure 2.1a shows a histogram of the birth weights indicating a spread around modal group 2800-3000 grams; Figure 2.1b indicates slightly higher birth weights for the boys than the girls; and Figure 2.1c shows an increasing relationship between weight and age. Together, these suggest that gestational age and sex are likely to be important for predicting weight.

Before considering possible models, Figure 2.2 again shows the relationship between weight and age but this time with the points coloured according to the baby's sex. This, perhaps, shows the boys to have generally higher weights across the age range than girls.

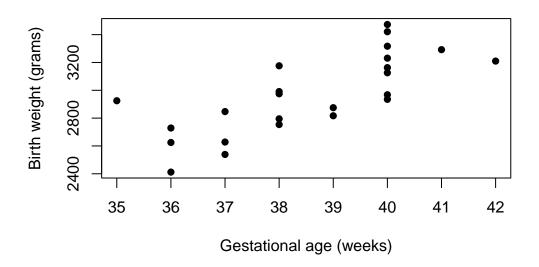
Of course, there are very many possible models, but here we will consider the following:

```
\begin{array}{lll} \hline \text{Model 0:} & \text{Weight} = \alpha \\ \hline \text{Model 1:} & \text{Weight} = \alpha + \beta. \\ \hline \text{Model 2:} & \text{Weight} = \alpha + \beta. \\ \hline \text{Model 3:} & \text{Weight} = \alpha + \beta. \\ \hline \text{Age} + \gamma. \\ \hline \text{Sex} + \delta. \\ \hline \text{Age.Sex} \\ \hline \end{array}
```

In these models, Weight is called the *response* variable (sometimes called the *dependent* variable) and Age and Sex are called the *covariates* or *explanatory* variables (sometimes called the *predictor* or *independent* variables). Here, Age is a continuous variable whereas Sex is coded as a *dummy* variable taking the value 0 for girls and 1 for boys; it is an example of a *factor*, in this case with just two *levels*: Girl and Boy.

¹Dobson and Barnett, 3rd edition, Table 2.3.





(c) Relationship beween variables

Figure 2.1: Birthweight and gestational age for 24 babies.

Table 2.1: Gestational ages (in weeks) and birth weights (in grams) for 24 babies (12 girls and 12 boys).

(a) Gi	rls		(b) Boys			
Gestational Age	Gestational Age Birth weight		Gestational Age	Birth weight		
40	3317		40	2968		
36	2729		38	2795		
40	2935		40	3163		
37	2754		35	2925		
42	3210		36	2625		
39	2817		37	2847		
40	3126		41	3292		
37	2539		40	3473		
36	2412		37	2628		
38	2991		38	3176		
39	2875		40	3421		
40	3231		38	3975		

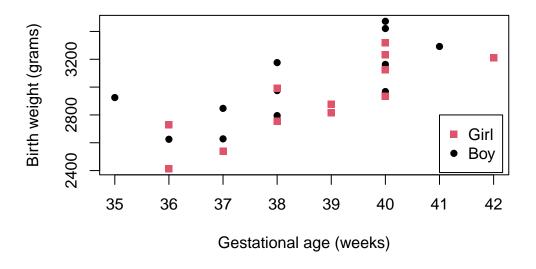


Figure 2.2: Birthweight and gestational age for 12 girls (red squares) and 12 boys (black dots).

Note that Model 0 is a special case of Model 1 (consider the situation when $\beta=0$) and that Model 1 is a special case of Model 2 (consider the situation when $\gamma=0$) and finally that Model 2 is a special case of Model 3 (consider the situation when $\delta=0$) – such models are called *nested*.

In these models, α , β , γ and δ are model parameters. Parameter α is called the *intercept* term; β is called the main effect of Age; and is interpreted as the effect on birth weight per week of gestational age. Similarly, γ is the main effect of Sex, interpreted as the effect on birth weight of being a boy (because girl is the baseline category).

Parameter δ is called the *interaction effect* between Age and Sex. Take care when interpreting an interaction effect. Here, it does not mean that age somehow affects sex, or vice-versa. It means that the effect of gestational age on birth weight depends on whether the baby is a boy or a girl.

