MATH3823 Generalized Linear Models

Robert G Aykroyd

2022-11-29T00:00:00+00:00

Table of contents

Preface

These lecture notes are produced for the University of Leeds module "MATH3823 - Generalized Linear Models" for the academic year 2022-23. 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

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 Gaussian linear models.
2. Introduction to generalized linear models, GLMs.
3. Logistic regression models.
4. Loglinear models, including contingency tables.

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 use lowercase y or y_i to denote both observed values or random variables, which should be clear from the context.

⚠ Warning

Statistical ethics and sensitive data
Please note that from time to time we will be using data sets from situations which might be perceived 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 common types of 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 effected by any of these situations, then please consider talking with the Student Counselling and Wellbeing service.

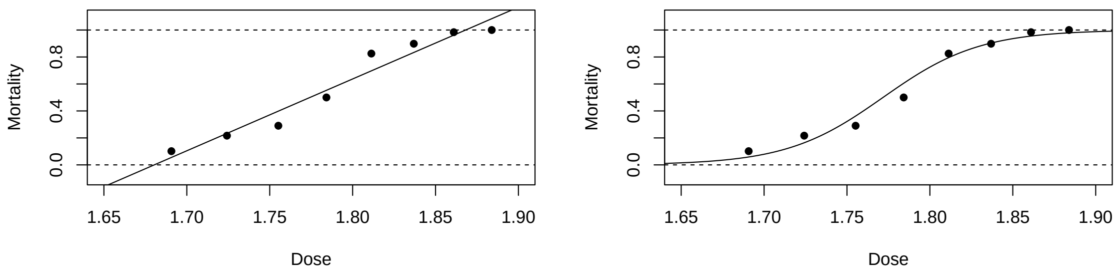
1.2 Motivating example

Table 1.1 shows data[1] 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!



(a) Linear model

(b) Logistic model

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.

[1] Dobson and Barnett, 3rd edn, p.127

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

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

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 2.1 takes its minimum when the parameters are given by

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

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 $\mathrm{E}[\hat{\alpha}] = \alpha$ and $\mathrm{E}[\hat{\beta}] = \beta$.

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, $\bar{r} = 0$. Further, it can be shown that this is an unbiased estimator, that is $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$.

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:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.9478     0.8717  -10.27 4.99e-05 ***
dose          5.3249     0.4857   10.96 3.42e-05 ***
---
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
```

You should have met *R* output like this in previous statistics modules, but if you need some revision then see Chapter 9.

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[2]: 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

1.1 Consider the following synthetic data:

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$
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 value of y when $x = 10$? Justify your answer.

1.2 Starting from Equation 1.1, derive the estimation equations given in Equation 1.2. Further, show that $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased estimators.

Hint: Check your MATH1712 lecture notes.

1.3 In an experiment conducted by de Silva et al. in 2020[3] data was obtained to investigate falling objects and gravity, as first consider by Galileo and Newton. A copy of the data is available on Minerva in the file: physics_from_data.csv

Read the data file into RStudio 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/):

[2] Also known as polychotomous.

[3] de 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.

Finding the intersercept
Finding the slope - Part 1
Finding the slope - Part 2

2 Essentials of Gaussian Linear Models

2.1 Overview

In many fields of application, we might assume the response variable is Gaussian, that is normally distributed; for example: heights, weights, log prices.

The data[1] 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 with modal group 2800-300grams; 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.

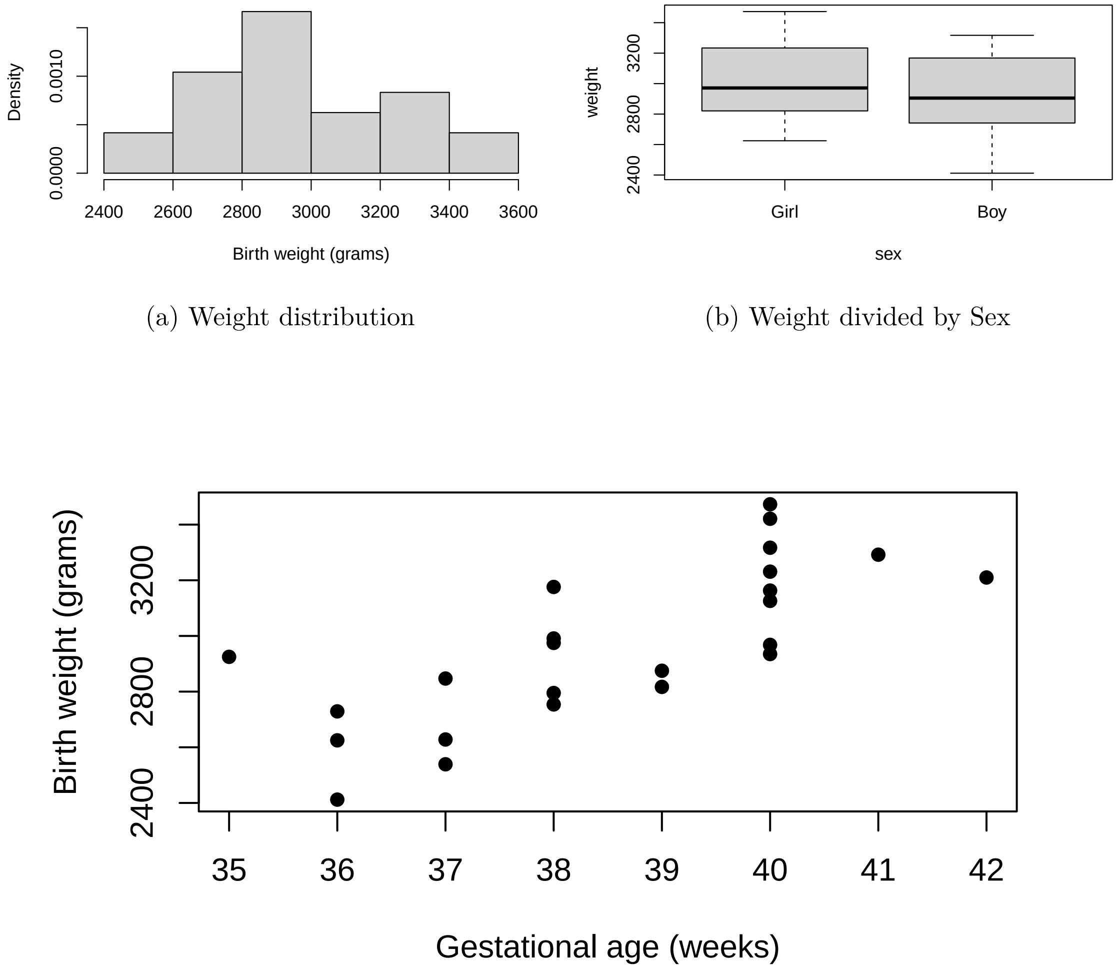
Before considering possible models, Figure 2.2 again shows the relationship between weight and age but with the points coloured according to the baby's sex. This, perhaps, shows the boys to have generally higher weights across the age range than girls.

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

$\texttt{Model 0}:$	$\texttt{Weight} = \alpha$
$\texttt{Model 1}:$	$\texttt{Weight} = \alpha + \beta.\texttt{Age}$
$\texttt{Model 2}:$	$\texttt{Weight} = \alpha + \beta.\texttt{Age} + \gamma.\texttt{Sex}$
$\texttt{Model 3}:$	$\texttt{Weight} = \alpha + \beta.\texttt{Age} + \gamma.\texttt{Sex} + \delta.\texttt{Age.Sex}$

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*: Boy and Girl.

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



(a) Weight distribution

(b) Weight divided by Sex

(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) Girls

Gestational Age	Birth weight
40	3317
36	2729
40	2935
37	2754
42	3210
39	2817
40	3126
37	2539
36	2412
38	2991
39	2875
40	3231

(b) Boys

Gestational Age	Birth weight
40	2968
38	2795
40	3163
35	2925
36	2625
37	2847
41	3292
40	3473
37	2628
38	3176
40	3421
38	3975

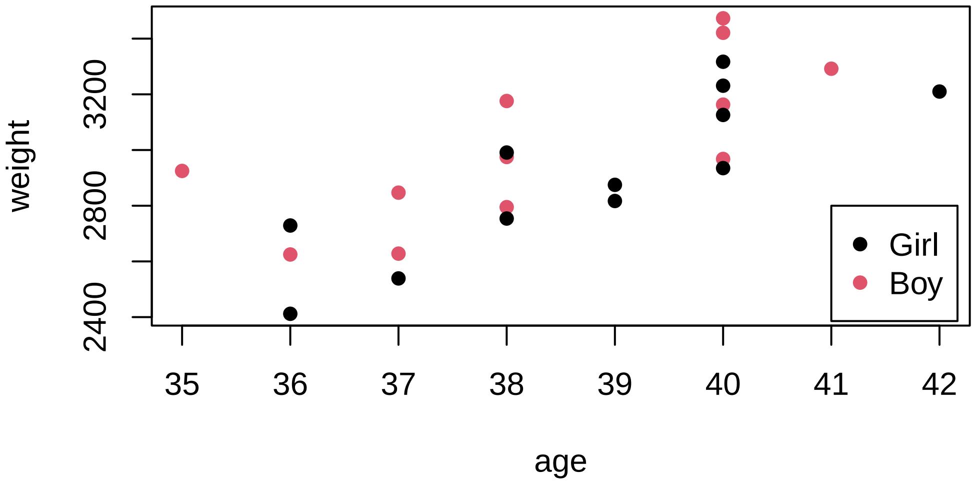


Figure 2.2: Birthweight and gestational age for 12 girls (red dots) and 12 boys (black dots).

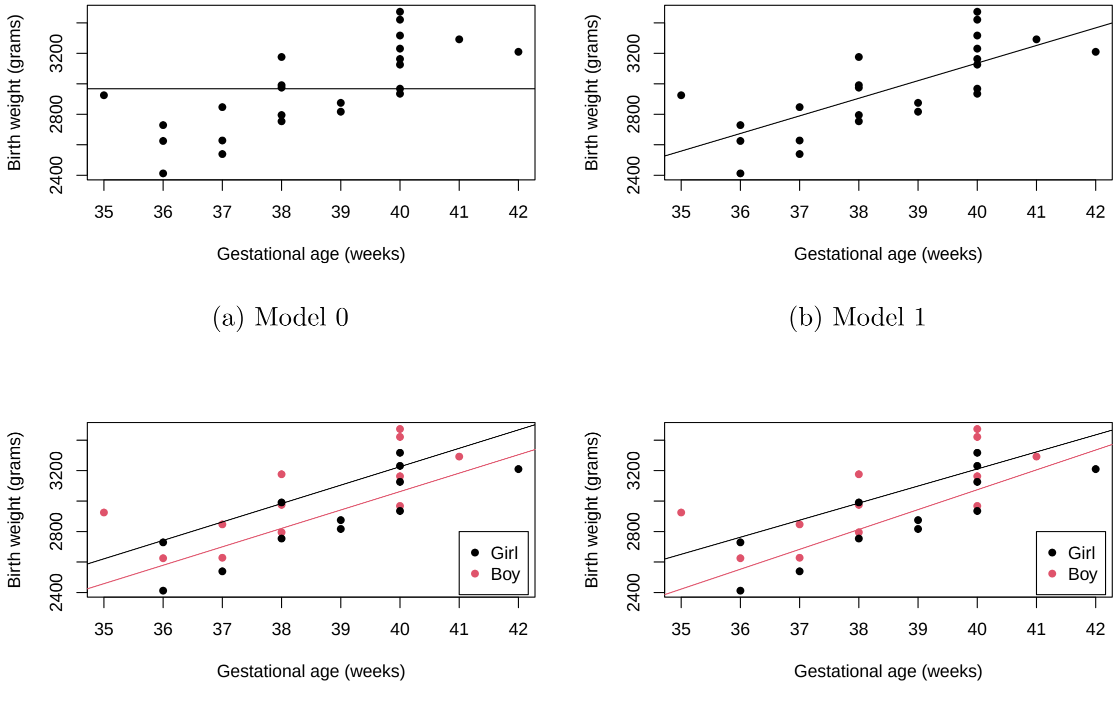
Note that `Model 0` is a special case of `Model 1` (consider the case when $\beta = 0$) and that `Model 1` is a special case of `Model 2` (consider the case when $\gamma = 0$) and finally that `Model 2` is a special case of `Model 3` (consider the case when $\delta = 0$) – such models are called *nested.*

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 9.1.

Which model should we use?



(a) Model 0

(b) Model 1

(c) Model 2

(d) Model 3

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 9.1, 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 can use a statistical test to help. In particular, we can choose between the models using F-tests.

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

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:

$$H_0 : \texttt{Model 0 is true;} \qquad H_1 : \texttt{Model 1 is true.}$$

Under the null hypothesis H_0, the difference between R_0 and R_1 will be purely random, so the between-model 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:

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

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:

