MODELLING FORCES OF INFECTION FOR MUMPS, RUBELLA AND PARVOVIRUS: A STATISTICAL PERSPECTIVE

Emanuele Del Fava

April 22, 2007

Ai miei genitori

Abstract

The force of infection is a fundamental epidemiological parameter of infectious diseases. For many infectious diseases it is assumed that the force of infection is age-dependant. Although the force of infection can be estimated directly from a follow up study, it is much more common to have cross-sectional seroprevalence data from which the seroprevalence and the force of infection can be estimated. Here we propose to model the seroprevalence with four different parametric models: a nonlinear least squares model proposed by Farrington (1990); a logistic model, estimated using the generalized linear models; two fractional polynomial models of different order. We illustrate the methods using three seroprevalence samples, taken by the literature, regarding the following infectious diseases: mumps, rubella and parvovirus.

Besides, in order to determine the optimal sample size for a serological survey, we show the serious problems of the standard confidence interval for a binomial proportion and we introduce some alternative confidence intervals proposed by the literature.

Contents

List of Tables

List of Figures

Chapter 1

Serological surveys

1.1 What is a serological survey?

The aim of a serological survey is to find out the presence of antibodies, produced by the organism in response to a specific antigen, responsible for the disease the researcher is studying. The organism can produce two kinds of antibodies:

1. *IgM*: these antibodies are produced at the very beginning of the infection, but they are present in the blood serum for a very short time;
2. *IgG*: these antibodies are produced later than IgM, but they remain in the organism for a very long period, even after the disease is disappeared.

So, if a person presents in its own blood some antibodies against a specific infection, it means that:

- the person has experimented the infection before or during the survey;
- otherwise, the person is vaccinated against the infection.

With a serological survey, the researcher wants to know the percentage of people, belonging to a certain cohort, who present antibodies against a specific disease. Together with this percentage, it is interesting to know some characteristics of the people object of the survey, in order to study the relationship between the presence of antibodies and these characteristics, called *variables*. Usually, the cohort is based on the *age* of people and the variables of interest vary from study to study, depending on the aim of the research.

1.2 Follow up and cross-section studies

There are two types of serological surveys.

1.2.1 *Follow up study*

In this study, the researcher takes a sample of *seronegative* people (people who do not present any antibodies against the infection) belonging to the same age-cohort and then he follows this cohort for a certain period. During this period, he notices if a person gets the infection (and so its organism begins to produce antibodies) or not: this is the "period *at risk*", because during this time the subject risks to acquire the infection.

At the end of the survey the researcher knows which people in the sample are *seropositive* (people with antibodies against the infection) or not, and the age at the infection for seropositive people. In this way, he can evaluate the percentage of people with antibodies at every age.

With a follow up study, it is possible to estimate two important parameters of great interest for the epidemiologist:

- the *prevalence*, which is a frequence and measures the percentage of people with antibodies at a specific moment in a population;
- the *force of infection*, which is a risk rate (or incidence rate) and measures the number of new infections during a certain period in the population.

The follow up study has the advantage of reducing "systematic errors", because the researcher can control personally the quality of data during the survey; however, the great disadvatage of this study is that we need a lot amount of time to complete the survey, even the entire life of people, because we have to follow the appearance of the events.

1.2.2 *Cross-section study*

In this study, the researcher takes a sample of people, stratified for the age of subjects, and notices how these individuals are distributed at a specific moment (the time of survey), at every age, between seropositive and seronegative.

The cross-section study has the advantage of requiring much less time than a follow up study, because at the beginning of the study the events of interest have already happened and so the researcher has only to record these events. Besides, from these studies it is

possible to estimate directly the seroprevalence, but not the force of infection.

Of course, cross-section studies are much more common than follow up studies and so, although the force of infection and the prevalence can be estimated directly from a follow up study, it is normal to estimate these parameters from cross-sectional *seroprevalence* data (data about the prevalence of antibodies in the blood serum of the individual).

If our data are obtained from a cross-section study and we want to estimate the force of infection, we have to assume that our data are representative of longitudinal changes in seroprevalence with age.

1.3 The catalytic model

The key quantity governing the trasmission of infection within a given population is the force of infection. This is defined as the instantaneous per capita rate at which *susceptibles* (people that have not contracted the infection yet) acquire infection. It reflects the degree of contact with potential for transmission of infection between susceptibles in the population. Since contact is age dependent, typically higher in children than infants or adults, the force of infection is itself a function of age. Besides, the force of infection, which is an incidence rate as mortality, also depends on calendar time and so the acquisition of an infection could be represented on a Lexis diagram with calendar time on the horizontal axis and age on the vertical axis [4].

These notions may be formalized using a set of differential equations, which aim is to describe the flow of individuals from the healthy stage to a disease stage. Letting $P(a,t)$ denote the probability that an individual, susceptible at birth, remains susceptible at age a and calendar time t and denoting the force of infection by $\ell(a,t)$ and the age-specific death rate by $m(a,t)$, we have:

$$\ell(a,t) + m(a,t) = -\frac{1}{P(a,t)}\left[\frac{\partial}{\partial a}P(a,t) + \frac{\partial}{\partial t}P(a,t)\right]. \tag{1.1}$$

We now have to make some assumptions:

- we assume *time homogeneity*, so the force of infection and the fraction of susceptible individuals only depend on age and not on calendar time, $\frac{\partial}{\partial t}P(a,t) = 0$ and $\frac{\partial}{\partial t}\ell(a,t) = 0$ (although this assumption seems very crude in most practical situations, it provides a convenient starting point for an exposition of statistical theory and it is reasonable if the disease is in a stedy state);

- the disease is irreversible, meaning that the immunity is assumed to be lifelong;
- the mortality caused by the infection is negligible and can be ignored;
- the natural death rate is zero up to the life expectancy and thereafter infinity;
- the population considered is in dynamic equilibrium;
- the disease is in a steady state.

So, Eq. 1.1 can be rewritten in the following form:

$$\ell(a) = -\frac{1}{P(a)}\frac{\partial P(a)}{\partial a}. \tag{1.2}$$

If $F(a)$ denotes the cumulative distribution function of age at infection, we have:

$$\ell(a) = \frac{1}{1 - F(a)}\frac{\partial F(a)}{\partial a} \tag{1.3}$$

and the general solution of this differential equation is the following:

$$F(a) = 1 - \exp\left\{-\int_0^a \ell(s)ds\right\}. \tag{1.4}$$

The seroprevalence, or simply "prevalence", is given by $1 - P(a)$, but from Eq. 1.2 and Eq. 1.3, we have that $F(a) = 1 - P(a)$, so $F(a)$ is also the prevalence, that is to say the probability that an individual at age a has already been infected.

1.3.1 The catalytic model as a risk model

Eq. 1.2 and Eq. 1.3 specify a so-called *catalytic model*, first defined by Muench [5]. The catalytic model is fundamentally a *risk model* and so it can be modelled by the typical functions of a risk model. For an introduction to risk models, see Yamaguchi [6].

The Probability Density Function

Given that T represents the timing of occurence of the event "infection", that is to say the age at infection, the probability density function (pdf) $f(a)$ expresses the unconditional instantaneous probability of having the event:

$$f(a) = \lim_{a \to 0} \frac{Pr(a < T \leq a + \Delta a)}{\Delta a}. \tag{1.5}$$

The Survivor Function

The survivor function $P(a)$ expresses the probability of not having the event prior to age a:

$$P(a) = Pr(T \geq a); \tag{1.6}$$

besides, the survivor function is the cumulative distribution function of the pdf $f(a)$:

$$P(a) = \int_a^{\infty} f(s)ds, \tag{1.7}$$

and so, inverting Eq. 1.7, we have the following equation for the pdf $f(a)$:

$$f(a) = -\frac{\partial P(a)}{\partial a}. \tag{1.8}$$

The Hazard Function

The hazard function $\ell(a)$ describes the istantaneous risk of having the infection at age a, given that the infection did not occur before age a:

$$\ell(a) = \lim_{a \to 0} \frac{Pr(a < T \leq a + \Delta a | T \geq a)}{\Delta a}; \tag{1.9}$$

if we consider Eq. 1.5 and Eq. 1.6, we can see that the hazard function is given by the ratio between the pdf and the survivor function:

$$\ell(a) = \frac{f(a)}{P(a)}; \tag{1.10}$$

then, if we consider Eq. 1.8 and rewrite the hazard function, we obtain Eq. 1.2:

$$\ell(a) = -\frac{\partial P(a)}{\partial a}\frac{1}{P(a)}. \tag{1.11}$$

In addition, the hazard function has also the following form:

$$\ell(a) = -\frac{\partial \log P(a)}{\partial a}, \tag{1.12}$$

by the definition of the first derivative of a logarithmic function.

The Cumulative Hazard Function

This is the cumulative hazard function:

$$G(a) = \int_0^a \ell(s)ds. \tag{1.13}$$

Finally, we have a direct relationship between $P(a)$ and $G(a)$. In effect, if we invert Eq. 1.12, after some passages we have that:

$$P(a) = \exp(-G(a)) = \exp\left\{-\int_0^a \ell(s)ds\right\}. \tag{1.14}$$

From Eq. 1.14 and knowing that $F(a) = 1 - P(a)$, we retrieve Eq. 3.12:

$$F(a) = 1 - P(a) = 1 - \exp\left\{-\int_0^a \ell(s)ds\right\}. \tag{1.15}$$

Chapter 2

Basic principles of the sampling for serological surveys

We have previously seen that cross-section studies require less time than follow up ones, because at the time of the survey the events (the acquisition of the disease) are already occured.

2.1 Current Status Data

Data from a cross-section study are often called *current status data*. We have current status data when we want to measure the time of occurence of some event (here, an infection) for a sample of individuals, but all we can obtain is limited to a single observation of whether or not the event has occured for each subject at the time of the survey. The resulting observations are censored with respect to time of event occurence:

1. *left censored*, if the event has occured, but we do not know when;
2. *right censored*, if the event has not occured at the time of the survey yet.

For the ith individual from a sample of size n, the goal of the researcher is to observe the random variables V_i and Z_i, where V_i is the time of event occurence and Z_i is a (1 x p) vector of covariates. The data actually collected consist of observations (y_i, t_i, z_i) of the random variables (Y_i, T_i, Z_i), where T_i is the time that the individual is observed (the time of the survey) and

$$Y_i = \begin{cases} 1 & \text{if } V_i \leq T_i \\ 0 & \text{if } V_i > T_i. \end{cases} \tag{2.1}$$

The information about V_i is limited to the binary indicator Y_i observed at T_i. In terminology of survival analysis, all observations of V_i are either left censored ($Y_i = 1$) or right censored ($Y_i = 0$) at T_i.

In the case of a serological survey, if we assumed the time homogeneity, the only covariate of interest Z_i is the age of the subject: in this case, the researcher do not know when occured the infection for the ith individual; he only knows if at time T_i the subject is seropositive ($Y_i = 1$) or seronegative ($Y_i = 0$) and what is the age of the subject at this time.

The distribution function of the random variable V_i is denoted by $F(z_i)$. Then the conditional probabilities associated with the random "response" indicator Y_i can be written:

$$Pr(Y_i = 1|T_i = t_i, Z_i = z_i) = F(t_i|z_i) \tag{2.2}$$

and

$$Pr(Y_i = 0|T_i = t_i, Z_i = z_i) = 1 - F(t_i|z_i). \tag{2.3}$$

If we consider the notation introduced in Section 1.3, we can write:

$$F(t_i|z_i) = F(a) \tag{2.4}$$

and

$$1 - F(t_i|z_i) = 1 - F(a) = P(a). \tag{2.5}$$

2.2 Dependance of seropositive proportion on age

Until now, we have always taken into account the dependence of the seropositive proportion on the age of the individual, without giving an explanation of this fact.

The knowledge of this dependance is the result of several studies, conducted in many developed western countries, on the seroepidemiology of some typical childhood infectious diseases (but which can cause serious problem in adulthood) as measles, mumps, rubella and other diseases caused by VZV (varicella-zoster virus). Some of these studies are Farrington [1] on data from UK, Thiry *et al.* [2] on data from Flanders (Belgium), Mossong *et al.* [7] on data from Luxembourg, Cohen *et al.* [8] on data from Israel. All these studies arrive to conclusions which are similar between them:

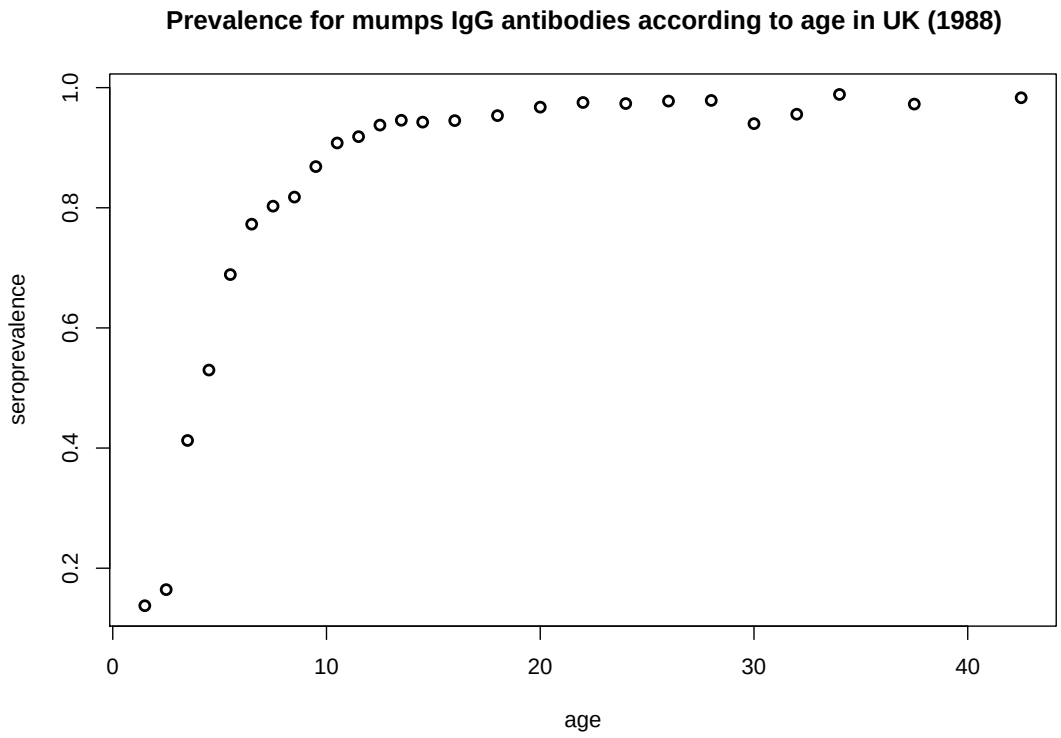


Figure 2.1: Elaboration from data by Farrington [1]

- seroprevalence estimates are significantly associated with age;
- age-specific seroprevalence rises rapidly in the first age classes and then becomes stable in the adolescence;
- the force of infection reaches its maximum in pre-school children and then decreases with age.

For example, we report two graphs, the first one, Fig. 2.1, representing the age-specific seroprevalence of mumps in UK and the second one, Fig. 2.2, representing the age-specific seroprevalence of VZV in Belgium. Although these two sets of data refer to different infections (i.e. mumps and VZV), to two different countries (i.e. United Kingdom and Belgium) and to two different periods (1988 for UK and 2000 for Belgiumd), graphs are very similar to each other.

Then the likeness between these results allows the researchers to consider them as a starting point for similar studies conducted in other countries with the same socio-

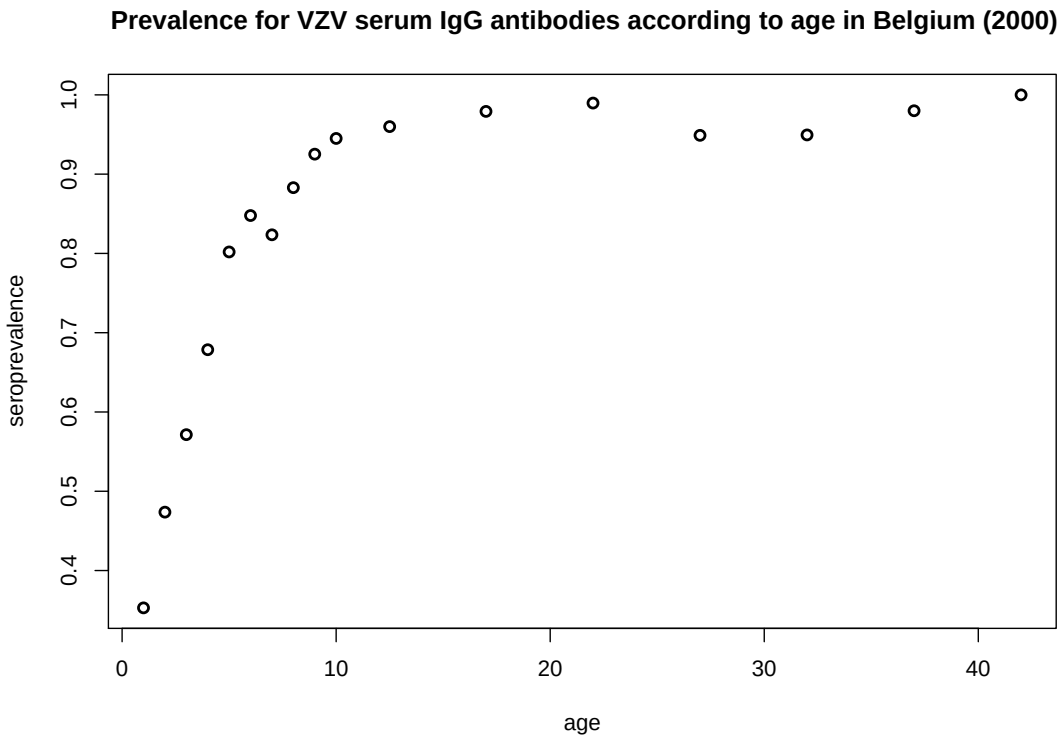


Figure 2.2: Elaboration from data by Thiry *et al.* [2]

economic conditions (and probably the same hygienic-sanitary conditions).