These models can be fitted to the data using (Ordinary) *Least Squares* to produce the results presented in Figure 2.3.

Which model should we use?

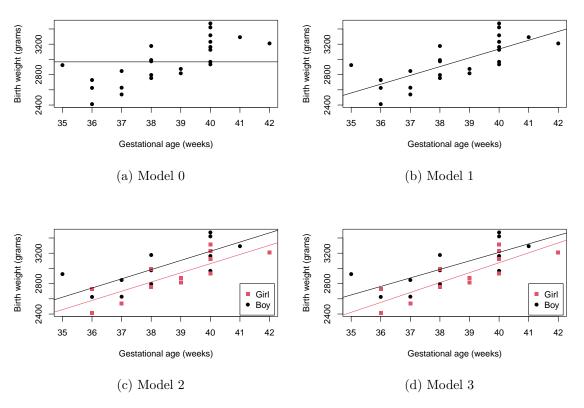


Figure 2.3: Birthweight and gestational age data with superimposed fitted regression lines from various competing models.

We know from previous modules that statistical tests can be used to check the importance of regression coefficients and model parameters, but it is also important to use the graphical results, as in Figure 2.3, to guide us.

Model 0 says that there is no change in birth weight with gestational age which means that we would use the average birth weight as the prediction whatever the gestational age – this makes no sense. As we can easily see from the scatter plot of the data, the fitted line in this case is clearly inappropriate.

Model 1 does not take into account whether the baby is a girl or a boy, but does model the relationship between birth weight and gestational age. This does seem to provide a good fit and might be adequate for many purposes. Recall from Figure 2.1b and Figure 2.2, however, that for a given gestational age the boys seem to have a higher birth weight than the girls.

Model 2 does take the sex of the baby into account by allowing separate intercepts in the fitted lines – this means that the lines are parallel. By eye, there is a clear difference between these two lines but it might not be important.

Model 3 allows for separate slopes as well as intercepts. There is a slight difference in the slopes, with the birth weight of the girls gradually catching-up as the gestational age increases. It is difficult to see, however, if this will be a general pattern or if it is only true for this data set – especially given the relatively small sample size.

Here, it is not clear by eye which of the fitted models will be the best and hence we should use a statistical test to help. In particular, we can choose between the models using F-tests.

Let y_i denote the value of the dependent variable Weight for individual $i=1,\ldots,n,$ and let the four models be indexed by k=0,1,2,3.

Let R_k denote the residual sum of squares (RSS) for Model k:

$$R_k = \sum_{i=1}^n (y_i - \hat{\mu}_{ki})^2, \tag{2.1}$$

where $\hat{\mu}_{ki}$ is the fitted value for individual *i* under Model *k*. Let r_k denote the corresponding residual degrees of freedom for Model *k* (the number of observations minus the number of model parameters).

Consider the following hypotheses:

$$H_0: \text{Model } 0 \text{ is true}; \quad H_1: \text{Model } 1 \text{ is true}.$$

Under the null hypothesis H_0 , the difference between R_0 and R_1 will be purely random, so the between-models mean-square $(R_0-R_1)/(r_0-r_1)$ should be comparable to the residual mean-square R_1/r_1 . Thus our test statistic for comparing Model 1 to the simpler Model 0 is:

$$F_{01} = \frac{(R_0 - R_1)/(r_0 - r_1)}{R_1/r_1}. (2.2)$$

It can be shown that, under the null hypothesis H_0 , the statistic F_{01} will have an F-distribution on $r_0 - r_1$ and r_1 degrees of freedom, which we write: $F_{r_0 - r_1, r_1}$. Under the alternative hypothesis H_1 , the difference $R_0 - R_1$ will tend to be larger than expected under H_0 , and so the observed value F_{01} will probably lie in the upper tail of the $F_{r_0 - r_1, r_1}$ distribution.