```
(Intercept)         age
 -1484.9846    115.5283


Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value   Pr(>F)
age        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) gives: $R_0 - R_1 = 1013799$ with $r_1 - r_1 = 1$ 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 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. Thus we would conclude that the effect of gestational age is statistically significant in these data.

Next, consider the following hypotheses:

$$H_0 : \texttt{Model}\ 1 \text{ is true}; \qquad H_1 : \texttt{Model}\ 2 \text{ is true}.$$

Under the null hypothesis H_0, the difference between R_1 and R_2 will be purely random, so the between-model mean-square $(R_1 - R_2)/(r_1 - r_2)$ should be comparable to the residual mean-square R_2/r_2. Thus our test statistic for comparing Model 2 to the simpler Model 1 is:

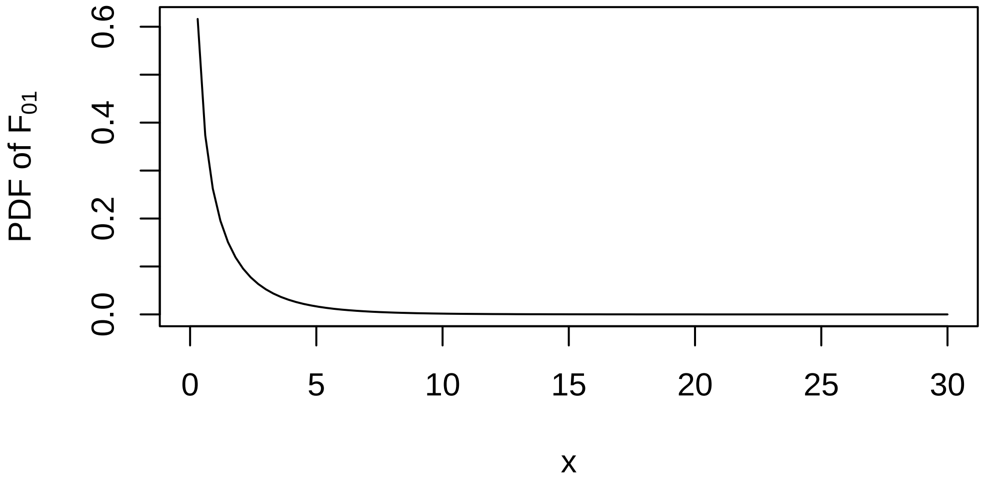


Figure 2.4: Probability density function of F_{01} distribution.

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

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) 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 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. Thus we would conclude that the effect of the sex of the baby, after controlling for gestational age, is statistically significant in these data.

2.2 Types of Gaussian linear model

We consider how Gaussian 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 $x = (x_1, x_2, \dots, x_p)$ plus a random error $\epsilon \sim N(0, \sigma^2)$, where '$\sim$' 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 Gaussian 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 ($x = 1$ or 2)	Two-sample t-test $y = \alpha + \delta\ I(x = 2) + \epsilon$
1	Polytomous, k levels ($x = 1, \dots, k$)	One-way ANOVA$y = \alpha + \sum_{j=1}^{k} \delta_j\ I(x = j) + \epsilon$
>1	Qualitative	p-way ANOVA

For the two-sample t-test model[2], 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

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

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 $j = 1, \dots, k$ – 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 $(\delta_1, \delta_2, \dots, \delta_k)$ 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 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

- $\mathbf{Y}$ 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.
- 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_j, do the following:

   a. If a variable x_j is quantitative, add a column to X containing the values of x_j.
   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 ℓth 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:

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

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 $i = 1, \dots, n$ to label the observations (rows of X) and $j = 1, \dots, p$ to label the columns of X.

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 $(x_i, y_i), i = 1, \dots, n$, 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 ith row of Equation 2.6 has the same meaning as Equation 2.5:

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

2.4.2 Example: One-way ANOVA

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

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

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 ith 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 ith row of X is

$$\alpha + \sum_{j=1}^{2} \delta_j I(g_i = j)$$

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 (α, δ_2 and δ_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. $\text{poly}(E; \ell)$: 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:

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

  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:

  $$\texttt{Wellbeing} \sim \texttt{Age} * \texttt{Income}$$

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

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

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

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

- Profit is regressed on amount invested:

  $$\texttt{Profit} \sim \texttt{Investment} - 1$$

  (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

1. An exta model which could have been considered for the Birthweight data example would be one that say 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. Consider a new data set...
3.