2.3 Sampling method for serological surveys

In a serological survey, as for every other survey, a fundamental step is the sampling phase. The construction of the sample is a critical moment. A good sample must be composed in accordance with the following principles:

- the elements of the sample are selected randomly, that is to say the probability of every population unit to be included in the sample is known, is different from zero and is positive, although this probability can vary from unit to unit;
- the sample must be representative of the population, that is to say it has to be composed in accordance with the variability of the parameter of interest in the population.

If the sample is in accordance with these principles, then it is possible to make inference correctly, applying the new information from the sample to population. Otherwise, it is not correct to make inference and the new information is valid only for the specific sample.

In the case of a serological survey, we have some information, derived from previous studies (Section 2.2), which can be used in the construction of the sample.

Because of seroprevalence is significantly associated with age, that is to say seropositive proportion varies between age classes, as we can see from Fig. 2.1 and Fig. 2.2, we are interested in the estimation of these proportions. We expect that the true value of the parameter, $F(a)$, is different for every class: it can vary from very low values, near 0, in the first years of life, arriving to very high values, near 1, when the individual enters in adulthood.

2.3.1 Simple Random Sampling

Considering an age-stratified population, for every age class $h = 1, 2, \ldots, H$, it is possible to extract a simple random sample (SRS) and from this sample estimate the proportion $\hat{p}_h = F(a)$ of seropositive individuals. Our dataset is composed by status current data, with the following response variable Y_{ih}:

$$Y_{ih} = \begin{cases} 0 & \text{if the infection for the } i\text{th subject of age } h \text{ did not occur} \\ 1 & \text{if the infection for } i\text{th subject of age } h \text{ occured} \end{cases} \tag{2.6}$$

and the explanatory variable Z_i, which is the age of the individual.

For every age class h, the Y_{ih} variable is extracted from a Bernoulli distributed population, whose structure is represented in Tab. 2.1:

The sample extracted from this population has a similar structure, represented in Tab. 2.2:

We are interested in the estimation of proportions π_h $(h = 1, 2, \ldots, H)$ of seropositive individuals in the population:

$$\pi_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}; \tag{2.7}$$

Age class	Elements					Size	Mean	Variance
1	Y_{11}	$\cdots$	Y_{1i}	$\cdots$	Y_{1N_1}	N_1	π_1	$\pi_1(1-\pi_1)$
2	Y_{21}	$\cdots$	Y_{2i}	$\cdots$	Y_{2N_2}	N_2	π_2	$\pi_2(1-\pi_2)$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	Y_{h1}	$\cdots$	Y_{hi}	$\cdots$	Y_{hN_h}	N_h	π_h	$\pi_h(1-\pi_h)$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	Y_{H1}	$\cdots$	Y_{Hi}	$\cdots$	Y_{HN_H}	N_H	π_H	$\pi_H(1-\pi_H)$

Table 2.1: Description of an age-stratified population

Age class	Elements					Size	Mean	Variance
1	y_{11}	$\cdots$	y_{1i}	$\cdots$	y_{1n_1}	n_1	$\hat{\pi}_1$	$\hat{\pi}_1(1-\hat{\pi}_1)$
2	y_{21}	$\cdots$	y_{2i}	$\cdots$	y_{2n_2}	n_2	$\hat{\pi}_2$	$\hat{\pi}_2(1-\hat{\pi}_2)$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	y_{h1}	$\cdots$	y_{hi}	$\cdots$	y_{hn_h}	n_h	$\hat{\pi}_h$	$\hat{\pi}_h(1-\hat{\pi}_h)$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	y_{H1}	$\cdots$	y_{Hi}	$\cdots$	y_{Hn_H}	n_H	$\hat{\pi}_H$	$\hat{\pi}_H(1-\hat{\pi}_H)$

Table 2.2: Description of an age-stratified sample

The maximum likelihood estimator

To get an estimate of π_h, we need an *estimator*. An estimator of the parameter π_h is a statistic, i.e. a known function of observable random variables, whose values are used to estimate the parameter π_h and which is a function and a random variable at the same time.

In the event of a binary random variable, the method to estimate the parameter π_h is the *maximum likelihood method*. The statistic defined with this method is called *maximum likelihood* (ML) estimator. Firstly, let us define a likelihood function, that is the joint probability density function of the n random variables from the sample and is a function of the parameter π_h:

$$\ell_h(\pi_h; y_{1h}, \ldots, y_{nh}) = f_{y_{1h},\ldots,y_{nh}}(y_{1h}, \ldots, y_{nh}; \pi_h). \tag{2.8}$$

The maximum likelihood estimate (MLE) of the unknown parameter of the population, $\hat{p}_h$, is the value of π_h corresponding to the maximum of $\ell_h(\pi_h; y_{ih}, \ldots, y_{in})$, i.e. the MLE is the value of π_h that is "most likely" to have produced the data $\{y_{ih}\}$. In case of a serological survey, for the hth age class, we have a random sample of size n_h extracted from a Bernoulli distribution:

$$f_h(y_h;\pi_h) = \pi_h^{y_h}(1-\pi_h)^{1-y_h}. \tag{2.9}$$

The likelihood function is:

$$\begin{aligned} \ell_h(\pi_h; y_{1h},\ldots,y_{nh}) &= \prod_{i=1}^{n_h} \pi_h^{y_{ih}}(1-\pi_h)^{1-y_{ih}} && (2.10)\\ &= \pi_h^{\sum_i y_{ih}}(1-\pi_h)^{n_h-\sum_i y_{ih}}. && (2.11) \end{aligned}$$

To maximize $\ell_h(\pi_h)$, we evaluate its first derivative with respect to the parameter π_h and set it to 0:

$$\frac{\partial \ell_h(\pi_h)}{\partial \pi_h} = 0, \tag{2.12}$$

or, which is the same, we maximize the natural logarithm of the likelihood function:

$$\frac{\partial \ln \ell_h(\pi_h)}{\partial \pi_h} = \frac{\partial L_h(\pi_h)}{\partial \pi_h} = 0. \tag{2.13}$$

The logarithm is a monotone increasing function, so that $\ell(\theta)$ and $L(\theta)$ have their maxima for the same value of θ. This transformation is sometimes necessary, because it makes the evaluation easier. So, $L_h(\pi_h)$ is:

$$L_h(\pi_h) = \sum_i y_{ih} \ln \pi_h + (n_h - \sum_i y_{ih}) \ln(1-\pi_h). \tag{2.14}$$

Now, let us maximize the log-likelihood function setting to 0 the score function that we have obtained:

$$\frac{\partial L_h(\pi_h)}{\partial \pi_h} = \sum_i y_{ih}\frac{1}{\pi_h} - (n_h - \sum_i y_{ih})\frac{1}{1-\pi_h} = 0. \tag{2.15}$$

Thus the likelihood equation is:

$$\frac{\sum_i y_{ih}}{\pi_h} = \frac{n_h - \sum_i y_{ih}}{1-\pi_h}; \tag{2.16}$$

after some simple calculations, we obtained the ML estimator:

$$p_h = \frac{\sum_i y_{ih}}{n_h}. \tag{2.17}$$

Now let us see which are, generally, the *large-sample properties* of ML estimators,

that is to say some properties defined for a sample size tending to infinity.

1. **Consistency**: as the sample size increases, the ML estimate converges to the true parameter value, that is

$$\lim_{n\to\infty} Pr\{|\hat{\pi}_h - \pi_h| < \varepsilon\} = 1, \tag{2.18}$$

   where ε is a sufficiently small positive value.

2. **Invariance**: if $f(\pi_h)$ is an invertible function of the unknown parameter of the distribution, then the MLE of $f(\pi_h)$ is $f(\hat{\pi}_h)$, i.e. the MLE of a function of the parameters is simply that function evaluated at the MLE. For example, the MLE of $\sqrt{\pi} = (\hat{\pi})^{1/2}$.

3. **Asymptotic normality and efficiency**: as the sample size increases, the sampling distribution of the MLE converges to a normal and (generally) no other estimation procedure has a smaller variance. Hence, for sufficiently large sample sizes, estimates obtained via maximum likelihood typically have the smallest confidence intervals.

4. **Variance**: for large sample sizes, the variance of an ML estimator is approximately the negative of the reciprocal of the second derivative of the log-likelihood function,

$$Var(p) \approx -\left[E_{\pi_h}\left[\frac{\partial^2}{\partial \pi^2} L(\pi; \mathbf{y})\right]\right]^{-1}. \tag{2.19}$$

   This is just the reciprocal of the curvature of the log-likelihood surface at the MLE. The flatter the likelihood surface around its maximum value (the MLE), the larger the variance; the steeper the surface, the smaller the variance. The minus sign appears because the second derivative is negative (downward curvature) at the maximum of the likelihood function.

Now let us see which is the asymptotical variance of the estimator p_h, in accordance with the fourth property of MLEs. Firstly, let us take the second derivative of the log-likelihood:

$$\begin{aligned}\frac{\partial^2}{\partial \pi_h^2} L_h(\pi_h; \mathbf{y})^2 &= -\sum_i y_{ih} \frac{1}{\pi_h^2} - (n - \sum_i y_{ih}) \frac{1}{(1-\pi_h)^2} \\ &= \frac{-\sum_i y_{ih} + 2\pi_h \sum_i y_{ih} - n_h \pi_h^2}{\pi_h^2 (1-\pi_h^2)}. \end{aligned} \tag{2.20}$$

Then, let us evaluate the expected value of Eq. 2.20:

$$\begin{aligned} & E\left[\frac{-\sum_i y_{ih} + 2\pi_h \sum_i y_{ih} - n_h \pi_h^2}{\pi_h^2 (1-\pi_h^2)}\right] \\ = & \frac{-\sum_i E[y_{ih}] + 2\pi_h \sum_i E[y_{ih}] - n_h \pi_h^2}{\pi_h^2 (1-\pi_h^2)} \\ = & \frac{-n_h \pi_h + 2 n_h \pi_h^2 - n_h \pi_h^2}{\pi_h^2 (1-\pi_h^2)} \\ = & \frac{-n_h \pi_h + n_h \pi_h^2}{\pi_h^2 (1-\pi_h^2)} \\ = & \frac{-n_h \pi_h (1-\pi_h)}{\pi_h^2 (1-\pi_h^2)} \\ = & \frac{-n_h}{\pi_h (1-\pi_h)}. \end{aligned} \tag{2.21}$$

Eventually, let us take the negative of the reciprocal of Eq. 2.21:

$$Var(p_h) \approx \frac{\pi_h (1-\pi_h)}{n_h} \tag{2.22}$$

and so we have found the asymptotic variance of the estimator p_h of the parameter π_h.

Besides, it is possible to demonstrate that the estimator p_h is the correct estimator of Horvitz and Thompson:

$$
\begin{aligned}
p_{HT,h} &= \frac{1}{N_h}\sum_{i=1}^{n_h}\frac{y_{hi}}{P_{hi}} \\
&= \frac{1}{N_h}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} \\
&= \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} \\
&= p_h
\end{aligned} \tag{2.23}
$$

where every sampling unit y_{hi} is weighted for the reciprocal of the *probability of inclusion of first order* $P_{hi} = n_h/N_h$: for every population unit, its probability to be included in the sample is equal to the sampling fraction of the age class it belongs to.

2.3.2 Sampling distribution of the estimator p

In case of proportions, there are two important problems we have to deal with:

1. what is the optimal sample size?
2. which is the best confidence interval for a proportion?

To solve these problems, before we have to better understand the features of the sampling distribution of the estimator p.

Given the following sampling realization $\{y_1, y_2, \ldots, y_n\}$ of the random variable Y_i Bernoulli distributed

$$
Y_i \sim \text{Ber}(\pi, \pi(1-\pi)), \tag{2.24}
$$

we have that the estimator $p = \sum_i y_i/n$ of the parameter π is Bernoulli distributed too:

$$
p \sim \text{Ber}\left(\pi, \frac{\pi(1-\pi)}{n}\right). \tag{2.25}
$$

Being p a ML estimator, for its large-sample properties we have that the distribution of p tends asymptotically to the normal distribution. This happens because of the *Central Limit Theorem*:

Theorem 1 (Central Limit Theorem) *Let $f(\cdot)$ be a probability density function with mean μ and finite variance σ^2. Let $\overline{X}_n$ be the sampling mean of a random sample of size n extracted by $f(\cdot)$. Let the random variable Z_n defined by*

$$Z_n = \frac{\overline{X}_n - E[\overline{X}_n]}{\sqrt{Var[\overline{X}_n]}} = \frac{\overline{X}_n - \mu}{\sigma/n}. \tag{2.26}$$

Then the probability density function of Z_n tends in distribution to the standard normal distribution $N(0,1)$:

$$\frac{\overline{X}_n - \mu}{\sigma/n} \xrightarrow{d} N(0,1). \tag{2.27}$$

This theorem tells us that the limit distribution of Z_n is a standard normal distribution or, that is the same, $\overline{X}_n$ is asymptotically distributed as a normal random variable with mean μ and variance σ^2/n. It is interesting to notice that the theorem nothing says about the form of the original probability density function $f(\cdot)$.

In our case, we have that

$$Z_n = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \xrightarrow{d} N(0,1), \tag{2.28}$$

or, that is the same,

$$p \xrightarrow{d} \mathrm{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right). \tag{2.29}$$

However, differently from a sampling mean from a continuous variable, in case of a sampling mean from a discrete variable (which is the case of a proportion), the approximation to the normal distribution is more problematic. The plain difference between the sampling distribution of the estimator p and the sampling distribution of the estimator $\bar{y} = (\sum_i y_i)/n$, with Y_i which is a continuous variable, is in their variance.

In effect, the variance of $\bar{y}$ depends on the value of the population variance, whatever is its expected value μ:

$$Var(\bar{y}) = \frac{\sigma^2}{n}. \tag{2.30}$$

On the contrary, the variance of p depends on the true value of the parameter π in the population:

$$Var(p) = \frac{\pi(1-\pi)}{n}. \tag{2.31}$$

This fact has an important consequence. From the one hand, when we estimate the parameter μ from the realization of a quantitative variable, the dispersion of the estimator $\bar{y}$ do not depend on the value assumed by μ.

From the other hand, when we estimate the parameter π from the realization of a qualitative variable, the dispersion of the estimator p depends on the value of π. When the unknown parameter π gets closer 0 or 1, the variance $\pi(1-\pi)$ of X_i becomes smaller: because of this fact, the precision of the estimator p increases and so it should be sufficient a moderate sample size n to estimate π. When π is near $1/2$ instead, the variance $\pi(1-\pi)$ becomes larger and so the precision of the estimator decreases: now, it should be necessary gather a larger sample.

However, since this is the case of a Bernoulli distributed variable and Bernoulli is a discrete distribution, there is something other which is important to notice. When π is near $1/2$, the sampling distribution of p approximates more quickly the Normal, because of the symmetrical distribution: from this point of view, it should be sufficient a moderate sample size. When π gets closer 0 or 1 instead, the sampling distribution of p becomes skewer and so it reaches slowlier the approximation with the Normal.

Obviously, the two preceding facts go in opposite directions. Let us make two examples.

- If $\pi = 0.5$, then the variance of p is $\pi(1-\pi)/n = 0.25/n$. This is the case of maximum variance and so the estimator is less accurate; however the sampling distribution of p approximates more quickly the Normal.

- If $\pi = 0.2$, then the variance of p is $\pi(1-\pi)/n = 0.16/n$. In this case, the variance is smaller and so the estimator is more accurate; however, the sampling distribution of p reaches the Normal slowlier.

The resolution of this question is fundamental to determine the optimal sample size. To make this, we have to see for which values of n the distance between the binomial distribution and the Normal distribution becomes very small.

The Cramer-Von Mises criterion

To evaluate the distance between two different distribution, we can use the *Cramer-Von Mises criterion*. This test is used to judge the goodness of fit of a probability distribution $f*(x)$ compared to a given distribution $f(x)$ and is given by

$$W^2 = \int_{-\infty}^{\infty} [F*(x) - F(x)]dF(x), \tag{2.32}$$

where $F*(x)$ is the cumulative distribution function of the pdf whose adequacy we want to test and $F(x)$ is the cumulative distribution function of the theoretical pdf. In

practice, this criterion is given by a sort of euclidean distance between the two distributions.