Returning to the birth weight data, we obtain the following output from R when we fit Model 1:

```
(Intercept) age
-1484.9846 115.5283
```

Analysis of Variance Table

```
Response: weight
```

Df Sum Sq Mean Sq F value Pr(>F) ge 1 1013799 1013799 27.33 3.04e-05 ***

Residuals 22 816074 37094

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Thus we have parameter estimates: $\hat{\alpha} = -1484.98$ and $\hat{\beta} = 115.5$. The Analysis of Variance (ANOVA) table gives: $R_0 - R_1 = 1013799$ with $r_0 - r_1 = 1$ and $R_1 = 816074$ with $r_1 = 22$.

If we wanted R_0 and r_0 then we can either fit Model 0 or get them by subtraction.

The F_{01} statistic, Equation 2.2, is then

$$F_{01} = \frac{113799/1}{816074/22} = 27.33,$$

which can be read directly from the ANOVA table in the column headed 'F value'.

Is $F_{01}=27.33$ in the upper tail of the $F_{1,22}$ distribution? (See Figure 2.4 and note that 27.33 is very far to the right.) The final column of the ANOVA table tells us that the probability of observing $F_{01}>27.33$ is only 3.04×10^5 – this is called a p-value. The *** beside this p-value highlights that its value lies between 0 and 0.001. This indicates that we should reject H_0 in favour of H_1 – there is very strong evidence for the more complicated model. Thus we would conclude that the effect of gestational age is statistically significant in these data.

Next, consider the following hypotheses:

$$H_0: \mathtt{Model}\ 1 \ \mathrm{is}\ \mathrm{true}; \quad H_1: \mathtt{Model}\ 2 \ \mathrm{is}\ \mathrm{true}.$$

Under the null hypothesis H_0 , the difference between R_1 and R_2 will be purely random, so the between-models mean-square $(R_1 - R_2)/(r_1 - r_2)$ should be comparable to the residual



Figure 2.4: Probability density function of ${\cal F}_{01}$ distribution.

mean-square R_2/r_2 . Thus our test statistic for comparing Model 2 to the simpler Model 1 is:

$$F_{12} = \frac{(R_1 - R_2)/(r_1 - r_2)}{R_2/r_2}. (2.3)$$

Under the null hypothesis H_0 , the statistic F_{12} will have an F-distribution on r_1-r_2 and r_2 degrees of freedom, which we write: $F_{r_1-r_2,r_2}$. Under the alternative hypothesis H_1 , the difference R_1-R_2 will tend to be larger than expected under H_0 , and so the observed value F_{12} will probably lie in the upper tail of the $F_{r_1-r_2,r_2}$ distribution.

Returning to the birth weight data, we obtain the following output from R (where sexM denotes Boy):

```
(Intercept) age sexM
-1773.3218 120.8943 163.0393
```

Analysis of Variance Table

Response: weight

Df Sum Sq Mean Sq F value Pr(>F)
age 1 1013799 1013799 32.3174 1.213e-05 ***
sex 1 157304 157304 5.0145 0.03609 *
Residuals 21 658771 31370

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Thus we have parameter estimates: $\hat{\alpha} = -1773.3$, $\hat{\beta} = 120.9$ and $\hat{\gamma} = 163.0$, the latter being the effect of being a boy compared to the baseline category of being a girl.

The Analysis of Variance (ANOVA) table gives: $R_1 - R_2 = 157304$ with $r_1 - r_2 = 1$, and $R_2 = 658771$ with $r_2 = 21$. The F_{12} statistic, Equation 2.3, is then

$$F_{12} = \frac{157304/1}{658771/21} = 5.0145,$$

which can be read directly from the ANOVA table in the column headed 'F value'. Is $F_{12} = 5.01$ in the upper tail of the $F_{1,21}$ distribution?

The final column of the ANOVA table tells us that the probability of observing $F_{12} > 5.01$ is only 0.03609 – this is called a p-value. The * beside this p-value highlights that its value lies between 0.01 and 0.05. This indicates that we should reject H_0 in favour of H_1 – there is evidence for the more complicated model. Thus we would conclude that the effect of the sex of the baby, after controlling for gestational age, is statistically significant in these data.