3 GLM Theory

3.1 Motivating examples

We cannot always assume that the dependent variable y is normally distributed. For example, for the beetle mortality data in Table 1.1, suppose each beetle subjected to a dose x_i has a probability p_i of being killed. Then the number of beetles killed y_i out of a total number m_i at dose-level x_i will have a $\text{Bin}(m_i, p_i)$ distribution:

$$\Pr(y_i;\ p_i, m_i) = \begin{pmatrix} m_i \\ y_i \end{pmatrix} p_i^{y_i}(1 - p_i)^{m_i - y_i} \tag{3.1}$$

where y_i takes values in $\{0, 1, \dots, m_i\}$.

Table 3.1 contains seasonal data on tropical cyclones for 13 seasons. Suppose that, within season i, there is a constant probability $\lambda_i dt$ of a cyclone occurring in any short time-interval dt. Then the total number of cyclones y_i during season i will have a Poisson distribution with mean λ_i, that is $y_i \sim \text{Po}(\lambda_i)$:

$$\Pr(y_i;\ \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \tag{3.2}$$

where y_i takes values in $\{0, 1, 2, \dots\}$.

Table 3.1: Numbers of tropical cyclones in $n = 13$ successive seasons[1]

Season	1	2	3	4	5	6	7	8	9	10	11	12	13
No of cyclones	6	5	4	6	6	3	12	7	4	2	6	7	4

In these two examples, we have non-normal data and would like to know whether and how the dependent variable y_i depends on the covariate x_i or i.

Generalized linear models provide a modelling framework for data analysis in the non-normal setting. We will revisit the beetle mortality and cyclone data sets after describing the structure of a generalized linear model.

[1] Dobson and Barnett, 3rd edn, Table 1.2

3.2 The GLM structure

A *generalized linear model* relates a continuous or discrete response variable y to a set of explanatory variables $x = (x_1, \dots, x_p)$. The model contains three parts:

Random part: The probability function of y is assumed to belong to the *two-parameter exponential family* of distributions with parameters θ and ϕ:

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\}, \tag{3.3}$$

where $\phi > 0$. Here, θ is called the *canonical* or *natural* parameter of the distribution and ϕ is called the *scale* parameter. We show below that the mean $\mathrm{E}[y]$ depends only on θ, and $\mathrm{Var}[y]$ depends on ϕ and possibly also θ. Various choices for functions $b(\cdot)$ and $c(\cdot)$ produce a wide variety of familiar distributions (see below). Sometimes we may set $\phi = 1$; then Equation 3.3 is called the *one-parameter exponential family.*

Further, note that in some references to generalized linear models (such as Dobson and Barnett, 3rd edn.), ϕ does not appear at all in the exponential family formula Equation 3.3, instead it is absorbed into θ and $b(\theta)$.

In this module, we will generally assume that each observation y_i, $i = 1, \dots, n$, is *independently* drawn from an exponential family where θ depends on the covariates for each unit of observation i. Thus we write

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}.$$

Note the subscripts on both y and θ.

Systematic part: This is a *linear predictor*:

$$\eta = \sum_{j=1}^{p} \beta_j x_j. \tag{3.4}$$

Link function: This is an isomorphic function providing the link between the linear predictor η and the mean $\mu = \mathrm{E}[y]$:

$$\eta = g(\mu), \quad \text{and} \quad \mu = g^{-1}(\eta) = h(\eta). \tag{3.5}$$

Here, $g(\mu)$ is called the *link function*, and $h(\eta)$ is called the *inverse link function.*

We will now discuss each of these parts in more detail.

3.3 The random part of a GLM

We begin with some examples of exponential family members.

3.3.1 Example: Poisson distribution

If y has a Poisson distribution with parameter λ, $y \sim \text{Po}(\lambda)$, then y takes values in $\{0, 1, 2, \dots\}$ and has probability mass function:

$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\left\{y \log\lambda - \lambda - \log y!\right\}, \tag{3.6}$$

which has the form of Equation 3.3 with components as in Table 3.2.

Table 3.2: Exponential model components for the Poisson

θ	ϕ	$b(\theta)$	$c(y,\phi)$
$\log\lambda$	1	$\lambda = e^\theta$	$-\log y!$

For example, to model the cyclones data in Table 3.1, we might simply assume that the number of cyclones in each season has a Poisson distribution, assuming a constant rate λ across all seasons i. That is $y_i \sim \text{Po}(\lambda)$.

3.3.2 Example: Binomial distribution

Let y have a Binomial distribution, (write $y \sim \text{Bin}(m, p)$ with m fixed. Then y is discrete, taking values in $\{0, 1, \dots, m\}$, and has probability mass{#tbl-GLM-poisson} function:

$$f(y) = \binom{m}{y} p^y (1-p)^{m-y} = \binom{m}{y}\left(\frac{p}{1-p}\right)^y (1-p)^m$$

which can be re-written as

$$f(y) = \exp\left\{y \text{ logit } p + m\log(1-p) + \log\binom{m}{y}\right\}, \tag{3.7}$$

which has the form of Equation 3.3 with,

$$\theta = \text{logit } p = \log\left(\frac{p}{1-p}\right),$$

and with components as in Table 3.3.

Table 3.3: Exponential model components for the Binomial

θ	ϕ	$b(\theta)$	$c(y,\phi)$
logit p	1	$m\log(1+e^\theta)$	$\log\binom{m}{y}$

Where it can be shown that $-m\log(1-p) = m\log(1+e^\theta)$ – see Exercises.

3.3.3 Example: Normal distribution

Let y have a Normal distribution with mean μ and variance σ^2. Then y takes values on the whole real line and has probability density function

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\}, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2}{2\sigma^2}+\frac{y\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu-\mu^2/2}{\sigma^2}+\left[\frac{-y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right]\right\}, \end{aligned}$$

which has the form of Equation 3.3 with components as in Table 3.4.

Table 3.4: Exponential model components for the Gaussian

θ	ϕ	$b(\theta)$	$c(y,\phi)$
μ	σ^2	$\theta^2/2$	$-\frac{y^2}{2\phi}-\frac{1}{2}\log(2\pi\phi)$

Where it can be shown that $\frac{-y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)=-\frac{y^2}{2\phi}-\frac{1}{2}\log(2\pi\phi)$ – see Exercises.

From the usual regression point of view, we write $y=\alpha+\beta x+\epsilon$, with $\epsilon\sim N(0,\sigma^2)$. From the point of view of a generalized linear model, we write $y\sim N(\mu,\sigma^2)$ where $\mu(x)=\alpha+\beta x$.

3.4 Moments of exponential-family distributions

It is straightforward to find the mean and variance of Y in terms of $b(\theta)$ and ϕ. Since we want to explore the dependence of E[Y] on explanatory variables, this property makes the exponential family very convenient.

Proposition 3.1. *For random variables in the exponential family:*

$$E[Y]=b'(\theta), \quad \textit{and} \quad Var[Y]=b''(\theta)\phi. \tag{3.8}$$

Proof We give the proof for a continuous random variables. For the discrete case, replace all integrals by sums.

Starting with the simple property that all probability density functions integrate to 1, we have

$$1=\int \exp\left\{\frac{y\theta-b(\theta)}{\phi}+c(y,\phi)\right\}dy$$

and then differentiating both sides with respect to θ gives

$$0 = \int \left[\frac{y - b'(\theta)}{\phi}\right] \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}\ dy. \tag{3.9}$$

Next, using the definition of the exponential family to simplify the equation gives

$$0 = \int \left[\frac{y - b'(\theta)}{\phi}\right] f(y;\theta)\ dy$$

and expanding the brackets leads to

$$0 = \frac{1}{\phi}\left(\int yf(y;\theta)dy - b'(\theta)\int f(y;\theta)\ dy\right).$$

The first integral is simply the expectation of Y and the second is the integral of the probability density function of Y, and hence

$$0 = \frac{1}{\phi}\left(\mathrm{E}[Y] - b'(\theta)\right)$$

which implies that

$$\mathrm{E}[Y] = b'(\theta), \tag{3.10}$$

which proves the first part of the proposition.

Differentiating Equation 3.9 by parts and then using the definition of the exponential family to simplify again yields

$$0 = \int \left\{-\frac{b''(\theta)}{\phi} + \left[\frac{y - b'(\theta)}{\phi}\right]^2\right\} f(y;\theta)\ dy$$

and using Equation 3.10 gives,

$$0 = -\frac{b''(\theta)}{\phi} + \int \left[\frac{y - \mathrm{E}[Y]}{\phi}\right]^2 f(y;\theta)\ dy$$

$$0 = -\frac{b''(\theta)}{\phi} + \frac{\mathrm{Var}[Y]}{\phi^2}$$

which implies that

$$\mathrm{Var}[Y] = \phi\ b''(\theta).$$

which proves the second part of the proposition.

Together, these two results allow us to write down the expectation and variance for any random variable once we have shown that it is a member of the exponential family.