In our case, we want to test if the binomial distribution, i.e. $F*(x)$, fits well the Normal distribution, i.e. $F(x)$, and for which values of the sample size n this happens. Of course, the Normal must have the same mean and variance of the binomial. So, if the binomial is distributed as $Bin(\pi, n)$, then the Normal will be distributed as $N(n\pi, n\pi(1-\pi))$.

We have used this test for three different values of the probability of success π of the binomial distribution: 0.01, 0.5 and 0.95. For every case we have plot the value W^2 of test against the sample size n.

The first graph in Tab. 2.3 shows what happens to the euclidean distance W^2 when $\pi = 0.01$, that is the event is very rare. We have that the distance reaches quickly a peak at $n = 20$ ($W^2 = 0.0673$) and then decreases slightly until 0.0221 at $n = 200$. Thus, when the event is very rare the binomial distribution fits badly approximates the Normal distribution.

The second graph in Tab. 2.3 shows what happens to the distance W^2 when $\pi = 0.5$: now the binomial distribution is perfectly symmetrical and so we expect that it will reach quickly the approximation with the Normal. In effect, we have that W^2 decreases until very low values, of order 10^{-7} - 10^{-11}. However, this decrease shows some jumps towards very low values: the first jump happens in the range $33 \leq n \leq 44$, the second in the range $75 \leq n \leq 132$ and the third begins with $n = 145$.

Eventually, the third graph shows what happens to W^2 when $\pi = 0.95$, that is the event is very frequent. In this case we have that the convergence of the binomial to the Normal happens at low values of n: when $n = 20$, we have already that the distance W^2 is of order 10^{-5}.

Therefore, we have seen that when the event is not very frequent ($\pi < 0.5$) the binomial badly converges to the Normal, so we need too high values of the sample size n to obtain correct estimates of the prevalence. On the contrary, when the event is frequent ($\pi \geq 0.5$), the binomial quickly converges to the Normal distribution for values of n greater than 80.

2.3.3 The optimal sample size

We have previously seen that the choice of the optimal sample size is a critical moment in the organization of a survey, because there are both statistical and economic constraints

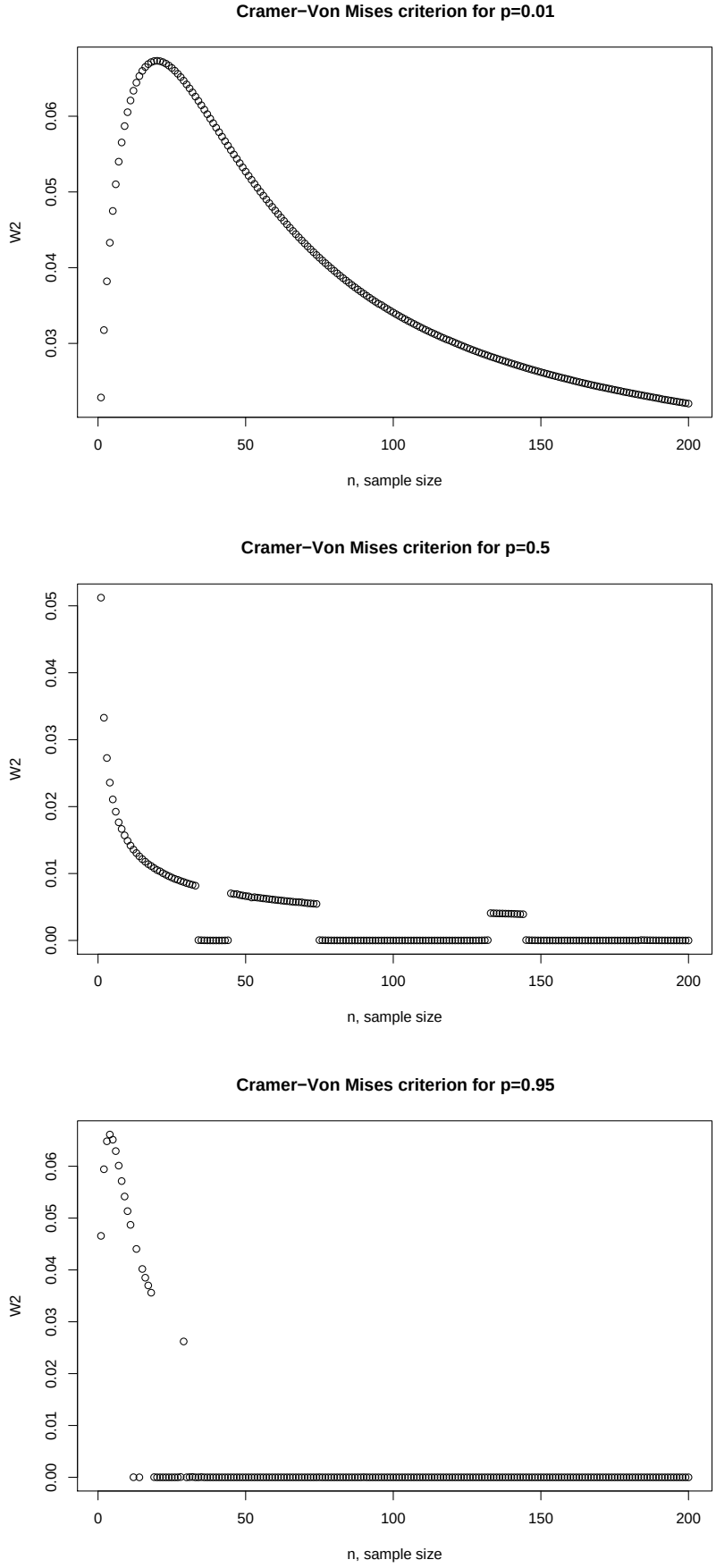


Figure 2.3: Cramer-Von Mises criterion for $n = 1$ to 200 and $\pi = 0.01$, $\pi = 0.5$ and $\pi = 0.95$

that have to be respected.

From the one hand, the smaller the size is, the lower the efficiency of the estimate is, that is the variance of the estimator p_h increases at the reduction of the sample size n and so it becomes more difficult to make inference correctly.

From the other hand, however, we need to consider the high costs of a serological survey, i.e. the costs for the collection of specimen sera and the analysis of these with very expensive machines, so the major the size is, the major the costs of the survey are.

In the previous subsection we have seen that, if the probability of success is below 0.5, we can need a very large sample size; if the probability is higher than 0.5, then we can obtain good estimates with values of n greater than 80. Therefore, in general, the question is: what is the minimum sample size, for every age class, to estimate correctly the proportion p_h?

A classical formula to determine the optimal sample size derives from the inversion of the standard interval for a binomial proportion, also known as the *Wald interval* since it comes from the Wald large sample test for the binomial case:

$$\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \tag{2.33}$$

if the population is virtually infinite (sampling with replacement). The formula for the calculation of the optimal sample size is:

$$n = \frac{4\kappa^2 \hat{\pi}(1-\hat{\pi})}{A^2}, \tag{2.34}$$

where $\kappa = z_{\alpha/2}$ and A is the size of the standard interval confidence (the difference between the upper bound and the lower bound). For the value of the estimate $\hat{\pi}$, there are two possibilities:

- if there are some information about the true value of the parameter π in the population, i.e. from previous similar studies, then it is possible to use this value for the formula;
- if there are not any information about the true value of the parameter π in the population, then it is possible to use $\hat{\pi} = 0.5$, which corresponds to maximum variance (and so to maximum size).

However, there are some problems related to this formula, because of the Wald test from which it derives. A caveat for Eq. 2.34 is that the phenomenon under study is not rare, i.e. $\pi \approx 0.5$, so it is not correct to use the formula when π is near 0 or 1. This

advice reflects the concern that the actual coverage probability of the Wald interval is poor for π near the extremes of the interval [0,1].

Therefore, to know which is the optimal sample size, we have to understand before why the standard interval is not a good confidence interval and which other intervals are better to use.

2.4 Interval Estimation for a binomial proportion

Generally, when constructing a confidence interval, we wish the actual coverage probability to be close to the nominal confidence level. Because of the discrete nature of the binomial distribution, we cannot always achieve the exact nominal confidence level $1-\alpha$ if a randomized procedure is not used. Thus our objective is to construct nonrandomized confidence intervals (CI) for π such that the coverage probability $C(\pi, n)$ is:

$$C(\pi, n) = Pr(\pi \in CI) \approx 1 - \alpha, \tag{2.35}$$

where α is the prespecified significance level.

2.4.1 The Wald interval

The Wald interval for the estimate $\hat{\pi}$ is based on a normal approximation and is obtained by inverting the acceptance region of the Wald large-sample normal test for a general problem:

$$\left| \frac{\hat{\theta} - \theta}{\hat{se}(\hat{\theta})} \right| \leq z_{\alpha/2}, \tag{2.36}$$

where θ is a generic parameter, $\hat{\theta}$ is the estimate of θ and $\hat{se}(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$. In the binomial case, if we want to test whether the estimate $\hat{\pi} = X/n$ is significantly equal to π, we have:

$$\left| \frac{\frac{X}{n} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \right| \leq z_{\alpha/2}; \tag{2.37}$$

so, the standard CI is

$$CI_S = \hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{2.38}$$

This CI is usually presented along with some justification based on the Central Limit Theorem.

However this CI has some serious problems: its actual coverage probability $C(\pi, n)$ is often far from the nominal coverage probability level $1 - \alpha$, even when n is large or π is far from 0 or 1.

In effect, we have that the actual coverage probability of the Wald CI contains non-negligible oscillation as both π and n vary. As shown in Brown *et al.* [9], from the one hand, there exist some "lucky" pairs (π, n) such that $C(\pi, n)$ is very close to or larger than the nominal level $1 - \alpha$. On the other hand, there exist "unlucky" pairs (π, n) such that $C(\pi, n)$ is much smaller than the nominal level.

In the following subsections, we report some examples of the inadequacy of the standard interval.

Example 1

Fig. 2.4 plots the coverage probability of the nominal 95% standard interval for $\pi = 0.2$. The number of trials n varies from 25 to 100. It is clear from the plot that the oscillation is significant and the coverage probability does not steadily get closer to the nominal confidence level as n increases. For instance, $C(0.2, 31) = 0.948$ and $C(0.2, 98) = 0.923$. So the coverage probability is significantly closer to 0.95 when $n = 31$ then when $n = 98$.

Example 2

Lucky n	16	22	25	30	35	42	49
$C(0.5, n)$	0.952	0.956	0.957	0.967	0.954	0.956	0.961
Unlucky n	10	12	15	18	23	33	40
$C(0.5, n)$	0.876	0.849	0.886	0.902	0.905	0.897	0.897

Table 2.3: Standard interval; lucky n and unlucky n for $10 \leq n \leq 50$ and $\pi = 0.5$

Now consider the case of $\pi = 0.5$. Since $\pi = 0.5$, we may think that, for $n > 20$, all is well, because the binomial distribution approximates the normal one more rapidly. We evaluate the exact coverage probability of the 95% standard interval for $10 \leq n \leq 50$. In Tab. 2.3 we list the values of "lucky" n, defined as $C(\pi, n) \geq 0.95$, and the values of "unlucky" n, defined as $C(\pi, n) \leq 0.91$. The conclusions do not respect the legitimate

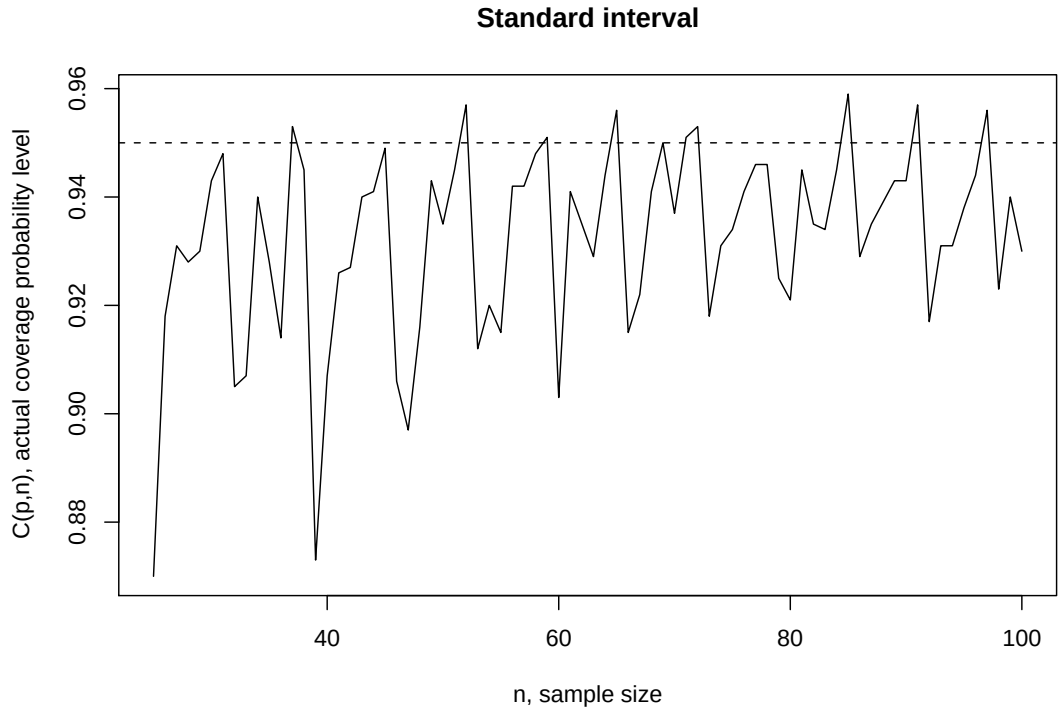


Figure 2.4: Standard interval: oscillation phenomenon for fixed $\pi = 0.2$, variable n=25 to 100 and nominal coverage probability at 95% (dashed line).

expectations that an unsuspecting user could have. For example, for $n = 22$ we have $C(0.5, 22) = 0.956$, while for $n = 23$ we have that $C(0.5, 23) = 0.905$. Indeed, the unlucky values of n arise suddenly: although $\pi = 0.5$, the coverage is still 0.897 at $n = 40$. This illustrates the inconsistency, unpredictability and poor performance of the Wald interval.

Example 3

Unlucky n	592	954	1279	1583	1877
$C(0.005, n)$	0.788	0.855	0.879	0.897	0.891

Table 2.4: Standard interval; late arrival of unlucky n for small π

Let us move π really close to the boundary, say $\pi = 0.005$. Such π are relevant in certain practical applications, as in the case of a serological survey, where the seropositive proportion at the very early age can be very close to 0. Since π is so small, now one may fully expect that the coverage probability of the Wald interval is very poor. Fig. 2.5 and

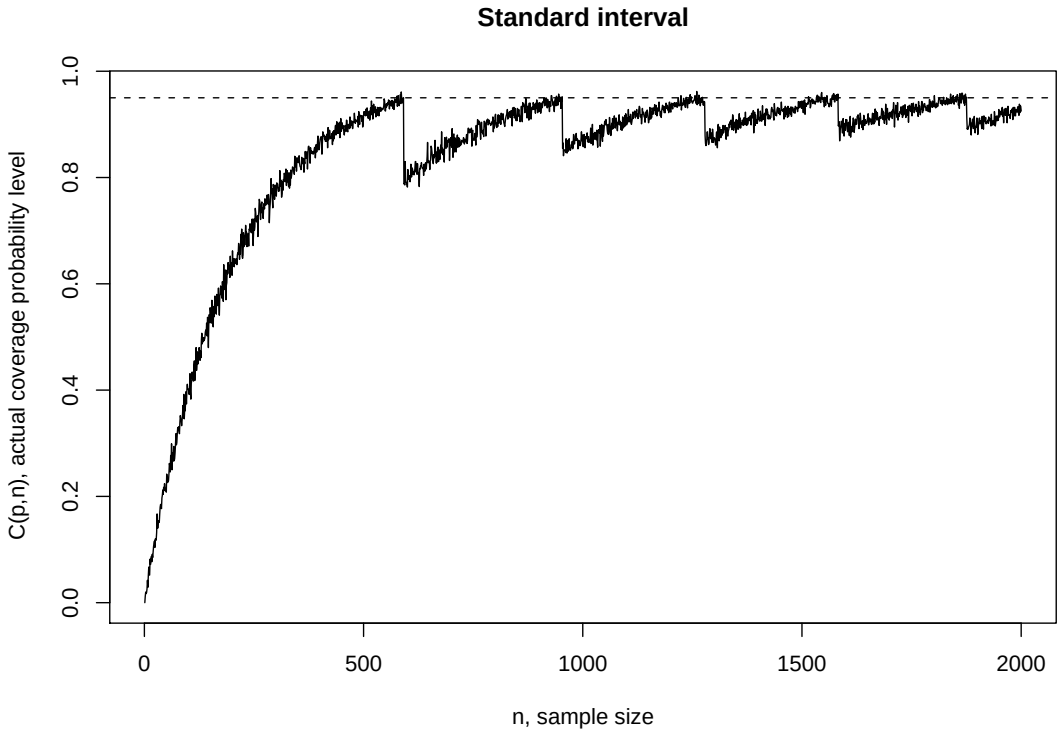


Figure 2.5: Standard interval: oscillation in coverage for small π: $\pi = 0.005$, variable n=1 to 2000 and nominal coverage probability at 95% (dashed line).