To complete the analysis, we should now compare Model 2 with Model 3 - see Exercises.

2.2 Types of normal linear model

Here we consider how normal linear models can be set up for different types of explanatory variable. The dependent variable y is modelled as a linear combination of p explanatory variables $x=(x_1,x_2,\ldots,x_p)$ plus a random error $\epsilon \sim N(0,\sigma^2)$, where '~' means 'is distributed as'. Several models are of this kind, depending on the number and type of explanatory variables. Table 2.3 lists some types of normal linear models with their explanatory variable types.

Table 2.3: Types of normal linear model and their explanatory variable types where indicator function I(x = j) = 1 if x = j and 0 otherwise.

p	Explanatory variables	Model
1	Quantitative	Simple linear regression $y = \alpha + \beta x + \epsilon$
>1	Quantitative	Multiple linear regression $y = \alpha + \sum_{i=1}^{p} \beta_i x_i + \epsilon$
1	Dichotomous $(x = 1 \text{ or } 2)$	Two-sample t-test $y = \alpha + \delta I(x = 2) + \epsilon$
1	Polytomous, k levels $(x = 1, \dots, k)$	One-way $\text{ANOVA} y = \alpha + \sum_{i=1}^{k} \delta_i \ I(x=j) + \epsilon$

Table 2.3: Types of normal linear model and their explanatory variable types where indicator function I(x = j) = 1 if x = j and 0 otherwise.

>1	Qualitative	p-way ANOVA	
ŕ -	SQ -2	p way mino vii	

For the two-sample t-test model², observations in the two groups have means $\alpha + \beta_1$ and $\alpha + \beta_2$. Notice, however, that we have three parameters with only two group sample means and hence parameter estimation is not possible. To avoid this identification problem, we either impose a 'corner' constraint: $\beta_1 = 0$ and then β_2 represents the difference in the Group 2 mean relative to a baseline of Group 1. Alternatively, we may impose a 'sum-to-zero' constraint: $\beta_1 + \beta_2 = 0$, the values $\beta_1 = -\beta_2$ then give differences in the groups means relative to the overall mean. Table 2.4 shows the effect of the parameter constraint on the group means.

Table 2.4: Parameters in the two-sample t-test model after imposing parameter constraint to avoid the identification problem.

Constraint	Group 1 mean	Group 2 mean
$\beta_1 = 0$	α	$\alpha + \beta_2$
$\beta_1 + \beta_2 = 0$	$\alpha-\beta_2$	$\alpha + \beta_2$

For the general one-way ANOVA model with k groups, observations in Group j have mean $\alpha+\delta_j$, for $j=1,\ldots,k$ – that leads to k+1 parameters describing k group means. Again we can impose the 'corner' constraint: $\delta_1=0$ and then δ_j represents the difference in means between Group j and the baseline Group 1. Alternatively, we may impose a 'sum-to-zero' constraint: $\sum_{j=1}^k \delta_j = 0$ and again $(\delta_1,\delta_2,\ldots,\delta_k)$ then represents an individual group effect relative to the overall data mean.

2.3 Matrix representation of linear models

All of the models in Table 2.3 can be fitted by least squares (OLS). To describe this, a matrix formulation will be most convenient:

$$\mathbf{Y} = X\beta + \epsilon \tag{2.4}$$

where

- Y is an $n \times 1$ vector of observed response values with n being the number of observations.
- X is an $n \times p$ design matrix, to be discussed below.
- β is a $p \times 1$ vector of parameters or coefficients to be estimated.

²Notice that this is a special case of the one-way ANOVA when there are only two-groups.

• ϵ is an $n \times 1$ vector of independent and identically distributed (IID) random variables, which here $\epsilon \sim N(0, \sigma^2)$ and is called the "error" term.

2.4 Construction of the design matrix

Creating the design matrix is a key part of the modelling as it describes the important structure of investigation or experiment. The design matrix can be constructed by the following process.