Table 3.5: Summary of moment calculations via exponential family properties

	θ	$b(\theta)$	ϕ	$\mathrm{E}[Y] =$ $b'(\theta)$	$b''(\theta)$	$\mathrm{Var}[Y] =$ $b''(\theta)\phi$
Poisson, $Po(\lambda)$	$\log \lambda$	e^{θ}	1	$e^{\theta} = \lambda$	e^{θ}	$e^{\theta} \times 1 = \lambda$
Normal, $N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2	$\theta = \mu$	1	$1{\times}\sigma^2 = \sigma^2$

3.5 The systematic part of the model

The second part of the generalized linear model, the linear predictor, is given in as $\eta = \sum_{j=1}^{p} \beta_j x_j$, where x_j is the jth explanatory variable (with $x_1 = 1$ for the intercept). Now, for each observation y_i, $i = 1, \dots, n$, the explanatory variables may differ. To make explicit this dependence on i, we write:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \tag{3.11}$$

where x_{ij} is the value of the jth explanatory variable on individual i (with $x_{i1} = 1$). Rewriting this in matrix notation:

$$\eta = X\beta, \tag{3.12}$$

where now $\eta = (\eta_1, \dots, \eta_n)$ is a vector of linear predictor variables, $\beta = (\beta_1, \dots, \beta_p)$ is a vector of regression parameters, and X is an $n \times p$ design matrix.

Recall from that we are concerned with two kinds of explanatory variable:

Quantitative — for example, $x_j \in (-\infty, \infty)$ etc.

Qualitative — for example, $x_j \in \{A, B, C\}$ etc.

As discussed in , each quantitative variable is represented in X by an $n \times 1$ column vector. Each qualitative variable, with $k + 1$ levels, say, is represented by a dummy $n \times k$ matrix of 0's and 1's (one column, usually the first, being dropped to avoid identification problems).

3.6 The link function

On we saw that the contribution of randomness to an observation y might be described with a member of the exponential family. We also saw that the systematic part of y might be described using a linear predictor η of the explanatory variables. In we introduced the notion of a link function $\eta = g(\mu)$ to link these two parts together, where μ is the mean of y.

Rarely, the choice of link function $g(\mu)$ is motivated by theory underlying the data at hand. For example, in a dose–response setting, the appropriate model could possibly be motivated by the solution to a set of partial differential equations describing the flow through the body of a dose of a drug.

When there is no compelling underlying substantive theory, we typically choose a link function that will transform a restricted range of the dependent variable onto the whole real line. For example, when observations are measurements they are typically positive, so we have $\mu > 0$ and might choose the logarithmic link:

$$g(\mu) = \log(\mu). \tag{3.13}$$

When observations are binomial counts from $B(m, p)$, $0 < p < 1$, with mean $\mu = mp$, we might choose the *logit* link from

$$\eta = g(\mu) = \operatorname{logit}(\mu/m) = \operatorname{logit}(p) = \log\{p/(1-p)\} \tag{3.14}$$

or the *probit* link which is the inverse of the cumulative distribution function of the $N(0, 1)$ distribution:

$$\eta = g(\mu) = \Phi^{-1}(\mu/m) = \Phi^{-1}(p), \tag{3.15}$$

or the *complementary log-log (cloglog)* link:

$$\eta = g(\mu) = \log(-\log(1-\mu/m)) = \log(-\log(1-p)), \tag{3.16}$$

or the *cauchit* link which is the inverse of the cumulative distribution function of the Cauchy (t_1) distribution:

$$\eta = g(\mu) = \tan(\pi(\mu/m - \tfrac{1}{2})) = \tan(\pi(p - \tfrac{1}{2})). \tag{3.17}$$

shows these link functions for proportions fitted to the beetle mortality data. This demonstrates that the logit and probit links are very similar, that the complementary log-log link fits these data slightly better in the extremes, but that the cauchit link fits these data quite poorly in the extremes.

3.7 The canonical links

A mathematically and computationally convenient choice of link function $g(\mu)$ can be constructed by setting:

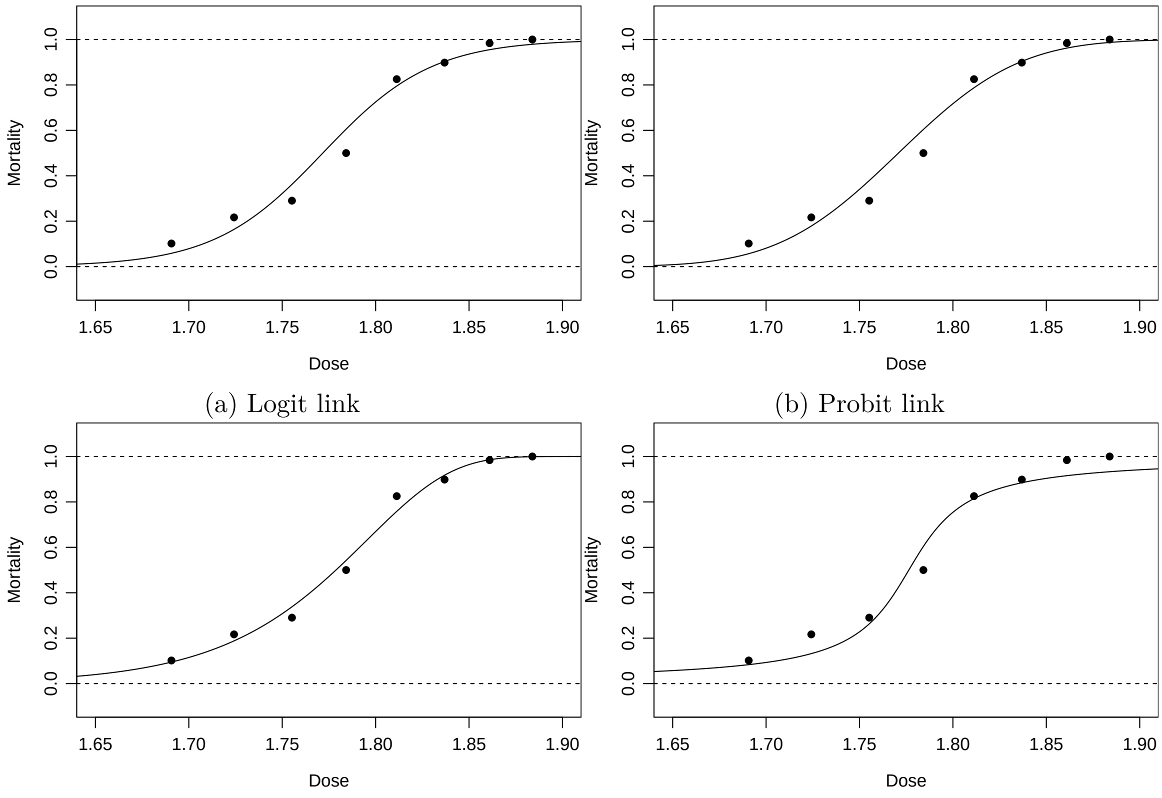
$$\theta = \eta, \tag{3.18}$$

where θ is the canonical parameter of the exponential family as defined in Equation 3.3. Then, Equation 3.8 shows that the mean μ is a function of θ. Therefore, Equation 3.18 indirectly provides a link between μ and η. That is, Equation 3.18 implicitly defines a link function $\eta = g(\mu)$.

What is the form of this $g(\cdot)$?

From Equation 3.8,

$$\mu = b'(\theta).$$



(a) Logit link

(b) Probit link

(c) cloglog link

(d) Cauchit link

Figure 3.1: Dose–response curves fitted to the beetle mortality data from Table 1.1 with different choices of link function.

So, provided function $b'(\cdot)$ has an inverse $(b')^{-1}(\cdot)$, we may write

$$\theta = (b')^{-1}(\mu). \tag{3.19}$$

Now, from Equation 3.5, $g(\mu) = \eta$, so using Equation 3.18:

$$g(\mu) = \theta = (b')^{-1}(\mu), \tag{3.20}$$

from Equation 3.19. This makes explicit the $g(\mu)$ that is implicitly asserted by Equation 3.18. Equation 3.20 is called the *canonical* link function.

Proposition 3.2. *For the canonical link function,*

$$g'(\mu) = 1/b''(\theta).$$

Proof: From Proposition 3.1, $\mu = E[Y] = b'(\theta)$, so

$$\frac{\mathrm{d}\mu}{\mathrm{d}\theta} = b''(\theta).$$

From Equation 3.20, for the canonical link function, we have $\theta = g(\mu)$, so

$$\frac{\mathrm{d}\theta}{\mathrm{d}\mu} = g'(\mu).$$

Now $\mathrm{d}\theta/\mathrm{d}\mu = (\mathrm{d}\mu/\mathrm{d}\theta)^{-1}$ and hence

$$g'(\mu) = 1/b''(\theta).$$

Which proves the proposition.