Tab. 2.4 show that there are still surprises and indeed we now begin to see a new kind of erratic behaviour. The oscillation of the coverage probability does not show until rather large n. In effect, the coverage probability makes a slow ascent all the way until $n = 591$ ($C(0.005, 591) = 0.944$) and then dramatically drops to 0.793 when $n = 592$. Fig. 2.5 shows that thereafter the oscillations manifest in full force, in contrast with Example 1 and Example 2, where the oscillations started early on. In Tab. 2.4 we report the "unlucky" values of n after the sudden and deep drops of the coverage probability.

Example 4

Fig. 2.6 shows the coverage probability of the nominal 95% standard interval with fixed $n = 100$ and variable π from 0 to 1, with step 0.005. It can be seen from Fig. 2.6 that, in spite of the "large" sample size, significant change in coverage probability occurs at the varying of π. The magnitude of the oscillation increases significantly as π moves toward 0 or 1. Except for values of π quite near $\pi = 0.5$, the general trend of this plot

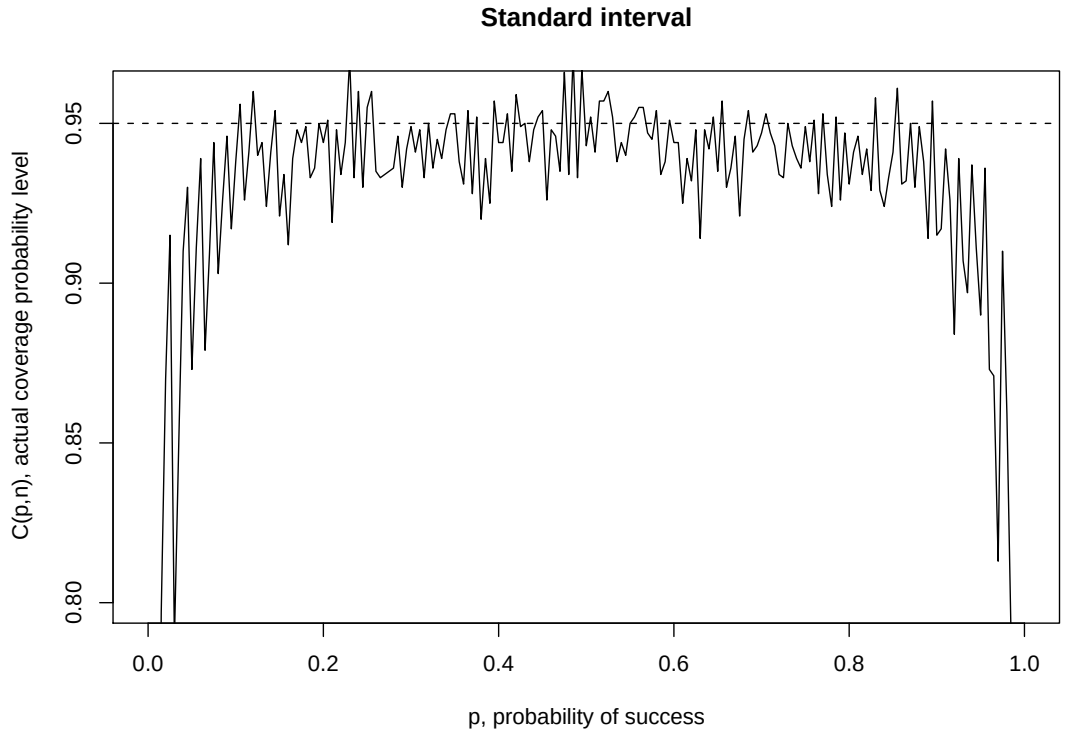


Figure 2.6: Standard interval: oscillation phenomenon for fixed $n = 100$, variable π and nominal coverage probability at 95% (dashed line).

is strikingly below the nominal coverage value of 0.95.

Example 5

Fig. 2.7 shows the coverage probability of the nominal 99% standard interval with $n = 30$ and variable π from 0 to 1, with step 0.005. In addition to the oscillation phenomenon similar to Fig. 2.6, a noticeable fact in this case is that the actual coverage probability never reaches the nominal level: it is always smaller than 0.99 and its average value is only 0.908.

It is evident from the previous examples that the actual coverage probability of the Wald interval can differ significantly from the nominal confidence level for moderate and even large sample sizes and not only when π is near 0 or 1. Fundamentally, there are two kinds of problems in the coverage probability of the standard interval:

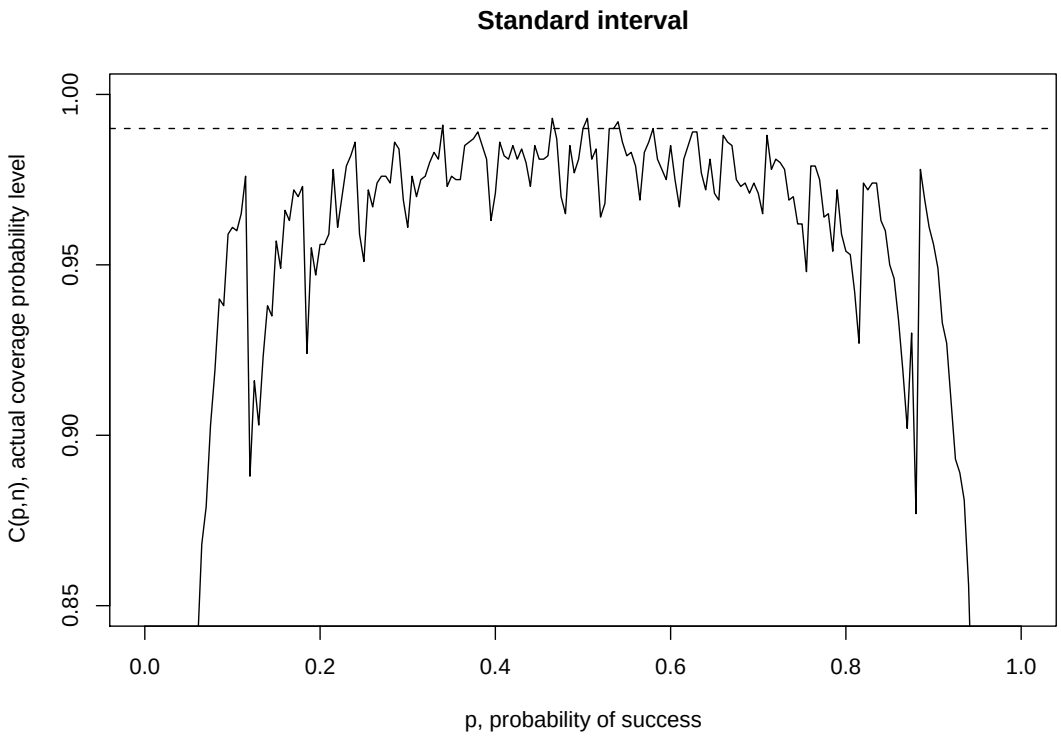


Figure 2.7: Standard interval: coverage of the nominal 99% standard interval for fixed $n = 30$ and variable π

- a systematic negative bias, whatever is the value of π for fixed n;
- an oscillatory behaviour at the varying of n for fixed π.

The reason for the bias

Examples 4 and 5 indicate that there is a systematic negative bias in the coverage probability of the standard interval. The bias is due mainly to the fact that the standard interval has the "wrong" center. The Wald interval is centered at $\hat{\pi} = X/n$. Although $\hat{\pi}$ is the maximum-likelihood estimate (MLE) and an unbiased estimate of π, the choice of using it as the center of a confidence interval causes a systematic negative bias in the coverage. As we can see with the alternative confidence intervals, by simply recentering the interval at $\tilde{\pi} = (X + \kappa^2/2)/(n + \kappa^2)$, where $\kappa = z_{\alpha/2}$, one can increase the coverage probability significantly for π away from 0 or 1 and eliminate the systematic bias.

We know, from the application of the Central Limit Theorem, that

$$Z_n = \frac{n^{1/2}(\hat{\pi} - \pi)}{\sqrt{\pi(1-\pi)}} \sim N(0,1). \tag{2.39}$$

The standard interval is based on the hypothesis that the Wald test W_n is asymptotically standard normally distributed:

$$W_n = \frac{n^{1/2}(\hat{\pi} - \pi)}{\sqrt{\hat{\pi}(1-\hat{\pi})}} \sim N(0,1). \tag{2.40}$$

The problem is all in the difference between Z_n and W_n: we assume that using $\hat{\pi}$ rather than π, the distribution of W_n is the same of Z_n. But we do not know nothing about the distribution of $\hat{\pi}(1-\hat{\pi})$.

Let us take the case of the Wald test applied to the sampling mean $\bar{x}$. The variable Z_n is

$$Z_n = \frac{n^{1/2}(\bar{x} - \mu)}{\sigma} \sim N(0,1); \tag{2.41}$$

the respective Wald test W_n is:

$$W_n = \frac{n^{1/2}(\bar{x} - \mu)}{\sqrt{s^2}}, \tag{2.42}$$

where $s^2 = \sum_i (x_i - \bar{x})^2/n - 1$ is the correct sampling variance. In this case, we can say something about the distribution of W_n. We know that $(\bar{x} - \mu)$ is standard normally distributed and s^2 has a χ^2 distribution with k degrees of freedom. So, from their ratio we have a t-distributed W_n, which approximates the Normal for $n > 20$.

By this comparison we can see that, what happens for the variance s, whose distribution is known, does not occur for $\hat{\pi}(1-\hat{\pi})$, whose distribution is unknown.

In practice, even for quite large values of n, the actual distribution of W_n is significantly nonnormal. Thus, the very premise on which the standard interval is based is seriously compromised for moderate and even quite large values of n. For instance, asymptotically, W_n has bias 0, variance 1, skewness 0 and kurtosis 3. For moderate n, however, the deviations of the bias, variance, skewness and kurtosis of W_n from their respective asymptotic values are often significant and cause a nonnegligible negative bias in the coverage probability of the standard confidence interval, so that the actual coverage $C(p,n)$ rarely reaches the nominal coverage $1-\alpha$.

We can analytically demonstrate the bias in the distribution of W_n by standard

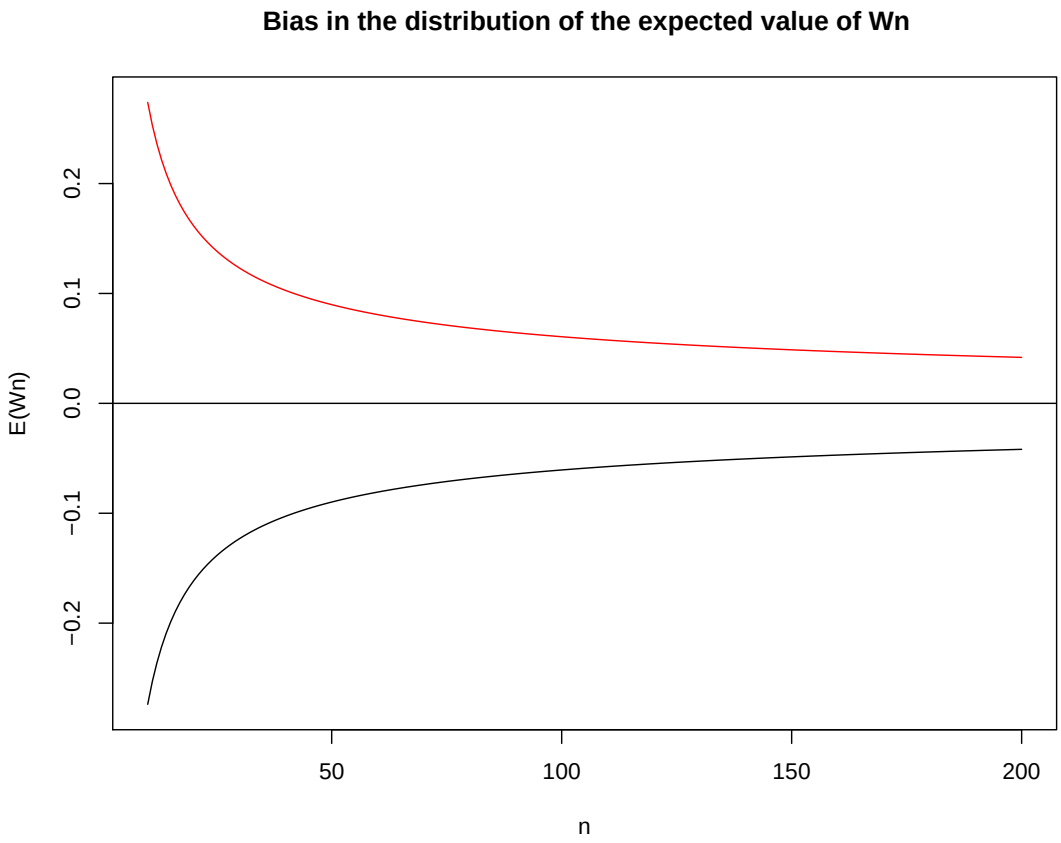


Figure 2.8: Bias in the distribution of the expected value of W_n with $\pi = 0.25$ (black line) and $\pi = 0.75$ (red line)

expansions. Let us write W_n in function of Z_n. After some algebraic passages, we have:

$$W_n(Z_n) = \frac{Z_n}{\sqrt{1 + (1 - 2\pi)Z_n/\sqrt{n\pi(1-\pi)} - Z_n^2/n}}. \tag{2.43}$$

A standard Taylor expansion and formulas for central moments of the binomial distribution then yield an approximation to the bias:

$$E[W_n(Z_n)] \approx \frac{\pi - 1/2}{\sqrt{n\pi(1-\pi)}} \left(1 + \frac{7}{2n} + \frac{9(\pi - 1/2)^2}{2n\pi(1-\pi)}\right). \tag{2.44}$$

It can be seen from Eq. 2.44 and from Fig. 2.8 that W_n has negative bias for $\pi < 0.5$, positive bias for $\pi > 0.5$ and no bias for $\pi = 0.5$. Also from the observation of the plots

of the first derivatives of $E[W_n]$ with respect to p and n, it is possible understand its behaviour.

From Fig. 2.9, we can see that the first derivative of $E[W_n]$ gets closer to 0 (without never reaching it) when p is near $1/2$. From Fig. 2.10, we can see that the first derivative of $E[W_n]$ tends asymptotically to 0 when n increases.

Therefore, ignoring the oscillation effect, one can expect to increase the coverage probability by shifting the center of the standard interval towards $1/2$, for which $E[W_n] = 0$.

The reason for the oscillation

It is evident from Examples 1, 2 and 3 that the actual coverage probability of the Wald interval can differ significantly from the nominal confidence level at realistic and even larger than realistic sample sizes: indeed, the actual coverage probability oscillates in a significant way near the nominal coverage. The error comes from two sources: *discreteness* and *skewness* in the underlying distribution, that is the binomial distribution. For a two-sided interval, the rounding error due to discreteness is asymptotically dominant: it is of the order $1/\sqrt{n}$ and decreases when n increases. On the contrary, the error due to skewness is secondary and is of the order $1/n$ (minor than $1/\sqrt{n}$), but still important fo even moderately large n.

The oscillation in the coverage probability is caused by the discreteness of the binomial distribution, more precisely the interlaced structure of the binomial distribution. The cumulative distribution function contains jumps at integer points...