- 1. Begin with an X containing only one column: a vector of ones for the overall mean or intercept term (the α in Table 2.3).
- 2. For each explanatory variable x_i , do the following:
 - a. If a variable x_i is quantitative, add a column to X containing the values of x_i .
 - b. If x_j is qualitative with k levels, add k "dummy" columns to X, taking values 0 and 1, where a 1 in the ℓ th dummy column identifies that the corresponding observation is at level ℓ of factor x_j . For example, suppose we have a factor $\mathbf{x}_j = (M, M, F, M, F)$ representing the sex of n = 5 individuals. This information can be coded into two dummy columns of X:

$$\begin{bmatrix}
F & M \\
0 & 1 \\
0 & 1 \\
1 & 0 \\
0 & 1 \\
1 & 0
\end{bmatrix}$$

3. When qualitative variables are present, X will be singular – that is, there will be linear dependencies between the columns of X. For example, the sum of the two columns above is a vector of ones, the same as the intercept column. We resolve this identification problem by deleting some columns of X. This is equivalent to applying the corner constraint $\delta_1 = 0$ in the one-way ANOVA.

In the above example, after removing a column, we get:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

4. Each column of X represents either a quantitative variable, or a level of a qualitative variable. We will use $i=1,\ldots,n$ to label the observations (rows of X) and $j=1,\ldots,p$ to label the columns of X.

21

2.4.1 Example: Simple linear regression

Consider the simple linear regression model $y = \alpha + \beta x + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. Given data on n pairs $(x_i, y_i), i = 1, ..., n$, we write this as

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \text{for } i = 1, 2, \dots, n, \tag{2.5}$$

where the ϵ_i are IID $N(0, \sigma^2)$. In matrix form, this becomes

$$\mathbf{Y} = X\beta + \epsilon \tag{2.6}$$

with

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The *i*th row of Equation 2.6 has the same meaning as Equation 2.5:

$$y_i = 1 \times \beta_1 + x_i \times \beta_2 + \epsilon_i = \alpha + \beta x_i + \epsilon_i$$
, for $i = 1, 2, \dots, n$.

2.4.2 Example: One-way ANOVA

For one-way ANOVA with k levels, the model is

$$y_i = \alpha + \sum_{i=1}^k \delta_j \ I(x_i = j) + \epsilon_i, \quad \text{for } i = 1, 2, \dots, n,$$

where x_i denotes the group level of individual i. So if y_i is from the jth group then $y_i \sim N(\alpha + \delta_j, \sigma^2)$. Here α is the intercept and the $(\delta_1, \delta_2, \dots, \delta_k)$ represent the "main effects".

We can store the information about the levels of g in a dummy matrix $X^* = (x_{ij}^*)$ where

$$x_{ij}^* = \begin{cases} 1, & g_i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Then set $X = [1, X^*]$, where 1 is an *n*-vector of 1's. For the male–female example at (1.12), we have n = 5 and a sex factor:

$$g = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \alpha \\ \delta_1 \\ \delta_2 \end{bmatrix}.$$

Then the *i*th row of X becomes $\beta_1 + \beta_2 = \alpha + \delta_1$ if $g_i = 1$ and $\beta_1 + \beta_3 = \alpha + \delta_2$ if $g_i = 2$. That is, the *i*th row of X is

$$\alpha + \sum_{i=1}^{2} \delta_{j} I(g_{i} = j)$$

so this model can be written $Y = X\beta + \epsilon$. Here, X is singular: its last two columns added together equal its first column. Statistically, the problem is that we are trying to estimate two means (the mean response for Boys and the mean response for girls) with three parameters $(\alpha, \delta_2 \text{ and } \delta_2)$.

In practice, we often resolve this aliasing or identification problem by setting one of the parameters to be zero, that is $\delta_1 = 0$, which corresponds to deleting the second column of X).