3.7.1 Example: canonical link function for Poisson distribution

For the Poisson distribution $\mathrm{Po}(\lambda)$, we have from Table 3.2 that $b(\theta) = e^{\theta}$. Therefore,

$$b'(\theta) = e^{\theta},$$

so the inverse of function $b'(\cdot)$ exists and is the inverse of the exponential function, which is the logarithmic function. Then, applying Equation 3.20

$$g(\mu) = \log(\mu) \tag{3.21}$$

Thus the canonical link for the Poisson distribution is log.

3.7.2 Example: canonical link function for Normal distribution

For the Normal distribution $N(\mu, \sigma^2)$, we have from Table 3.4 that $b(\theta) = \theta^2/2$. Therefore

$$b'(\theta) = \theta$$

so the inverse of function $b'(\cdot)$ exists and is the inverse of the identity function, which is the identity function. (The identity function is that which maps a value onto itself.) Then, applying Equation 3.20,

$$g(\mu) = \mu.$$

Thus the canonical link for the Normal distribution is the identity function.

3.7.3 Range of canonical link functions

For many models, μ has a restricted range, but we would like η to have unlimited range. It turns out, for several members of the exponential family, that the canonical link function provides η with unlimited range. However, Table Table 3.6 shows that this is not always so.

Table 3.6: Canonical link functions and their ranges (see McCullagh and Nelder, 2nd Edn., p291 with †binomial distribution with index m and mean μ and ‡gamma distribution with mean μ (see Exercises for details).

$f(y)$	Range of μ	$b(\theta)$	$\mu = b'(\theta)$	Canonical link, $g(\mu)$	Range of η
Normal	$(-\infty, \infty)$	$\frac{1}{2}\theta^2$	θ	μ	$(-\infty, \infty)$
Poisson	$(0, \infty)$	e^θ	e^θ	$\log \mu$	$(-\infty, \infty)$
Binomial†	$(0, m)$	$m \log(1 - e^\theta)$	$m/(1 + e^{-\theta})$	$\text{logit}(\mu/m)$	$(-\infty, \infty)$
Gamma‡	$(0, \infty)$	$-\log(-\theta)$	$-\theta^{-1}$	$-\mu^{-1}$	$(-\infty, 0)$

3.7.4 Convenience of the canonical link function

Why is the canonical link function Equation 3.20 convenient? The assertion Equation 3.18 means that, in the exponential-family formula Equation 3.3, we can simply substitute the linear predictor

$$\eta = \sum_j \beta_j x_j$$

from Equation 3.4 in place of θ, to give:

$$f(y; \mathbf{x}, \beta, \phi) = \exp\left\{ \frac{y\left[\sum_j \beta_j x_j\right] - b\left(\left[\sum_j \beta_j x_j\right]\right)}{\phi} + c(y, \phi) \right\}, \tag{3.22}$$

where $\mathbf{x} = \{x_j, j = 1, \dots, p\}$ and $\beta = \{\beta_j, j = 1, \dots, p\}$.

Suppose we have n independent observations, $\{y_i,\ i = 1, \dots, n\}$. As discussed in Section~**??**, the explanatory variables $(x_1, \dots, x_p)$ will depend on~i, and so η will also depend on~i. Therefore, we attach subscript i to y and to each x_j, giving:

$$f(y_i; \{x_{ij}\}, \{\beta_j\}, \phi) = \exp\left\{ \frac{y_i \left[\sum_j \beta_j x_{ij}\right] - b\left(\left[\sum_j \beta_j x_{ij}\right]\right)}{\phi} + c(y_i, \phi) \right\}. \tag{3.23}$$

By independence, the joint distribution of all observations $\{y_i\} = \{y_i,\ i = 1, \dots, n\}$ is:

$$f(\mathbf{y}; X, \beta, \phi) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi),$$

so

$$\log f(\mathbf{y}; X, \beta, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi)$$

then substituting in

$$\log f(\mathbf{y}; X, \beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \left[\sum_j \beta_j x_{ij}\right] - b\left(\left[\sum_j \beta_j x_{ij}\right]\right)}{\phi} + c(y_i, \phi) \right\}$$

and finally simplifying to give

$$\log f(\mathbf{y}; X, \beta, \phi) = \frac{\sum_j \beta_j S_j - \sum_i b\left(\left[\sum_j \beta_j x_{ij}\right]\right)}{\phi} + \sum_i c(y_i, \phi) \tag{3.24}$$

where

$$S_j = \sum_{i=1}^{n} y_i x_{ij}.$$

Thus, in the log-likelihood Equation 3.24, it is only the first term that involves both the observations $\mathbf{y} = \{y_i,\ i = 1, \dots, n\}$ and the parameters $\beta = \{\beta_j, j = 1, \dots, p\}$, and this term depends on the observations only through the statistics $\mathbf{S} = \{S_j, j = 1, \dots, p\}$. These are called *sufficient statistics*, and their appearance in Equation 3.24 confers both theoretical and practical advantages.

3.8 Maximum likelihood estimation for generalized linear models

Throughout this module we use the principle of maximum likelihood estimation (MLE) to estimate regression parameters.

3.8.1 The i.i.d. case

Suppose we have n i.i.d. observations $\{y_i,\ i = 1, \dots, n\}$, where each y_i is sampled from the same exponential family density Equation 3.3

$$f(y_i; \theta, \phi) = \exp\left\{\frac{\theta y_i - b(\theta)}{\phi} + c(y_i, \phi)\right\}. \tag{3.25}$$

For simplicity we assume the canonical parameter θ does not depend on i. Later, we will consider the case where θ depends on i through covariates $\{x_{ij},\ j = 1, \dots, p\}$, as in Section Section 3.7.4.

By independence, the joint distribution of all the observations $\{y_i\} = \{y_i,\ i = 1, \dots, n\}$ is:

$$f(\{y_i\}; \theta, \phi) = \prod_{i=1}^{n} f(y_i; \theta, \phi).$$

So

$$\log f(\{y_i\}; \theta, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta, \phi) = \sum_{i=1}^{n} \left[\frac{\theta y_i - b(\theta)}{\phi} + c(y_i, \phi)\right].$$

Regarding the observations $\{y_i\}$ as constants (which they are, once we have them) and the scale parameter ϕ as a fixed *nuisance* parameter (whose value we may not know), the log-likelihood as a function of the parameter θ of interest is:

$$l(\theta; \{y_i\}, \phi) = n\,\frac{\theta \bar{y} - b(\theta)}{\phi} + \text{constant}, \tag{3.26}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

We estimate θ by maximizing the log likelihood – i.e. given the data $\{y_i,\ i = 1, \dots, n\}$, we estimate the value of θ to be that value for which the likelihood, and hence the log-likelihood, is greatest.

We maximize the log-likelihood by differentiating it and setting it to zero:

$$\frac{dl(\theta; \{y_i\}, \phi)}{d\theta} = n\,\frac{\bar{y} - b'(\theta)}{\phi}$$

and hence the MLE for θ, which we denote $\hat{\theta}$, satisfies

$$b'(\hat{\theta}) = \bar{y}. \tag{3.27}$$

Now, we showed in Proposition 3.1 that

$$\mathrm{E}[Y] = \mu = b'(\theta).$$

Let $\hat{\mu}$ denote the MLE of μ. Then

$$\hat{\mu} = b'(\hat{\theta}),$$

because the MLE of any function $\zeta = h(\theta)$ of the parameters is $\hat{\zeta} = h(\hat{\theta})$. Therefore, we have

$$\hat{\mu} = \bar{y}. \tag{3.28}$$

So we find that $\hat{\theta}$ is the value of θ for which the theoretical mean $\hat{\mu} = b'(\hat{\theta})$ matches the sample mean $\bar{y}$.

3.8.2 Accuracy of MLEs in the i.i.d. case

For our i.i.d. sample $\{y_i,\ i = 1, \dots, n\}$, we have $b'(\hat{\theta}) = \hat{\mu} = \bar{y}$. Let θ_0 be the true value of θ with corresponding mean μ_0, i.e.

$$b'(\theta_0) = \mu_0. \tag{3.29}$$

How accurate is $\hat{\theta}$? We know that

$$\mathrm{E}[\bar{Y}] = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(y_i) = \mu_0 = b'(\theta_0), \tag{3.30}$$

using Equation 3.29, and

$$\mathrm{Var}[\bar{Y}] = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(y_i)$$

because the observations are independent,

$$= \frac{1}{n}\, b''(\theta_0)\phi, \tag{3.31}$$

using the result Equation 3.8.