Let us try to understand at a more intuitive level why the coverage probability oscillates so significantly. By an easy calculation, one can show that the coverage probability $C(\pi, n) = Pr(\pi \in CI_s)$ equals $Pr(L_{\pi,n} \leq X \leq U_{\pi,n})$, where $L_{\pi,n}$ is the smallest integer larger than or equal to

$$\frac{n(\kappa^2 + 2n\pi) - \kappa n\sqrt{\kappa^2 + 4n\pi(1-\pi)}}{2(\kappa^2 + n)}, \tag{2.45}$$

and $U_{\pi,n}$ is the largest integer smaller than or equal to

$$\frac{n(\kappa^2 + 2n\pi) + \kappa n\sqrt{\kappa^2 + 4n\pi(1-\pi)}}{2(\kappa^2 + n)}. \tag{2.46}$$

What happens is that a small change in n or π can cause $L_{\pi,n}$ and/or $U_{\pi,n}$ to leap to the next integer value. For example, take the case $\pi = 0.5$ and $\alpha = 0.05$. When $n = 39$, we have $L_{0.5,39} = 14$ and $U_{0.5,39} = 25$; but when $n = 40$, $L_{0.5,40}$ leaps to 15,

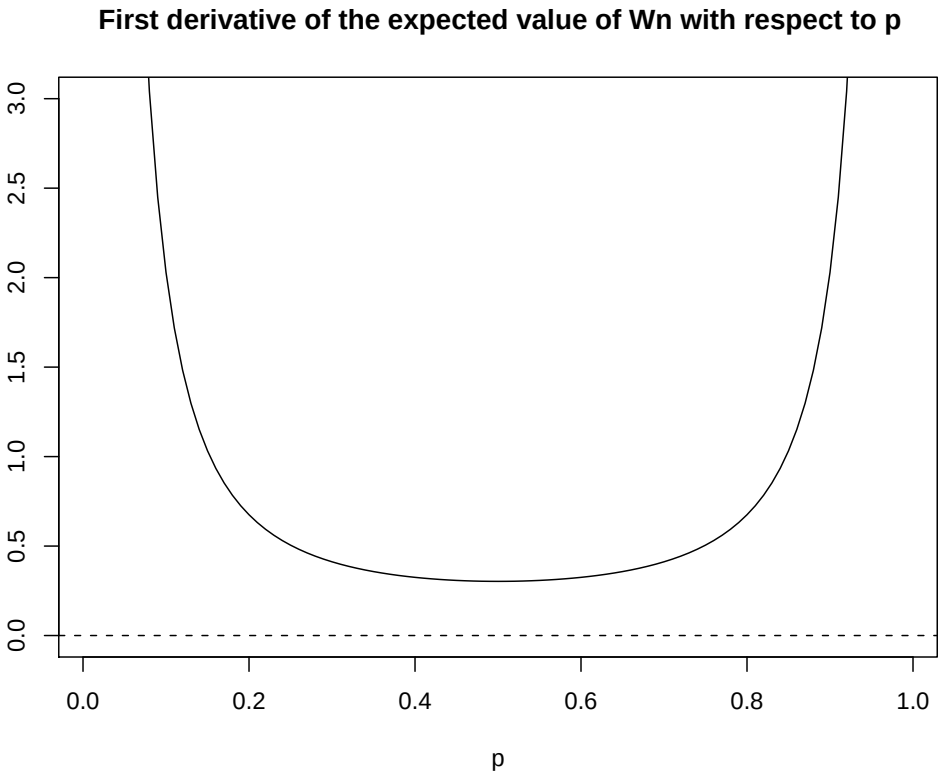


Figure 2.9: First derivative of $E[W_n]$ with respect to π

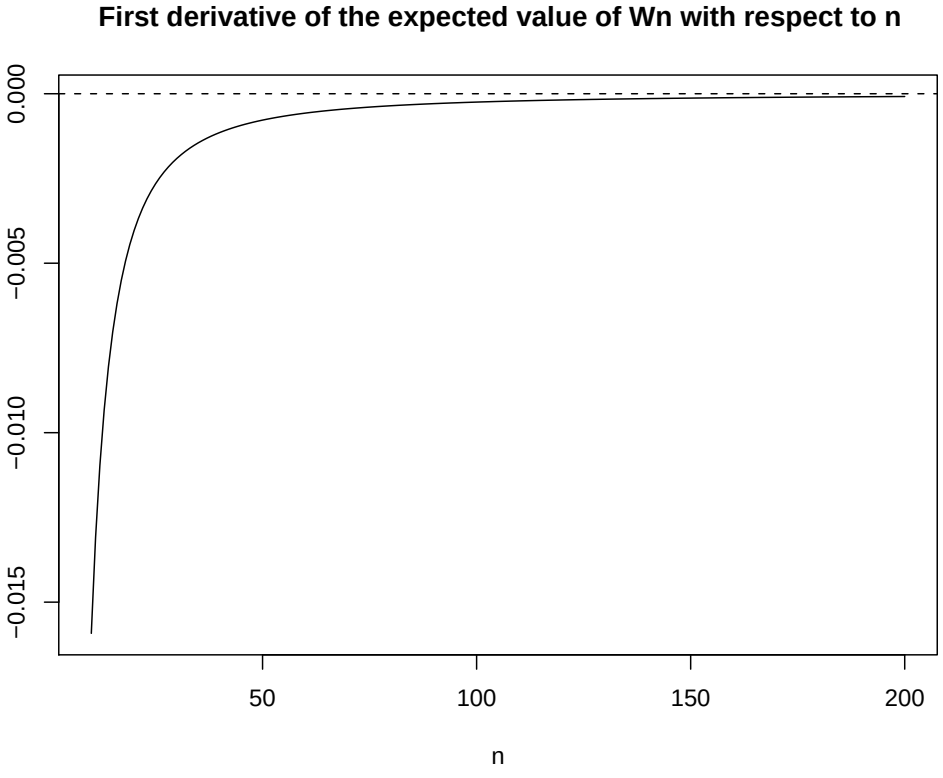


Figure 2.10: First derivative of $E[W_n]$ with respect to n

while $U_{0.5,39}$ remains 25. Thus the set of favorable values of X loses the point $X = 14$ even though n has increased from 39 to 40. This causes $n = 40$ to be an unlucky choice of n: in effect, from data for Example 2, $C(0.5, 39) = 0.935$, while $C(0.5, 40) = 0.897$. This also happens when n is kept fixed and π changes slightly and so we begin to see unlucky values of π.

2.4.2 Recommended alternative intervals

From the evidence of the preceding examples, it seems clear that the standard interval is just too risky. This brings us to the consideration of alternative intervals. We now analyze several such alternatives, each with its motivation.

The Wilson interval

An alternative to the standard interval is the confidence interval based on inverting the *Rao's tailed score test* of $H_0 : \mu = \mu_0$. Here, one accepts H_0 based on Rao's score test if and only if μ_0 is in this interval. The test is

$$Z_n = \left| \frac{\sqrt{n}(\hat{\pi} - \pi)}{\sqrt{\pi(1-\pi)}} \right| \leq \kappa, \tag{2.47}$$

and differs from Wald test because it uses the null standard error $(\pi(1 - \pi)/n)^{1/2}$ instead of the estimate standard error $(\hat{\pi}(1 - \hat{\pi})/n)^{1/2}$: indeed, score tests, and in particular their standard errors, are based on the log likelihood at the null hypothesis value of the parameter π, whereas Wald tests are based on the log likelihood at the maximum likelihood estimate (MLE) $\hat{\pi}$. If we solve the quadratic equation $(\hat{\pi} - \pi)^2 = \kappa\pi(1 - \pi)/n$ to find π, we have the confidence interval:

$$CI_W = \frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa\sqrt{n}}{n + \kappa^2} \sqrt{\hat{\pi}(1 - \hat{\pi}) + \kappa^2/4n}. \tag{2.48}$$

This interval was apparently introduced by Wilson [10] and so it is called the Wilson interval (or *score interval*, because it comes from the inversion of the score test).

As we told talking about the reason for the bias of the Wald interval, the Wilson interval is one of the alternatives which recenters the interval at $\tilde{\pi} = (X + \kappa^2/2)/(n + \kappa^2)$. This point $\tilde{\pi}$ is simply the weighted average of $\hat{\pi}$ and $1/2$, where n and κ^2 are the respective weights:

$$\tilde{\pi} = \frac{X + \kappa^2/2}{n + \kappa^2} = \frac{n\hat{\pi} + \kappa^2 \frac{1}{2}}{n + \kappa^2}. \tag{2.49}$$

It falls between $\hat{\pi}$ and $1/2$, with the weight given to $\hat{\pi}$ approaching 1 aymptotically. This midpoint shrinks the sample proportion towards 0.5, the shrinking being less sever as n increases.

The coefficient of κ in the term that is added to and subtracted from the midpoint to form the score confidence interval can be rewritten in the following way:

$$\sqrt{\frac{1}{n+\kappa^2}\left[\frac{n\hat{\pi}(1-\hat{\pi})+\kappa^2\frac{1}{2}\frac{1}{2}}{n+\kappa^2}\right]}. \tag{2.50}$$

We can see that this has the form of a weighted average of the variance of a sample proportion when $\pi = \hat{\pi}$ and the variance of a sample proportion when $\pi = 1/2$, using $n + \kappa^2$ in place of the usual sample size n.

The Wilson interval can be reccomended for use with nearly all sample sizes and parameter values. Coverage of this interval fluctuates acceptably near the nominal coverage $1 - \alpha$, except for π very near 0 or 1. See Fig. 2.11 and Fig. 2.12.

The Agresti-Coull interval

The standard interval CI_s is simple and easy to remember. For the purpose of classroom presentation and use in texts, it may be nice to have an alternative that has the familiar form $\hat{\pi} \pm z\sqrt{\hat{\pi}(1-\hat{\pi})/n}$, with a better and new choice of $\hat{\pi}$ rather than $\hat{\pi} = X/n$. Brown *et al.* [9] suggests that this can be accomplished by using the center of the Wilson region in place of $\hat{\pi}$. Given the following notation,

- $\widetilde{X} = X + \kappa^2/2$;
- $\tilde{n} = n + \kappa^2$;
- $\tilde{\pi} = \widetilde{X}/\tilde{n}$;
- $\tilde{q} = 1 - \tilde{\pi}$;

we define the Agresti-Coull interval for π by

$$CI_{AC} = \tilde{\pi} \pm \kappa\sqrt{\tilde{\pi}\tilde{q}/\tilde{n}}. \tag{2.51}$$

Both the Agresti-Coull and the Wilson interval are centered on the same value, $\tilde{\pi}$. It is easy to check that the Agresti-Coull interval is never shorter than the Wilson one. For the case when $\alpha = 0.05$, if we use the value 2 instead of 1.96 for κ, this interval is

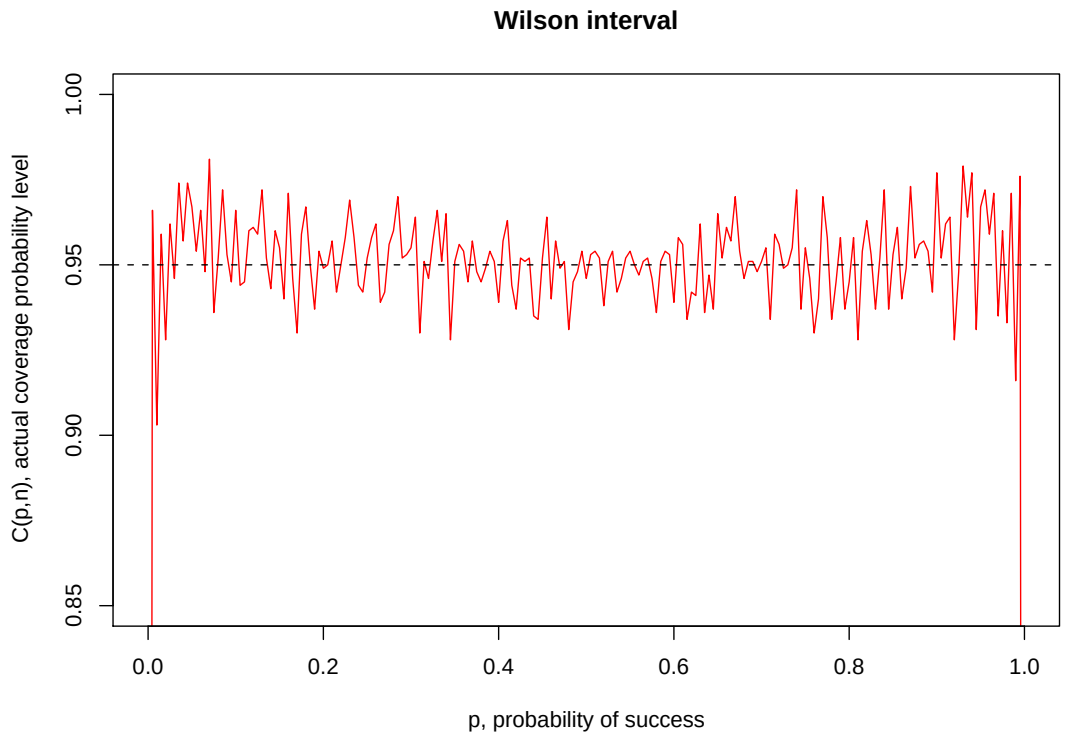


Figure 2.11: Wilson interval: coverage of the nominal 95% standard interval for fixed $n = 50$ and variable π

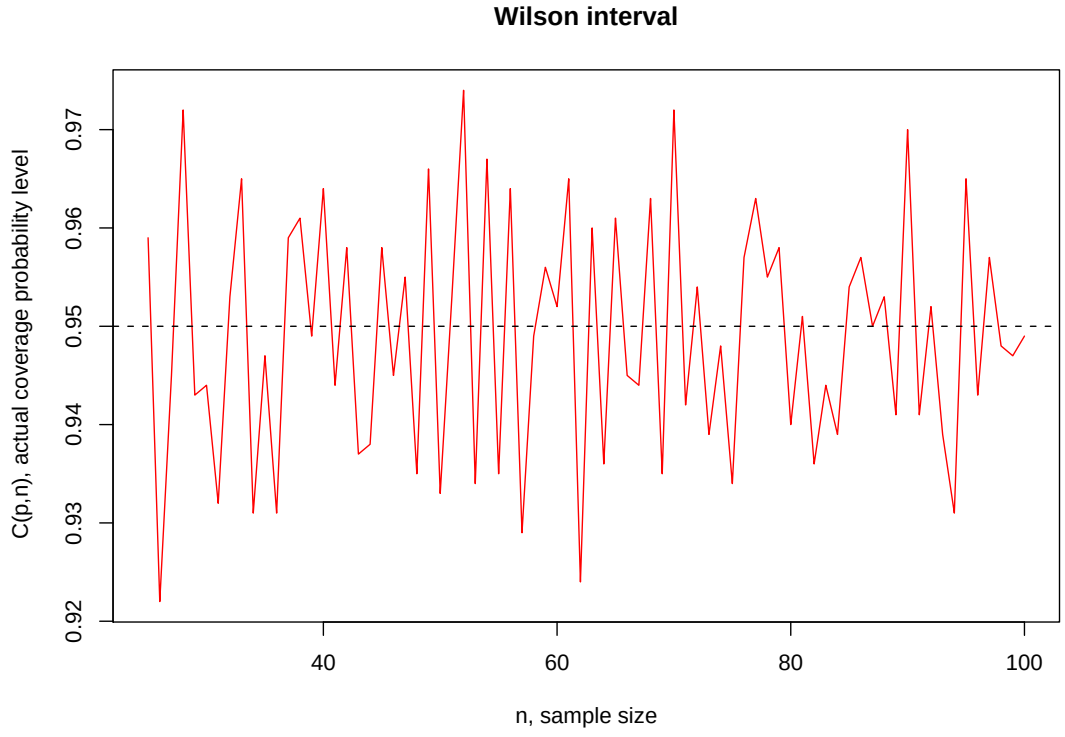


Figure 2.12: Wilson interval: coverage of the nominal 95% standard interval for fixed $\pi = 0.5$ and variable $n = 25$ to 100

the "add 2 successes and 2 failures" interval in Agresti and Coull [11]. For this reason, we call it the Agresti-Coull interval.

The Agresti-Coull interval has good minimum coverage probability. The coverage probability of the interval is quite conservative for π very close to 0 or 1. In comparison to the Wilson interval it is more conservative, especially for small n. See Fig. 2.13 and Fig. 2.14.

The Jeffreys prior interval

Beta distributions are the standard conjugate priors for binomial distributions and it is quite common to use beta priors for inference on π. Before going on, let us see what we intend for *conjugate prior distribution*.