2.5 Model shorthand notation

In R, a qualitative (categorical) variable is called a *factor*, and its categories are called *levels*. For example, variable Sex in the birthweight data (above) has levels coded "M" for 'Boy' and "F" for 'Girl'. It may not be obvious to R whether a variable is quantitative or qualitative. For example, a qualitative variable called **Grade** might have categories 1, 2 and 3. If **grade** was included in a model, R would treat it as quantitative unless we declare it to be a factor, which we can do with the command:

grade = as.factor(grade)

A convenient model-specification notation has been developed from which the design matrix X can be constructed. Below, E, F, ... denote generic quantitative (continuous) or qualitative (categorical) variables. Terms in this notation may take the following forms:

- a. 1: a column of 1's to accommodate an intercept term (the α 's of Table 2.3). This is included in the model by default.
- b. E: variable E is included in the model. The design matrix includes k_E columns for E. If E is quantitative, $k_E = 1$. If E is qualitative, k_E is the number of levels of E minus 1.
- c. E+F : both E and F are included the model. The design matrix includes k_E+k_F columns accordingly.
- d. E: F (sometimes $E \cdot F$): the model includes an interaction between E and F; each column that would be included for E is multiplied by each column for F in turn. The design matrix includes $k_E \times k_F$ columns accordingly.
- e. E * F: shorthand for 1 + E + F + E : F: useful for crossed models where E and F are different factors. For example, E labels age groups; F labels medical conditions.

- f. E/F: shorthand for 1+E+E:F: useful for nested models where F is a factor whose levels have meaning only within levels of factor E. For example, E labels different hospitals; F labels wards within hospitals.
- g. $\operatorname{poly}(E;\ell)$: shorthand for an orthogonal polynomial, wherein x contains a set of mutually orthogonal columns containing polynomials in E of increasing order, from order 1 through order ℓ .
- h. -E: shorthand for removing a term from the model; for example E * F E is short for 1 + F + E : F.
- i. I(): shorthand for an arithmetical expression (not to be confused with the indicator function of equation (1.10)). For example, I(E+F) denotes a new quantitative variable constructed by adding together quantitative variables E and F. This would cause an error if either E or F has been declared as a factor. What would happen in this example if we omitted the I() notation?

The notation uses "~" as shorthand for "is modelled by" or "is regressed on". For example,

• Weight is regressed on age-group and sex with no interaction between them:

$${\tt Weight} \sim {\tt Age} + {\tt Sex}$$

as for the birthweight data in Figure 1.2c.

• Well being is regressed on age-group and income-group, where income is thought to affect wellbeing differentially by age:

Wellbeing
$$\sim$$
 Age $*$ Income

• Class of degree is regressed on school of the university and on degree subject within the school:

$${\tt DegreeClass} \sim {\tt School/Subject}$$

• Yield of wheat is regressed on seed-variety and annual rainfall:

Yield
$$\sim$$
 Variety + poly(Rainfall, 2)

• Profit is regressed on amount invested:

Profit
$$\sim$$
 Investment -1

(no intercept term, that is a regression through the origin).

See Handout 4 for material on intrinsic aliasing to deal with singularity problem.

2.6 Exercises

2.1. An extra model which could have been considered for the Birthweight data example would be one that says that Weight is different for girls and boys, but does not depend on gestational age.

Write down the equation corresponding to this model. Then, load the birthweight data into RStudio and fit the model. How are the fitted model parameters related to the overall birthweight mean and the mean birthweights of the girls and boys? Is this a good fit to the data? Is Sex statistically significant?

2.2. In an experiment to investigate Ohm's Law, V = IR where V is Voltage, I is current and R is resistance of the material, the following data³ were recorded:

Table 2.5: Experimental verification of Ohm's Law

Voltage (Volts)	4	8	10	12	14	18	20	24
Current (mAmps)	11	24	30	36	40	53	58.5	70

Does this data support Ohm's Law? What is the resistance of the material used?



Warning

Warning of potentially sensitive material.

2.3. Data⁴ were recording on the number of deaths due to breast cancer in 301 counties in the US states of North Carolina, South Carolina and Georgia between 1950 and 1960. Also recorded was the adult white female population for each county. A copy of the data is available in the file Breast Cancer.txt.

³Aykroyd, P.J. (1956). Unpublished.

⁴Rice, J.A. (1995) Mathematical Statistics and Data Analysis (2nd Ed). Example 57, p235.