We can use Taylor's theorem to expand $b'(\hat{\theta})$ about θ_0:

$$\bar{y} = b'(\hat{\theta}) \approx b'(\theta_0) + (\hat{\theta} - \theta_0) b''(\theta_0),$$

which implies that

$$(\hat{\theta} - \theta_0) \approx b''(\theta_0)^{-1}\{b'(\hat{\theta}) - b'(\theta_0)\} = b''(\theta_0)^{-1}(\bar{y} - \mu_0), \tag{3.32}$$

using Equation 3.27 and Equation 3.29. We can use Equation 3.32 to get approximations to the mean and variance of $\hat{\theta}$:}

$$\mathrm{E}[\hat{\theta} - \theta_0] \approx b''(\theta_0)^{-1}\mathrm{E}(\bar{y} - \mu_0) = 0,$$

using Equation 3.30, so

$$\mathrm{E}(\hat{\theta}) \approx \theta_0, \tag{3.33}$$

and

$$\mathrm{Var}(\hat{\theta}) \approx \mathrm{E}\left[(\hat{\theta} - \theta_0)^2\right]$$

using Equation 3.33,

$$\mathrm{Var}(\hat{\theta}) \approx \mathrm{E}\left[\left(b''(\theta_0)^{-1}(\bar{y} - \mu_0)\right)^2\right]$$

using Equation 3.32,

$$\mathrm{Var}(\hat{\theta}) \approx \left(b''(\theta_0)\right)^{-2}\mathrm{Var}[(\bar{Y})]$$

using Equation 3.30,

$$\mathrm{Var}(\hat{\theta}) = \frac{\phi}{n\, b''(\theta_0)} \tag{3.34}$$

using Equation 3.31.

Thus we see that the first two derivatives of $b(\theta)$ play a key role in inference.

3.9 Exercises

3.1 In the binomial distribution, show that $-m\log(1-p) = m\log(1+e^{\theta})$ where $\theta = \text{logit } p$.

3.2 Use the results in Section 3.4 and the exponential family description of the Binomial distribution in Section 3.3.2 to show that the mean and variance of a $\text{Bin}(m, p)$ are mp and $mp(1-p)$.

Hint: $f'(g(x)) = f'(g(x))g'(x)$

4 GLM fitting

4.1 Maximum likelihood estimation for GLMs

4.2 Deviance

4.3 Residuals

4.4 Fitting generalized linear models in R

5 Logistic regression model

5.1 Overview

temporary

5.2 Application: dose–response experiments

5.3 Residuals and deviance

6 Loglinear models

6.1 Overview

6.2 General log-linear model

6.3 Margins of a contingency table

6.4 Conditioning on marginal totals

6.5 Maximum likelihood estimates

6.6 Analysis of Melanoma data

7 Extensions to Loglinear models

7.1 Overview

7.2 Multi-way contingency tables

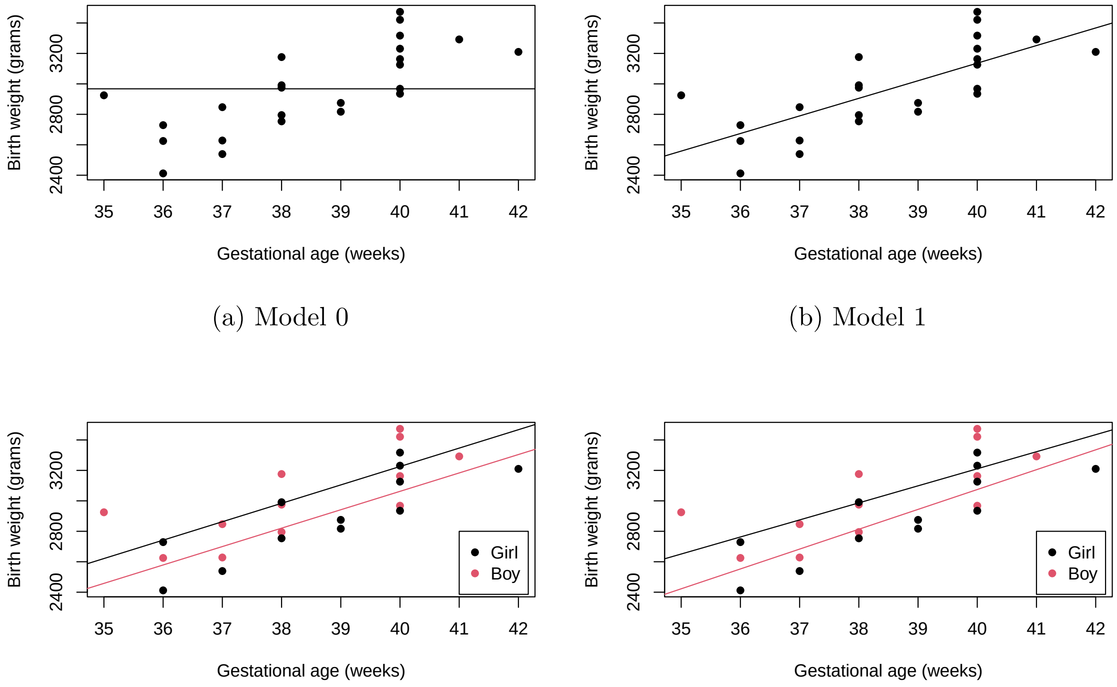
7.3 Product-multinomial models

8 Summary

9 Appendix: Background to analysis of variance

9.1 Analysis of variance

Consider the four models fitted to the birth weight data. Figure 9.1 shows the data set along with the corresponding fitted model as a single line, for the models which do not take `Sex` into account, and two lines, for the models which include `Sex`.



(a) Model 0

(b) Model 1

(c) Model 2

(d) Model 3

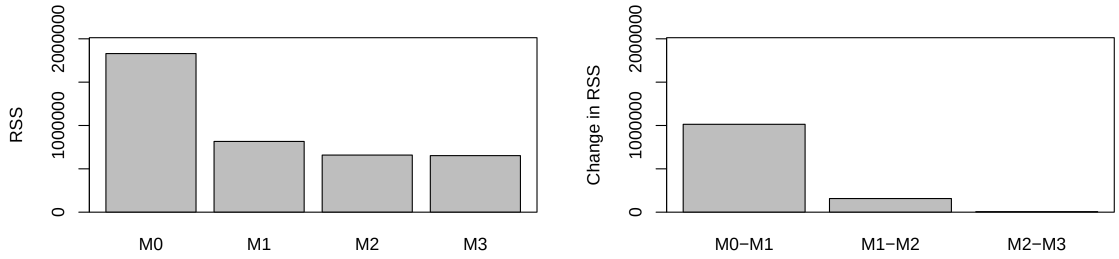
Figure 9.1: Birthweight and gestational age data with superimposed fitted regression lines from various competing models.

The *residual sum of squares* (RSS) takes into account the vertical distance between the fitted model and the data values. Let R_k denote the residual sum of squares for Model $k : R_k = \sum_{i=1}^{n}(y_i - \widehat{\mu}_{ki})^2$, where $\widehat{\mu}_{ki}$ is the fitted value for individual i under `Model` k and let r_k denote the corresponding *residual degrees of freedom* for `Model` k (the number of observations minus the number of model parameters). The table below shows these values for the four models fitted to the data.

Table 9.1: Summary of the residual sums of squares

`Model` k	R_k	r_k	$R_k - R_{k-1}$	$r_k - r_{k-1}$
0	1829873	23		
1	816074	22	1013799	1
2	658771	21	157304	1
3	652425	20	6346	1

The table also shows the change in residual sums of squares, $R_k - R_{k-1}$, which measures the improvement in the fit due to the extra parameters used in Model k compared to Model $k-1$. The RSS and changes in RSS values are also shown in Figure 9.2. It is clear that there is a substantial reduction in RSS moving from Model 0 to Model 1, but small reductions as further parameters are added to the model. We might guess that Model 1 will be the “best’ ’ model, but a it is not acceptable to base a choice on our personal subjective opinion but instead a sequence of hypothesis tests will be used.



(a) Residual sum of squares

(b) Change in RSS

Figure 9.2: Birthweight and gestational age data with superimposed fitted regression lines from various competing models.

Here, a sequence of three hypothesis tests is considered: Starting with

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

Which can be judged by comparing $R_1 - R_0 = 1013799$ which follows a $\sigma^2\chi^2$ distribution on $r_1 - r_0 = 1$ degrees of freedom ($(R_1 - R_0)/\sigma^2$ follows a χ^2_1distribution) with $R_1 = 816074$ which follows a $\sigma^2\chi^2$ distribution on $r_1 = 22$ degrees of freedom (R_1/σ^2 follows a

χ^2_{22} distribution. Fortunately, taking the ratio eliminates σ^2 giving the test statistics

$$F_{01} = \frac{(R_1 - R_0)/(r_1 - r_0)}{R_1/r_1} = \frac{1013799/1}{816074/22} = 27.33$$

If H_0 is true, then we would expect this to be close to 1. The 5%, 1% and 0.1% critical values for the distribution are 4.3, 7.95, 14.38, and the observed F statistics is much larger than all these and hence p-value < 0.001 meaning we reject H_0 in favour of H_1.