In Bayesian probability theory, a class of prior probability distributions $f(\theta)$ is said to be conjugate to a class of likelihood functions $f(x|\theta)$ if the resulting posterior distributions $f(\theta|x)$ are in the same family as $f(\theta)$. For example, the Gaussian family is conjugate to itself (or self-conjugate): if the likelihood function is Gaussian, choosing a Gaussian prior will ensure that the posterior distribution is also Gaussian. Consider the general problem of inferring a distribution for a parameter θ given some datum or data X. From Bayes' theorem, the posterior distribution is calculated from the prior $f(\theta)$ and the likelihood function $\theta \mapsto f(x \mid \theta)$ as

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \tag{2.52}$$

Let the likelihood function be considered fixed; the likelihood function is usually well-determined from a statement of the data-generating process. It is clear that different choices of the prior distribution $f(\theta)$ may make the integral more or less difficult to calculate, and the product $f(x|\theta)f(\theta)$ may take one algebraic form or another. For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameters): such a choice is a *conjugate prior*. A conjugate prior is an algebraic convenience: otherwise a difficult numerical integration may be necessary. All members of the exponential family have conjugate priors: for a random variable which is a Bernoulli trial with unknown probability of success p in [0,1], the usual conjugate prior is the beta distribution with

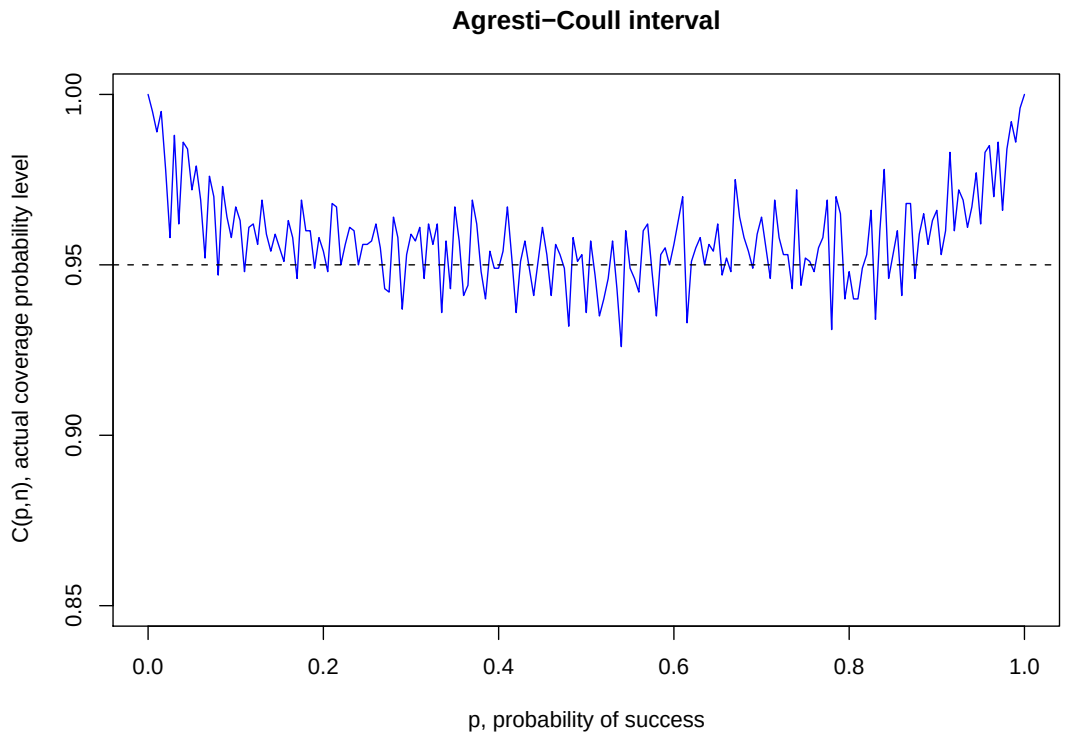


Figure 2.13: Agresti-Coull interval: coverage of the nominal 95% standard interval for fixed $n = 50$ and variable π

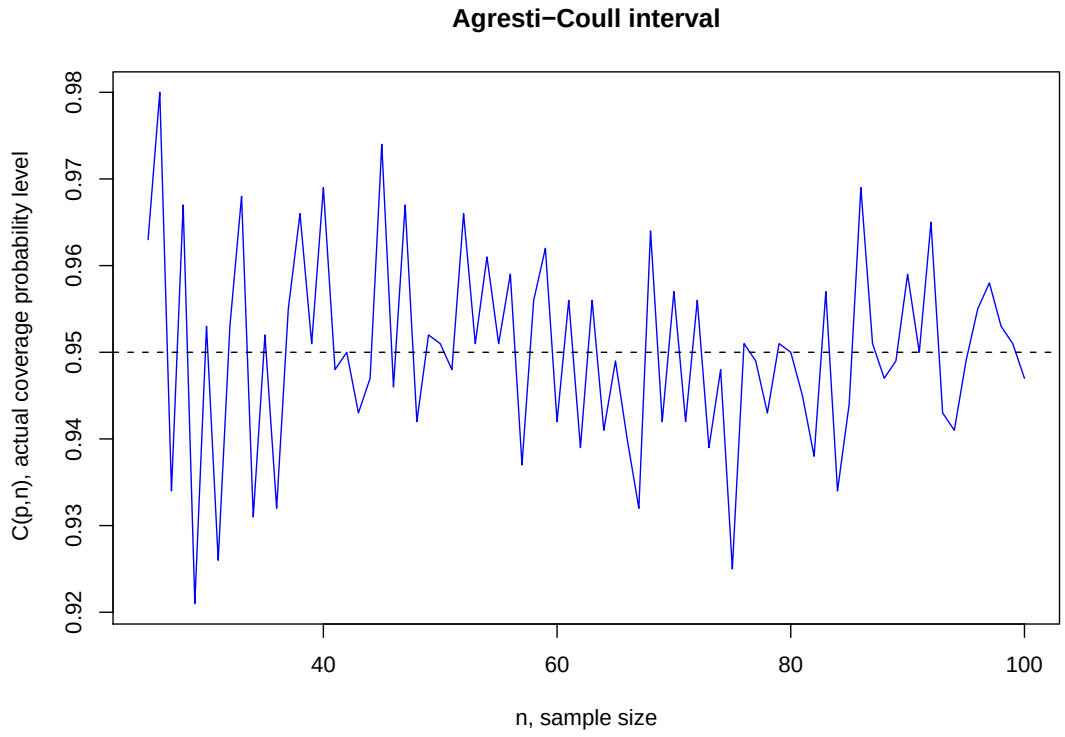


Figure 2.14: Agresti-Coull interval: coverage of the nominal 95% standard interval for fixed $\pi = 0.5$ and variable $n = 25$ to 100

$$f(p = x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \tag{2.53}$$

where α and β are chosen to reflect any existing belief or information (e.g. $\alpha = 1$ and $\beta = 1$ would give a uniform distribution) and $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is the Beta function acting as a normalising constant.

If we then sample this random variable and get s successes and t failures, we have the following likelihood function:

$$f(\pi = x|s,t) = \left(\begin{array}{c} s+t \\ s \end{array}\right) x^s(1-x)^t; \tag{2.54}$$

using beta distributions as the standard conjugate prior for binomial distributions, we have the following posterior distribution:

$$\begin{aligned} f(s,t|\pi = x) &= \frac{\left(\begin{array}{c} s+t \\ s \end{array}\right) x^{s+\alpha-1}(1-x)^{t+\beta-1}/B(\alpha,\beta)}{\int_{y=0}^{1}\left(\left(\begin{array}{c} s+t \\ s \end{array}\right) y^{s+\alpha-1}(1-y)^{t+\beta-1}/B(\alpha,\beta)\right)dy} \\ &= \frac{x^{s+\alpha-1}(1-x)^{t+\beta-1}}{B(s+\alpha,t+\beta)}, \end{aligned} \tag{2.55}$$

which is another Beta distribution with a simple change to the parameters.

Now, we can better understand the construction of the Jeffreys prior interval. Suppose $X \sim Bin(n,\pi)$ and suppose π has a prior distribution $Beta(a_1,a_2)$; then the posterior distribution of π is $Beta(X+a_1, n-X+a_2)$. Thus a $100(1-\alpha)\%$ equal-tailed Bayesian interval is given by

$$L = \left[B\left(\frac{\alpha}{2}; X+a_1, n-X+a_2\right)\right] \tag{2.56}$$

and

$$U = \left[B\left(1-\frac{\alpha}{2}; X+a_1, n-X+a_2\right)\right], \tag{2.57}$$

where $B(\alpha; m_1, m_2)$ denotes the α quantile of a $Beta(m_1,m_2)$ distribution.

The Jeffreys prior interval is a special case of a Bayesian interval. The Jeffreys prior

distribution is $Beta(1/2, 1/2)$. The $100(1-\alpha)\%$ equal-tailed Jeffreys prior interval is defined as

$$CI_J = [L_j(x), U_j(x)], \tag{2.58}$$

where $L_j(0) = 0$ and $U_J(n) = 1$ and otherwise

$$L_J(x) = \left[B_{\alpha/2} \left(X + \frac{1}{2}, n - X + \frac{1}{2} \right) \right], \tag{2.59}$$

$$U_J(x) = \left[B_{1-\alpha/2} \left(X + \frac{1}{2}, n - X + \frac{1}{2} \right) \right]. \tag{2.60}$$

The endpoints of the Jeffreys prior interval are the $\alpha/2$ and $1-\alpha/2$ quintiles of the $Beta(x + 1/2, n - x + 1/2)$ distribution.

The quality of the Jeffreys prior interval is qualitatively similar to that of CI_W over most of the parameter space [0,1]. The coverage has an unfortunate fairly deep spike near $\pi = 0$ and, simmetrically, another near $\pi = 1$. See Fig. 2.15 and Fig. 2.16.

There is also a modified version of the Jeffreys prior interval, which solves some problems of the interval. We have seen previously that the Jeffreys prior interval shows two downward spikes in the coverage function because $U_J(0)$ is too small and simmetrically $L_J(n)$ is too large. To remedy this, one may revise these two specific limits as

$$U_{M-J}(0) = \pi_l \text{ and } L_{M-J}(n) = 1 - \pi_l, \tag{2.61}$$

where π_l satisfies $(1-\pi_l)^n = \alpha/2$ or equivalently $\pi_l = 1 - (\alpha/2)^{1/n}$. Besides, it can be made an other *ad hoc* alteration:

$$L_{M-J}(1) = 0 \text{ and } U_{M-J}(n-1) = 1. \tag{2.62}$$

The Clopper-Pearson interval

The Clopper-Pearson "exact" confidence interval, proposed by Clopper and Pearson [12], is based on the inversion of the equal-tailed binomial test of $H_0 : \pi = \pi_0$, rather than its normal approximation. It has endpoints that are the solutions in p_0 to the equations

$$\sum_{k=x}^{n} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} = \frac{\alpha}{2} \tag{2.63}$$

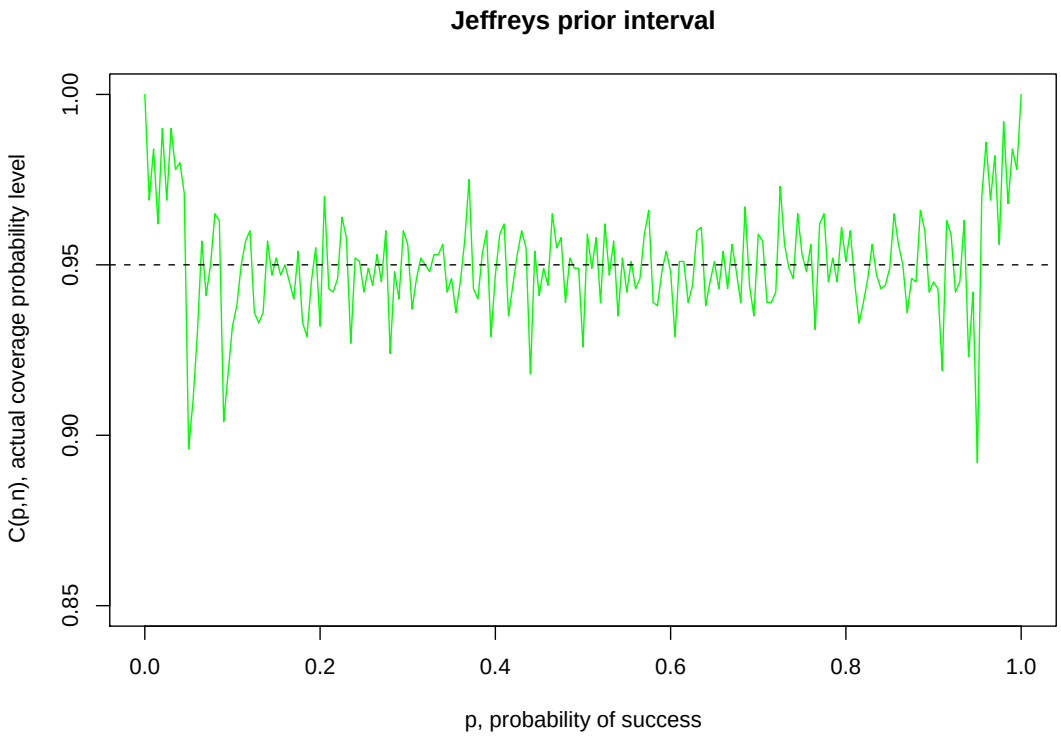


Figure 2.15: Jeffreys prior interval: coverage of the nominal 95% standard interval for fixed $n = 50$ and variable π

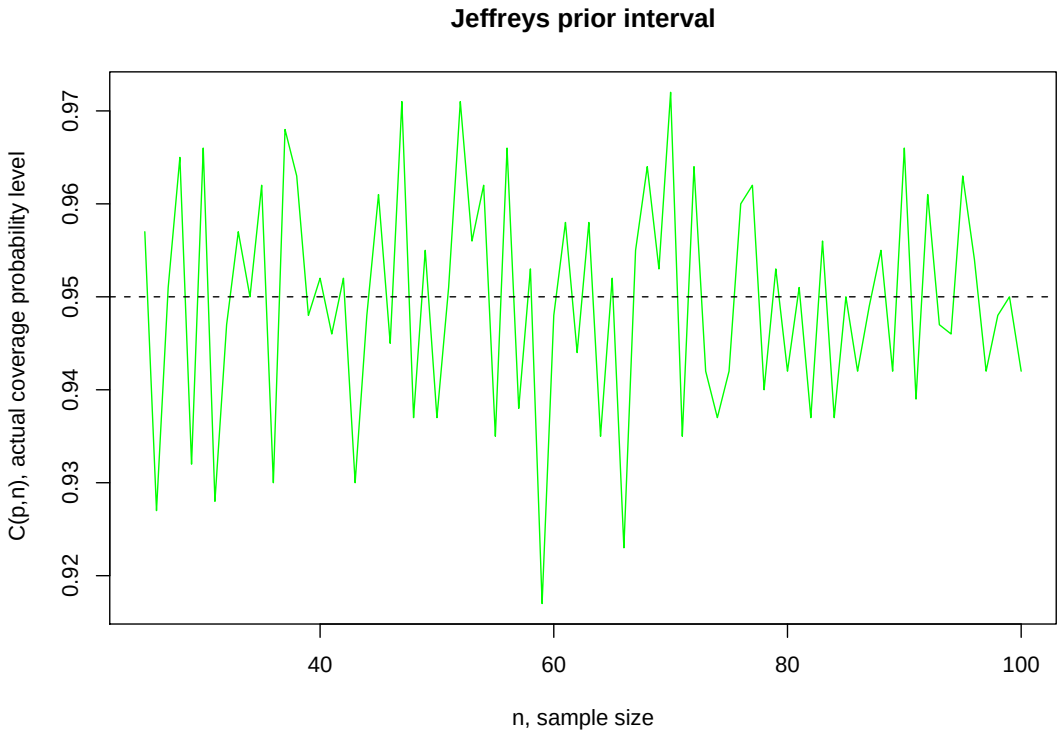


Figure 2.16: Jeffreys prior interval: coverage of the nominal 95% standard interval for fixed $\pi = 0.5$ and variable $n = 25$ to 100

and

$$\sum_{k=0}^{x} \left(\begin{array}{c} n \\ k \end{array} \right) \pi_0^k (1-\pi_0)^{n-k} = \frac{\alpha}{2}, \tag{2.64}$$

except that the lower bound is 0 when $x = 0$ and the upper bound is 1 when $x = n$. This interval estimator is guaranteed to have coverage probability of at least $1 - \alpha$ for every possible value of π. When $x = 1, 2, \ldots, n-1$, the confidence interval equals

$$\left[1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < \pi < \left[1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1} \tag{2.65}$$

and $F_{a,b,c}$ denotes the $1-c$ quantile from the F distribution with degrees of freedom a and b. Equivalently, the lower endpoint is the $\alpha/2$ quantile of a beta distribution $Beta(x, n-x+1)$ and the upper bound is the $1-\alpha/2$ quantile of a beta distribution $Beta(x+1, n-x)$.

The Clopper-Pearson exact interval is typically treated as the "gold standard", although this procedure is necessarily very conservative, because of the discreteness of the binomial distribution. For any fixed parameter value, the actual coverage probability can be much larger than the nominal confidence level unless n is quite large. See Fig. 2.17 and Fig. 2.18.

Coverage probability

Let us also evaluate the intervals in terms of their average coverage probability, the average being over π. Fig. 2.19 demonstrates the striking difference in the average coverage probability among the five intervals previously introduced. The standard interval performs poorly. The Clopper-Pearson "exact" interval is the more conservative in terms of average coverage probability, overall with the smaller values of n. The interval CI_{AC} is slightly conservative, but less than the Clopper-Pearson. Both the Wilson interval and the Jeffreys prior interval have excellent performances in terms of the average coverage probability; that of the Jeffreys prior interval is, if anything, slightly superior. The average coverage probability of the Jeffreys interval is really very close to the nominal level even for quite small n.