If H_0 had been accepted then the sequence would stops and `Model` 0 declared the best, whereas H_0 is rejected and the next test is considered

$$\text{Test 2} \quad H_0 : \texttt{Model}\ 1 \text{ is true}; H_1 : \texttt{Model}\ 2 \text{ is true.}$$

If H_0 is accepted here then the sequence stops and `Model` 1 is declared the best, whereas is H_0 is rejected then the last test is considered

$$\text{Test 3} \quad H_1 : \texttt{Model}\ 2 \text{ is true}; H_1 : \texttt{Model}\ 3 \text{ is true.}$$

If H_0 is accepted here then the sequence stops and `Model` 2 is declared the best, whereas if H_0 is rejected then `Model` 3 is declared the best.

9.2 Distributions derived from the Gaussian distribution

9.2.1 The Gaussian (normal) distribution

If $X \sim N(\mu, \sigma^2)$ then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, \qquad -\infty < x < \infty.$$

Properties:

1. The parameter $\mu = \mathrm{E}[X]$ is a location parameter and $\sigma^2 = \mathrm{Var}[X]$ is a scale parameter.
2. If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
3. If $X_i \sim N(\mu_i, \sigma_i^2), i = 1, ..., n$ (independent) then $\sum a_i X_i \sim N(\sum a_i \mu, \sum a_i^2 \sigma^2)$.
4. A special case is when $\mu = 0$ and $\sigma^2 = 1$ which is called the *standard normal* distribution.

9.2.2 The χ^2-distribution

If X has a Chi-squared distribution, $X \sim \chi^2_\nu$ then

$$f(x) = \frac{(\frac{1}{2})^{\frac{\nu}{2}} x^{\frac{\nu}{2}-1} e^{-\frac{1}{2}x}}{\Gamma\left(\frac{\nu}{2}\right)}, \qquad x \geq 0, \nu > 0 \text{ and integer.}$$

with $\mathrm{E}[X] = \nu$ and $\mathrm{Var}[X] = 2\nu$.

Properties:

1. The parameter ν is a shape parameter and is called the *degrees of freedom.* The pdf is positive skew, but becomes more symmetric as ν increases.
2. If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2_1$.
3. If $X_i \sim \chi^2_{\nu_i}, i = 1, ..., n$ (independent) then $\sum X_i \sim \chi^2_\nu$, where $\nu = \sum \nu_i$.
4. If $Z_i \sim N(0,1), i = 1, ..., n$ (independent) then $\sum Z_i^2 \sim \chi^2_n$.
5. This is a special case of the gamma distribution, with $\alpha = \nu/2$ and $\lambda = \frac{1}{2}$, that is $\gamma(\frac{\nu}{2}, \frac{1}{2})$.

9.2.3 The t- and F-distributions

If X has a t-distribution, $X \sim t_\nu$ then

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \qquad -\infty < x < \infty,$$

where $\nu > 0$ and integer.

Properties:

1. The parameter ν is called the *degrees of freedom.*
2. If $X \sim N(0,1)$ and $Y \sim \chi^2_\nu$ (independent) then
$$\frac{X}{\sqrt{Y/\nu}} \sim t_\nu.$$
3. If $X \sim t_\nu$ then $X^2 \sim F_{1,\nu}$.
4. $t_\nu \to N(0,1)$ as $\nu \to \infty$.

If X has an F-distribution, $X \sim F_{\nu_1,\nu_2}$ then

$$f(x) = \frac{\nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} x^{\frac{\nu_1}{2}-1}}{B(\frac{\nu_1}{2}, \frac{\nu_2}{2})(\nu_2 + \nu_1 x)^{\frac{\nu_2+\nu_1}{2}}} \qquad x \geq 0.$$

where $\nu_1, \nu_2 > 0$ and integer are know as the *degrees of freedom.*

Properties:

1. The parameters ν_1 and ν_2 are called the degrees of freedom.
2. If $X_1 \sim \chi^2_{\nu_2}$ and $X_2 \sim \chi^2_{\nu_2}$ (independent) then
$$\frac{X_1/\nu_1}{X_2/\nu_2} \sim F_{\nu_1,\nu_2}.$$
3. If $X \sim F_{\nu_1,\nu_2}$ then $1/X \sim F_{\nu_2,\nu_1}$, hence, $Pr(F_{\nu_1,\nu_2} < c) = Pr(F_{\nu_2,\nu_1} > 1/c)$.

10 A: Revision of vectors and matrices

10.1 A.1 Notation

Note that in these notes we use lowercase y or y_i to denote both observed values or random variables, which should be clear from the context.

Similarly, although all vectors are column vectors, for ease of writing, we may write simply as a horizontal list. Again, the meaning will be clear from context.

Examples of scalars:

$$x, y, \alpha, \beta, \gamma, \delta, \epsilon \quad \text{or} \quad x_i, y_i, \alpha_i, \beta_j, \gamma_j, \delta_j, \epsilon_i$$

Examples of $n \times 1$ vectors:

$$\mathbf{y} = (y_i) = (y_1, y_2, \dots, y_n), \mathbf{Y} = (Y_i) = (Y_1, Y_2, \dots, Y_n),$$
$$\epsilon = (\epsilon_i) = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$$

or as:

$$\mathbf{y} = (y_i) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{Y} = (Y_i) = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \epsilon = (\epsilon_i) = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Examples of $p \times 1$ vectors:

$$\mathbf{x} = (x_j) = (x_1, x_2, \dots, x_p), \mathbf{X} = (X_j) = (X_1, X_2, \dots, X_p),$$
$$\beta = (\beta_j) = (\beta_1, \beta_2, \dots, \beta_p)$$

or as

$$\mathbf{x} = (x_j) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \mathbf{X} = (X_j) = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

Examples of $n \times p$ matrices:

$$X = (X_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

10.2 Operations

Examples of addition:

$$\alpha + \beta = \begin{bmatrix} \alpha + \beta_1 \\ \alpha + \beta_2 \\ \vdots \\ \alpha + \beta_p \end{bmatrix} \qquad \mathbf{x} + \beta = \begin{bmatrix} x_1 + \beta_1 \\ x_2 + \beta_2 \\ \vdots \\ x_p + \beta_p \end{bmatrix}$$

Examples of multiplication:

$$\alpha\beta = \alpha \times \beta = \alpha \cdot \beta = \begin{bmatrix} \alpha\beta_1 \\ \alpha\beta_2 \\ \vdots \\ \alpha\beta_p \end{bmatrix} \qquad \mathbf{x} \cdot \beta = \mathbf{x}^T\beta = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$$

$$X\beta = \begin{bmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p \\ x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p \\ \vdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p \end{bmatrix}$$

10.3 Special vectors and matrices

Vector of zero's or one's:

$$\mathbf{0} = (0, 0, \dots, 0), \qquad \mathbf{1} = (1, 1, \dots, 1)$$

Similarly, although all vectors are column vectors, for ease of writing, we may write simply as a horizontal list. Again, the meaning will be clear from context.

Examples of scalars:

$$x, y, \alpha, \beta, \gamma, \delta, \epsilon \quad \text{or} \quad x_i, y_i, \alpha_i, \beta_j, \gamma_j, \delta_j, \epsilon_i$$

Examples of $n \times 1$ vectors:

$$\mathbf{y} = (y_i) = (y_1, y_2, \dots, y_n), \mathbf{Y} = (Y_i) = (Y_1, Y_2, \dots, Y_n),$$

$$\epsilon = (\epsilon_i) = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$$

or as:

$$\mathbf{y} = (y_i) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{Y} = (Y_i) = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \epsilon = (\epsilon_i) = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Examples of $p \times 1$ vectors:

$$\mathbf{x} = (x_j) = (x_1, x_2, \dots, x_p), \mathbf{X} = (X_j) = (X_1, X_2, \dots, X_p),$$
$$\beta = (\beta_j) = (\beta_1, \beta_2, \dots, \beta_p)$$

or as

$$\mathbf{x} = (x_j) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \mathbf{X} = (X_j) = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

Examples of $n \times p$ matrices:

$$X = (X_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

10.4 Operations

Examples of addition:

$$\alpha + \beta = \begin{bmatrix} \alpha + \beta_1 \\ \alpha + \beta_2 \\ \vdots \\ \alpha + \beta_p \end{bmatrix} \qquad \mathbf{x} + \beta = \begin{bmatrix} x_1 + \beta_1 \\ x_2 + \beta_2 \\ \vdots \\ x_p + \beta_p \end{bmatrix}$$

Examples of multiplication:

$$\alpha\beta = \alpha \times \beta = \alpha \cdot \beta = \begin{bmatrix} \alpha\beta_1 \\ \alpha\beta_2 \\ \vdots \\ \alpha\beta_p \end{bmatrix} \qquad \mathbf{x} \cdot \beta = \mathbf{x}^T\beta = x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p$$

$$X\beta = \begin{bmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1p}\beta_p \\ x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2p}\beta_p \\ \vdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \dots + x_{np}\beta_p \end{bmatrix}$$

10.5 Special vectors and matrices

Vector of zero's or one's:

$$\mathbf{0} = (0, 0, \dots, 0), \qquad \mathbf{1} = (1, 1, \dots, 1)$$

11 Appendix: Standard distributions

11.1 Notation

11.2 Basic rules in probability & statistics

- **Rules of expectation** Expectation is linear so for random variables X_1 and X_2, and constants a, b and c, the

  $$\mathrm{E}[aX_1 + bX_2 + c] = a\mathrm{E}[X_1] + b\mathrm{E}[X_2] + c.$$

  There is no change to this rule when X_1 and X_2 are independent.

- **Rules of variance** By definition, the $\mathrm{Var}[X]$ is given by

  $$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X]^2 - \mathrm{E}[X^2]$$

  with the variance of the sum of two random variables

  $$\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] - 2\mathrm{Cov}[X_1, X_2]$$

  where $\mathrm{Cov}[X_1, X_2]$ is the covariance which is defined in the following section. If, however, X_1 and X_2 are independent, then the above reduces to

  $$\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2].$$

- **Definition of covariance and correlation** By definition, the $\mathrm{Cov}[X_1, X_2]$ is given by

  $$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\mathrm{E}[X_2]$$

  and

  $$\mathrm{Cov}[X_1, X_2] = \frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{Var}[X_1]\mathrm{Var}[X_2]}}.$$

  In the case when X_1 and X_2 are independent random variables then $\mathrm{E}[X_1 X_2] = \mathrm{E}[X_1]\mathrm{E}[X_2]$ and so $\mathrm{Cov}[X_1, X_2] = 0$ and hence $\mathrm{Cov}[X_1, X_2] = 0$ also. This gives justification of the variance formula for the sum of independent random variables.

- **Joint distribution for independent events**. Suppose that X_1 and X_2 are a pair of independent on identically distributed random variables then their joint probability function is given by

  $$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2).$$

11.3 Standard distributions

Here, $f(x;\theta) \equiv f(x|\theta)$ is written to denote a probability function if X is discrete, or a probability density function if X is continuous, where θ is the parameter of the distribution.

- **Bernoulli**, with parameter θ $(0 < \theta < 1)$. A discrete random variable X with probability function

$$f(x;\theta) \quad = \quad \theta^x(1-\theta)^{1-x} \qquad\qquad x = 0, 1$$

  $\mathrm{E}[X] \ = \ \theta$ and $\mathrm{Var}[X] = \theta(1-\theta)$.

- **Geometric**, with parameter θ $(0 < \theta < 1)$. A discrete random variable \$X\$ with probability function

$$f(x;\theta) = \theta(1-\theta)^{x-1} \qquad\qquad x = 1, 2, ...$$

  $\mathrm{E}[X] = 1/\theta$ and $\mathrm{Var}[X] = (1-\theta)/\theta^2$.

- **Binomial**, with parameters n and θ (where n is a known positive integer and $0 < \theta < 1$). A discrete random variable X with probability function

$$f(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad\qquad x = 0, 1, \ldots, n$$

  $\mathrm{E}[X] = n\theta$ and $\mathrm{Var}[X] = n\theta(1-\theta)$.

- **Poisson**, with parameter θ $(\theta > 0)$. A discrete random variable X with probability function

$$f(x;\theta) \quad = \quad \frac{\theta^x e^{-\theta}}{x!} \qquad\qquad x = 0, 1, ...$$

  $\mathrm{E}[X] \ = \ \theta$ and $\mathrm{Var}[X] \ = \ \theta$.

- **Negative Binomial**, with parameters r and θ $(r > 0$ and $0 < \theta < 1)$. A discrete random variable X with probability function

$$f(x;r,\ \theta) \quad = \quad \binom{x-1}{r-1}\theta^r(1-\theta)^{x-r} \qquad\qquad x = r, r+1, ...$$

  $\mathrm{E}[X] \ = \ \frac{r}{\theta}$ and $\mathrm{Var}[X] \ = \ \frac{r(1-\theta)}{\theta^2}$.

- **Uniform**, with parameter θ $(\theta > 0)$. A continuous random variable X, with probability density function

$$f(x;\theta) \quad = \quad \frac{1}{\theta} \qquad\qquad 0 < x < \theta$$

  $\mathrm{E}[X] \ = \ \frac{\theta}{2}$ and $\mathrm{Var}[X] \ = \ \frac{\theta^2}{12}$.

- **Exponential**, with parameter λ ($\lambda > 0$).

  A continuous random variable X, with probability density function

  $$f(x;\lambda) \;=\; \lambda e^{-\lambda x} \qquad\qquad x > 0$$

  $\mathrm{E}[X] \;=\; \dfrac{1}{\lambda}$ and $\mathrm{Var}[X] \;=\; \dfrac{1}{\lambda^2}$.

- **Gamma** with parameters α and β ($\alpha, \beta > 0$). A continuous random variable X, with probability density function

  $$f(x;\alpha,\beta) \;=\; \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \qquad\qquad x > 0$$

  $\mathrm{E}[X] \;=\; \dfrac{\alpha}{\beta}$ and $\mathrm{Var}[X] \;=\; \dfrac{\alpha}{\beta^2}$.

  Note: $\Gamma(\alpha+1) \;=\; \alpha\Gamma(\alpha) \;\; \forall\alpha$ and $\qquad\qquad \Gamma(\alpha+1) \;=\; \alpha!$ for integers $\alpha > 1$.

  BEWARE some authors replace β by $1/\beta$ but still calling it Gamma(α, β). You always need to be clear which parameterization is being used.

- **Beta** with parameters α and β ($\alpha, \beta > 0$). A continuous random variable X, with probability density function

  $$f(x;\alpha,\beta) \;=\; \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad\qquad 0 < x < 1$$

  where

  $$B(\alpha,\beta) \;=\; \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \;=\; \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

  $\mathrm{E}[X] \;=\; \dfrac{\alpha}{\alpha+\beta}$ and $\mathrm{Var}[X] \;=\; \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

  Also referred to as the Beta(α, β) distribution.

- **Normal** with parameters μ and σ^2 ($-\infty < \mu < \infty$ and $\sigma^2 > 0$). A continuous random variable X, with probability density function

  $$f(x;\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi}\;\sigma}\;\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad\qquad -\infty < x < \infty$$

  $\mathrm{E}[X] \;=\; \mu$ and $\mathrm{Var}[X] \;=\; \sigma^2$.

  Also referred to as the N(μ, σ^2) distribution.

- **Pareto** with parameters θ and α ($\theta, \alpha > 0$). A continuous random variable X, with probability density function

$$f(x; \theta, \alpha) \quad = \quad \frac{\alpha\theta^\alpha}{x^{\alpha+1}} \qquad\qquad x > \theta$$

$$\mathrm{E}[X] \ = \ \frac{\alpha\theta}{(\alpha-1)} \ \text{ and } \ \mathrm{Var}[X] \ = \ \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} \quad (\alpha > 2).$$

- **Chi-Square** with parameter n (n is a positive integer). A continuous random variable X, with probability density function

$$f(x; n) \quad = \quad \left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{\Gamma(\frac{n}{2})} \qquad\qquad x > 0$$

$$\mathrm{E}[X] \ = \ n \ \text{ and } \ \mathrm{Var}[X] \ = \ 2n.$$

  Also referred to as the χ^2_n distribution, with n degrees of freedom.

  Note: A χ^2_n distribution is also a Gamma$(\frac{n}{2}, \frac{1}{2})$ distribution.

- **Student's t** with parameter n (n is a positive integer). A continuous random variable X, with probability density function

$$f(x; n) \quad = \quad \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \ \ \Gamma(\frac{n}{2}) \left[1 + \frac{x^2}{n}\right]^{\frac{n+1}{2}}} \qquad\qquad -\infty < x < \infty$$

$$\mathrm{E}[X] \ = \ 0 \ \ (n > 1) \ \text{ and } \ \mathrm{Var}[X] \ = \ \frac{n}{n-2} \quad (n > 2).$$

  Also referred to as the t distribution, with n degrees of freedom.

- **F** with parameters m and n (m, n are positive integers). A continuous random variable X, with probability density function

$$f(x; m, n) \quad = \quad \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma(\frac{m+n}{2}) \ x^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}) \ \left[1 + \frac{mx}{n}\right]^{\frac{m+n}{2}}} \qquad\qquad x > 0$$

$$\mathrm{E}[X] \ = \ \frac{n}{n-2} \ \ (n > 2) \ \text{ and } \ \mathrm{Var}[X] \ = \ \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (n > 4).$$

  Also referred to as the F distribution, with m and n degrees of freedom.

  Note: The order of m and n is important.

References