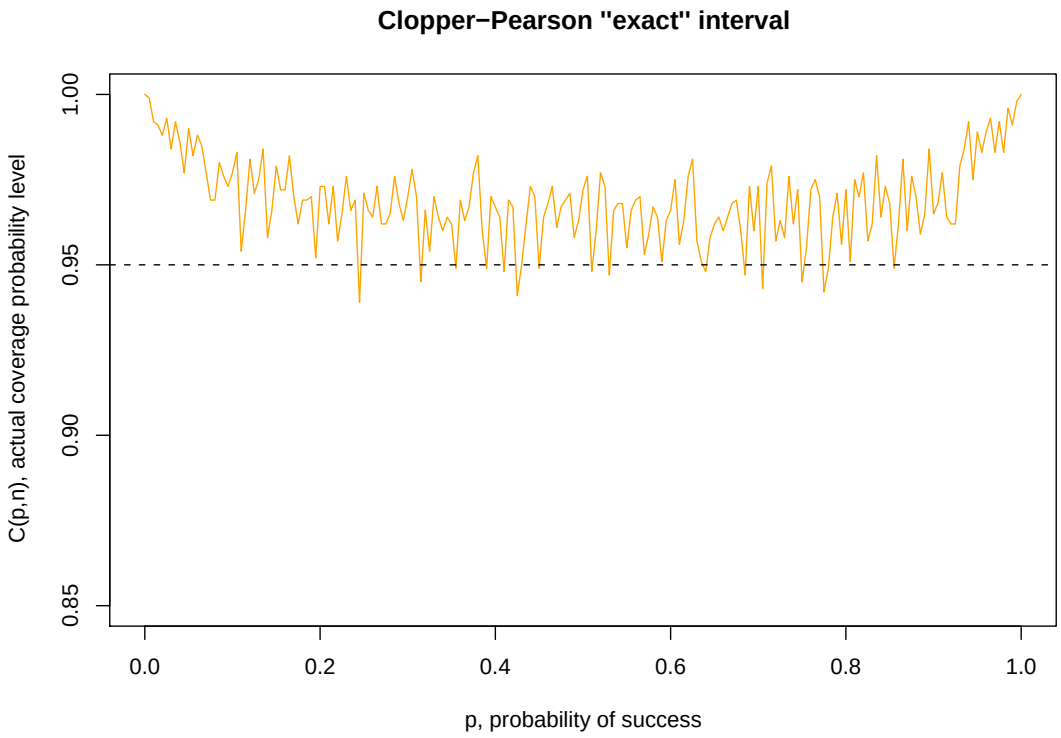


Figure 2.17: Clopper-Pearson "exact" interval: coverage of the nominal 95% standard interval for fixed $n = 50$ and variable π

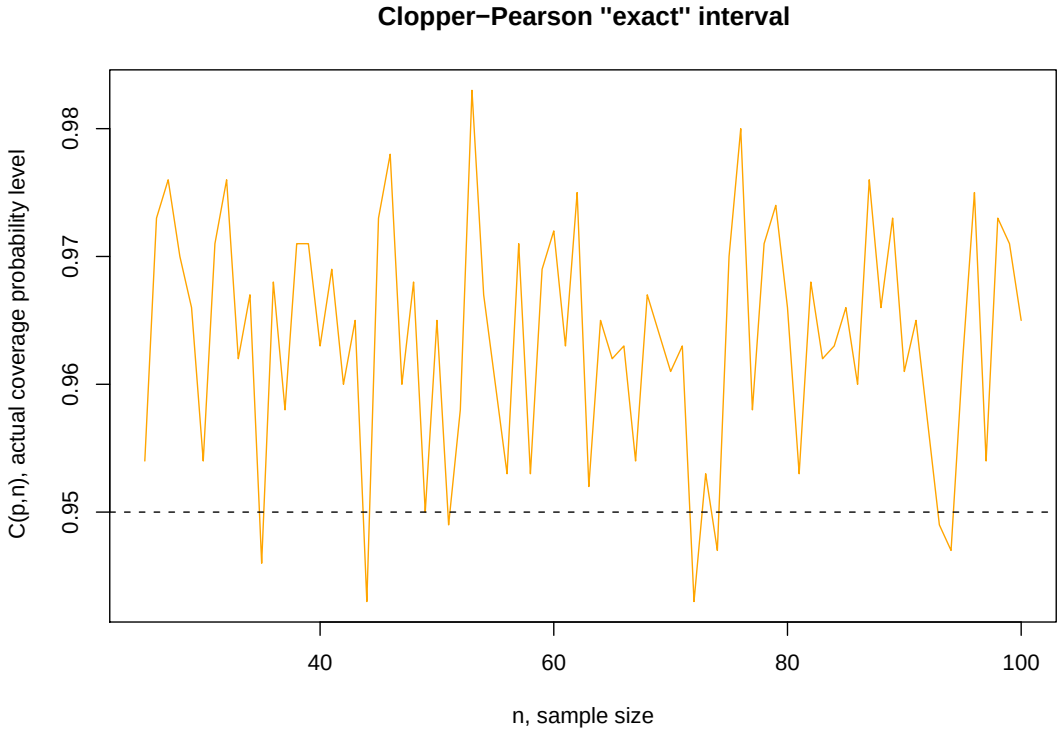


Figure 2.18: Clopper-Pearson "exact" interval: coverage of the nominal 95% standard interval for fixed $\pi = 0.5$ and variable $n = 25$ to 100

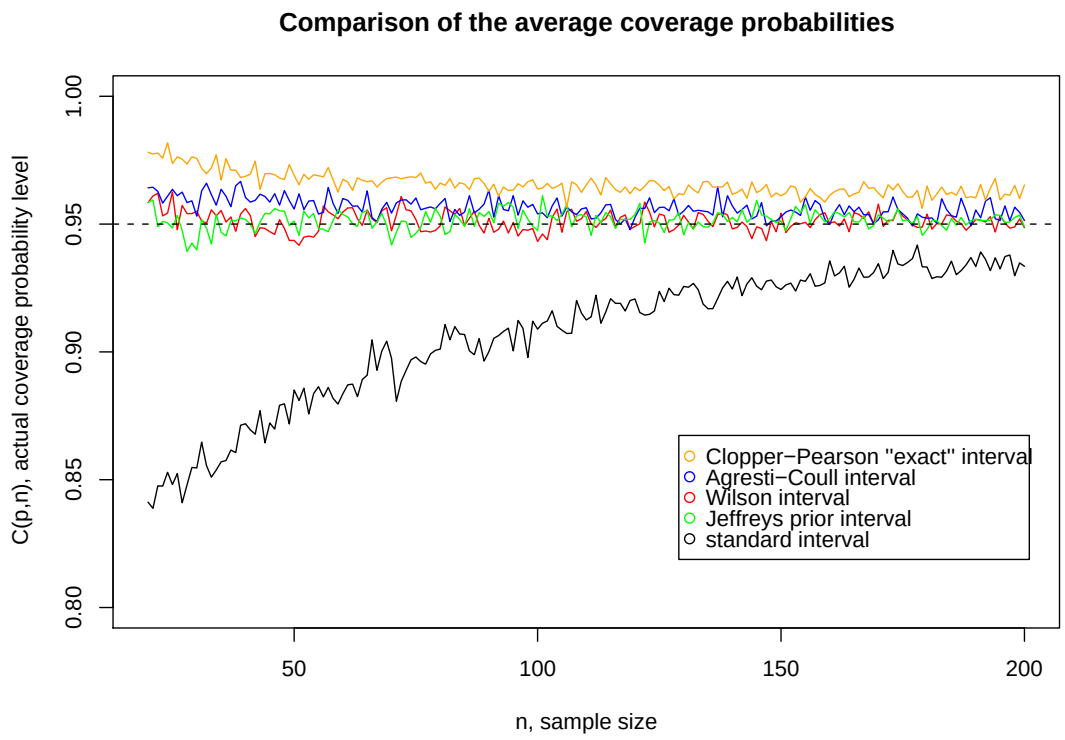


Figure 2.19: Comparison of the average coverage probabilities for the five CI over π, with $1 - \alpha = 0.95$ and $n = 20$ to 200

2.5 Deriving the optimal sample size from the confidence intervals

In subsection 2.3.3 we had remained that we would studied the problem of the poor coverage of the standard interval and some alternative confidence interval for the proportion π. In section 2.4, we have seen how the problems of the standard interval are and the following alternatives: the Wilson interval, the Agresti-Coull interval, the Jeffreys prior interval and the Clopper-Pearson interval.

We have seen previously that the optimal sample size is calculated by the inversion of a confidence interval, looking for that value of n which allows a certain value of the total size of error, chosen *a priori* by the researcher. This total size of error, $A(n, \pi)$ is given by the difference between the upper bound U and the lower bound L of the interval and it is a function of the sample size n and of the probability of success π. Also the value of π is chosen by the researcher:

- if we have some information about the true value of the proportion in the population we are studying, e.g. from previous studies, then we can use that value in order to calculate the optimal sample size;
- if we have not any information about the true value of the proportion in the population we are studying, then we can use the value which maximizes the variance and so maximizes the sample size. Being the proportion π binomial distributed

$$\frac{X}{n} \sim Bin\left(\pi, \frac{\pi(1-\pi)}{n}\right), \tag{2.66}$$

  the variance of X/n is maximized when $\pi = 0.5$.

A method to evaluate the function $A(n, \pi)$ can be the construction of its curves. For example, if we consider the Wilson interval, we have the curves presented in Fig. 2.20.

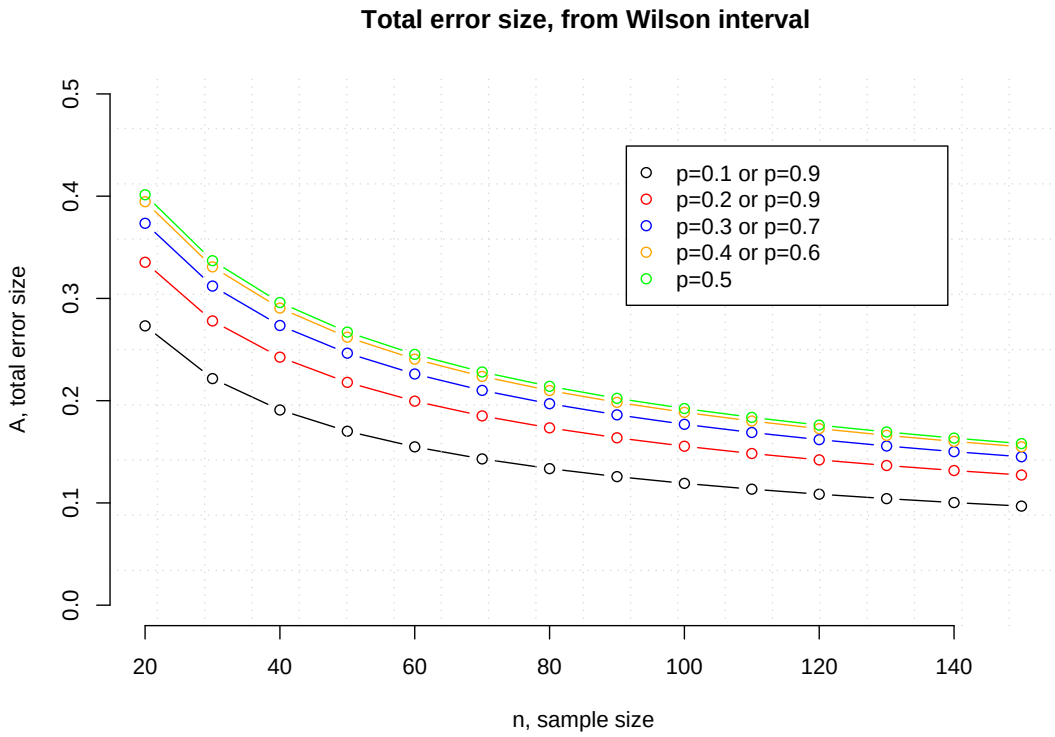


Figure 2.20: Curves of the function $A(n, \pi)$ from the Wilson interval

Chapter 3

Statistical estimation of serological curves

3.1 Modeling the force of infection and the prevalence

As we have seen in the first chapter, the force of infection cannot be estimated directly from cross-sectional data; however, we can obtain it indirectly in accord with the catalytic model, previously introduced.

From 1959, lots of authors have studied the problem of the estimation of the force of infection from cross-sectional seroprevalence data.

Muench [5] suggested to model the infection process with a catalytic model, in which the distribution of the time spent in the susceptible class is exponential with rate β. The force of infection, in this case β, is age independent. Under the catalytic model, we have that:

$$P(a) = \exp\left\{-\int_0^a \beta ds\right\} = e^{-\beta a} \tag{3.1}$$

and

$$f(a) = -\frac{\partial P(a)}{\partial a} = \beta e^{-\beta a}. \tag{3.2}$$

Thus $\ell(a) = \frac{f(a)}{P(a)}$, we finally have that:

$$\ell(a) = \frac{\beta e^{-\beta a}}{e^{-\beta a}} = \beta. \tag{3.3}$$

The prevalence $F(a) = 1 - P(a)$ is:

$$F(a) = 1 - e^{-\beta a}. \tag{3.4}$$

Griffiths [13] proposed a model for Measles in which the force of infection increases linearly in the age range $0 - 10$ (for the author this range includes over 95 per cent of cases). He developed a maximum likelihood method for estimating the parameters of a model in which $\ell(a)$ is assumed to be a linear function of age. The force of infection is:

$$\ell(a) = \beta_0(t + \beta_1), \qquad t > \tau, \tag{3.5}$$

where β_0, β_1 and τ are the parameters of the model. The prevalence is:

$$F(a) = 1 - \exp\left\{\frac{1}{2}\beta_0\{(\tau + \beta_1)^2 - (t + \beta_1)^2\}\right\}, \qquad t > \tau. \tag{3.6}$$

The author shares the subjects in k age classes and then maximizes the following likelihood function (N_j is the number of cases in the jth age class):

$$L = \prod_{j=1}^{k}\left[\frac{F(a_j) - F(a_{j-1})}{F(t_k)}\right]^{N_j}. \tag{3.7}$$

The maximum-likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\tau}$ are found by Newton-Raphson iterations.

Grenfell and Anderson [14] extended the model further and used conventional polynomial functions to model the force of infection. Their model assumes that:

$$P(a) = \exp\left\{\sum_i \beta_i a^i\right\}, \tag{3.8}$$

which implies that the force of infection is:

$$\ell(a) = \sum_i \beta_i i a^{i-1}. \tag{3.9}$$

Other authors, as Farrington [1], Farrington *et al.* [3] and Edmunds *et al.* [15], use

a non-linear model for the prevalence $F(a)$. The problem they want to solve is that the force of infection estimate turns negative if $F(a)$ is a non-monotone function. So, they define a non-negative force of infection, $\ell(a, \beta) \geq 0$ for all a, and estimate $F(a)$ under these constraints using a non-linear model.

Other parametric models, fitted within the framework of generalized linear models (GLM) with binomial error, were discussed by Becker [16], Diamond and McDonald [17] and Keiding *at al.* [18] who used models with complementary log - log link function in order to parameterize the prevalence and the force of infection as a Weibull. Becker [16] suggested to model a piecewise constant force of infection by fitting a model with log link function.

Considering the case where other covariates, in addition to age, are included in the model, Jewell and Van Der Laan [19] proposed, in the context of current status data, a proportional hazard model with constant force of infection which can be fitted as a GLM with complementary log - log link.

Grummer-Strawn [20] discussed two parametric models, the first being a Weibull proportional hazard model with complementary log - log link and the second being a log - logistic model with logit link function. For the latter, the proportionality in the model is interpreted as proportional odds.

A non-parametric method was discussed by Keiding [4] who used isotonic regression to estimate the prevalence and applied kernel smoothers to estimate the force of infection. Keiding *et al.* [18] proposed to model the force of infection using natural cubic splines. Shkedy *et al.* [21] proposed to use local polynomials to estimate both the prevalence and the force of infection.

Shiboski [22] proposed a semiparametric model, based on generalized additive models (GAM) [23], in which the dependency of the force of infection on the age is modelled non-parametrically and the covariate effect is the parametric component of the model. Depending on the link function, the model proposed by Shiboski assumes proportionality: proportional hazard, using the complementary log - log link; proportional odds, using logit and probit links. Other semiparametric models were proposed by Rossini and Tsiatis [24], Martinussen and Scheike [25] and Lin *et al.* [26].

3.2 A parametric model for the force of infection

One of the fundamental studies on the force of infection is Farrington [1]. In his article, Farrington develops a parsimonious parametric model for the seroprevalence and the force of infection for three diseases: measles, mumps and rubella. The hypothesis at the basis of the model are that already introduced in Section 1.3, talking about the catalytic model.

The main purpose in modelling the prevalence $F(a)$ using an underlying "catalytic" model of the force of infection is to study the age dependence of the force of infection. Polynomial models successfully capture the essential features of this age dependence, but are not always consistent with the properties of forces of infection. The only requirement of a consistent model is that the force of infection must be non-negative and must be low (or zero) at birth as a result of persisting maternal antibody.

3.2.1 Farrington's parametric model

Farrington, following preceding studies, notices that serological data for measles, mumps and rubella fit the pattern of an initial near-linear rise in the force of infection (see Griffiths [13]), followed by steady decline. The decline is well fitted by an exponential. This suggests a family of models based on the following exponentially restrained linear model:

$$\ell(a) = (b_1 a - b_3)e^{-b_2 a} + b_4. \tag{3.10}$$

We shall assume furthermore that $\ell(0) = 0$, to account for the protective effect of maternal antibodies at birth, and that $\ell(a)$ eventually decreases with age. It thus follows that $b_4 = b_3$ and $b_2 \geq 0$. Since $\ell(a)$ is always positive, we also have $b_1, b_3 \geq 0$. Farrington's basic model is thus:

$$\ell(a) = (b_1 a - b_3)e^{-b_2 a} + b_3 \qquad b_1, b_2, b_3 \geq 0. \tag{3.11}$$

Combining Eq. 3.11 and the general solution for $F(a)$

$$F(a) = 1 - \exp\left\{-\int_0^t \ell(s)ds\right\}, \tag{3.12}$$

we obtain the following expression for the cumulative distribution of the age at infection:

$$F(a) = 1 - \exp\left\{\frac{b_1}{b_2} a e^{-b_2 a} + \frac{1}{b_2}\left(\frac{b_1}{b_2} - b_3\right)(e^{-b_2 a} - 1) - b_3 a\right\}. \tag{3.13}$$

Note that b_3 is the long term residual value of the force of infection. If b_3 is 0, then the force of infection declines asymptotically to 0.

3.2.2 Application of the model to measles, mumps and rubella

Following Griffiths [13] and Farrington, we now seek to validate the use of the catalytic model and thus of Model. 3.11 by graphical methods.

The data come from Farrington *et al.* [3]: in this paper, there is a table with the numbers seropositive and seronegative for mumps and rubella by completed year of age, from 1 to 44. Besides, this table contains age-stratified seroprevalence data for parvovirus (1991). Differently from Farrington [1], we do not present the analysis for measles, but for parvovirus. The data have been aggregated into 26 age groups, so as to obtain roughly equal numbers in each.

Observed proportions seropositive

Firstly, we report scatterplots of the observed proportions seropositive for mumps, rubella and parvovirus. We can observe two different patterns.

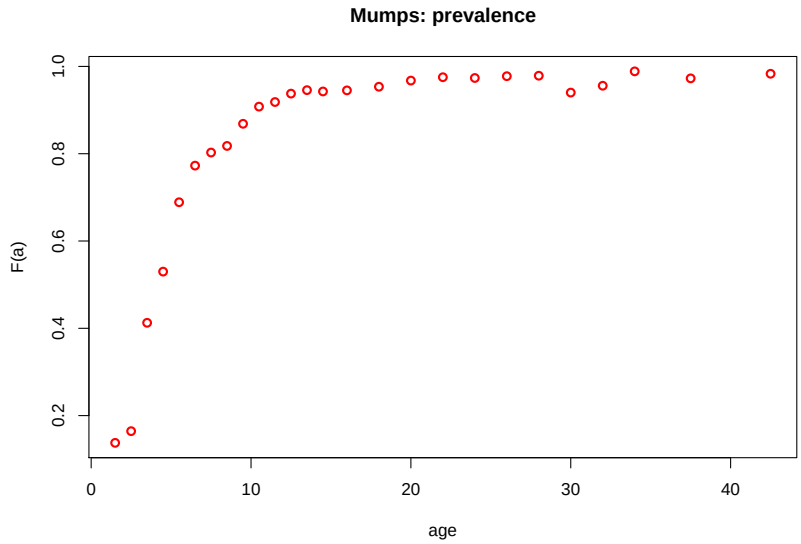
Mumps and rubella in Fig. 3.1 have a similar pattern with the prevalence that increases until 10-12 years old and then becomes steady, with proportions seropositive near 1. Parvovirus shows a different pattern, with the prevalence that increases until 14 years old, then slightly decreases and eventually increases without reaching a steady situation at 44 years old. Besides, here the proportions seropositive are lower than the case of mumps and rubella: the maximum prevalence is 0.82 at 43 years old.

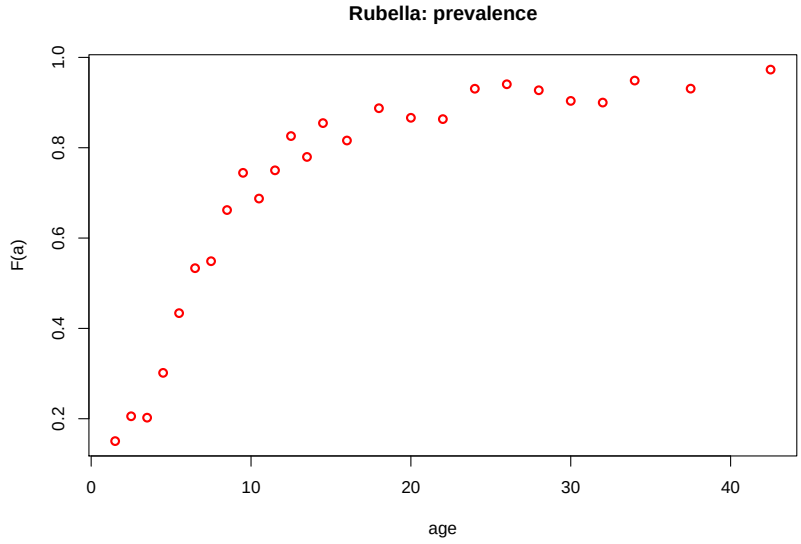
Cumulative hazard function

Then, from Eq. 3.12, we have that the integral of $\ell(a)$, which is the cumulative hazard function, is given by the function:

$$G(a) = \int_0^t \ell(s)\, ds = -\ln[1 - F(a)]. \tag{3.14}$$

In this case also, we have two different patterns.





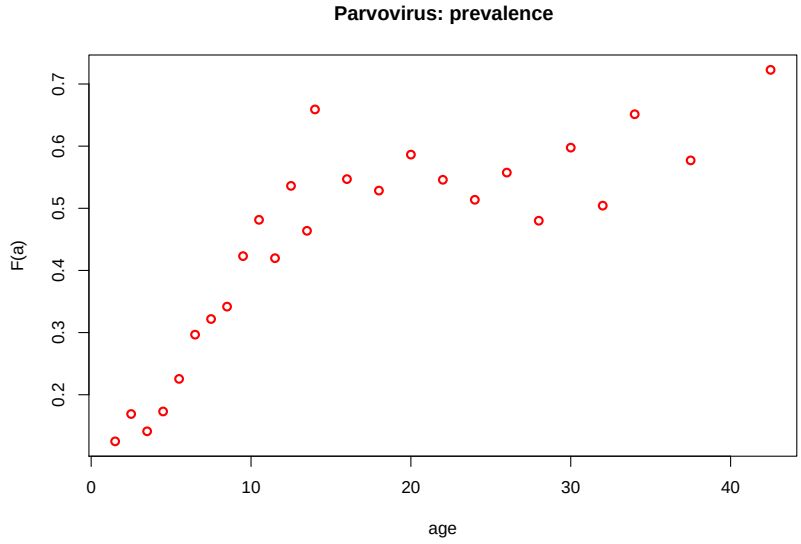
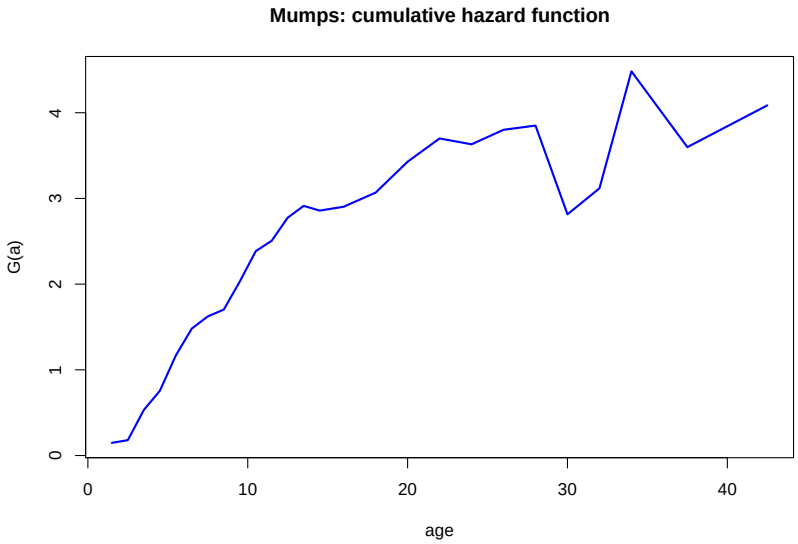
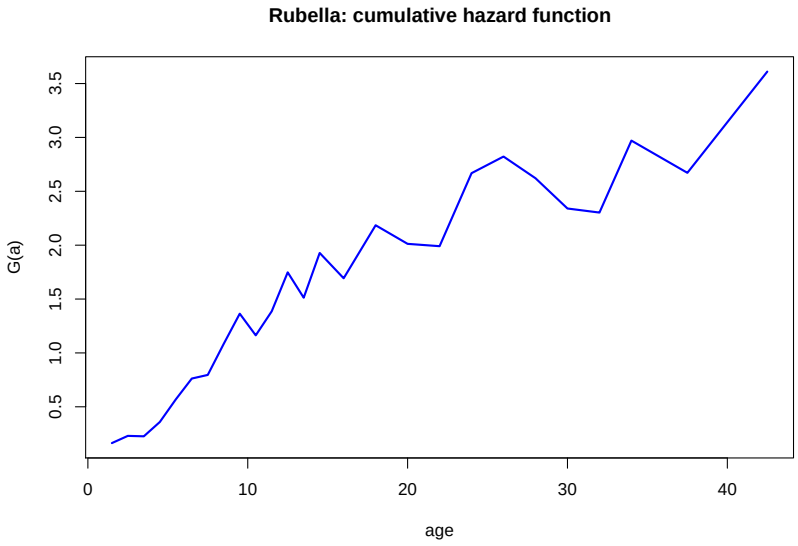


Figure 3.1: Observed proportions seropositive for mumps, rubella and parvovirus; elaboration from Farrington *et al.* [3]





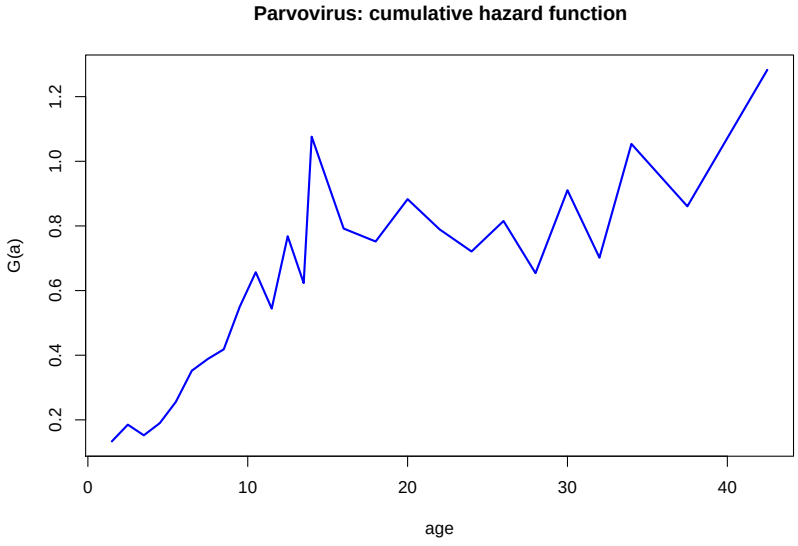


Figure 3.2: Cumulative hazard function for mumps, rubella and parvovirus; elaboration from Farrington *et al.* [3]

Mumps and rubella in Fig. 3.2 have a similar pattern with an initial steepening (consistent with a quadratic for lower values of a and hence linear $\ell(a)$) followed by a flattening out. These two plots suggest that the function $G(a)$ has a single point of inflexion, corresponding to a single maximum for $\ell(a)$; there is too much scatter in the data to allow the detection of any secondary peaks in the force of infection. For large values of a, $G(a)$ is approximately linear, with a shallow slope. If the Model 3.11 is correct, then this asymptotic slope is equal to c. Parvovirus has a similar pattern for lower values of a, while shows an increment in the last years. The graph suggests the presence of two points of inflexion, corresponding the first to a local maximum and the second to a local minimum.

In every case, the considerable scatter of the data, apparent in the saw-tooth appearence of the plots of $G(a)$, is inevitable in any study of age dependence based on current status data from an horizontal survey.

Empirical hazard function

Now, let us plot another function, an approximation of the hazard function:

$$L(a) = \frac{1}{1 - F(a)} \frac{\Delta F(a)}{\Delta a} \approx \frac{1}{1 - F(a)} \frac{\partial F(a)}{\partial a} = \ell(a), \tag{3.15}$$

considering an increment in a of $\Delta a = 1$ (year). Smoothing the observed prevalence data by means of a 3-point moving average, let us evaluate the empirical outlines of the forces of infection.

The empirical hazard function for mumps and for rubella, Fig. 3.3, are all consistent with a steep rise followed by a gradual decline. Differently, the empirical hazard function for parvovirus shows a gradual rise in the first years, followed by a steep decline and by an increment in the last years.

Linearization of the empirical hazard function

Eventually, let us introduce a linearization of the empirical hazard function:

$$R(a) = -\ln\left(\frac{|L(a)|}{a}\right). \tag{3.16}$$

These points should approximately lie on a straight line if the model 3.11 is correct. The absolute value in 3.16 is to deal with the few cases where $L(a)$ in 3.15 is negative: these negative values are due to residual scatter remaining after smoothing. To validate

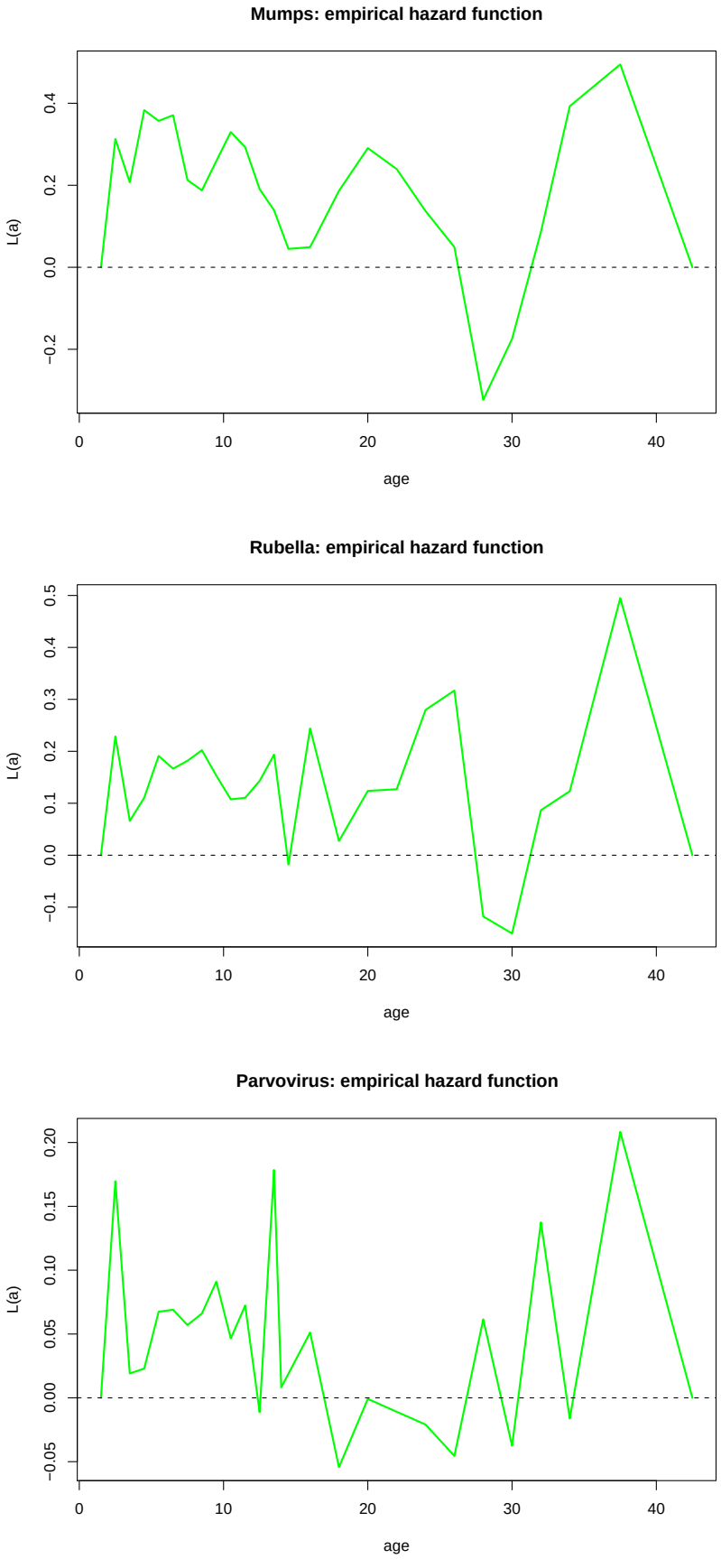


Figure 3.3: Empirical hazard function for mumps, rubella and parvovirus; elaboration from Farrington *et al.* [3]

the assumption of an exponential decline in $\ell(a)$, we shall temporarily assume for simplicity that $b_3 = 0$: this value of b_3 can be taken as its initial guess in the model we are going to estimate.

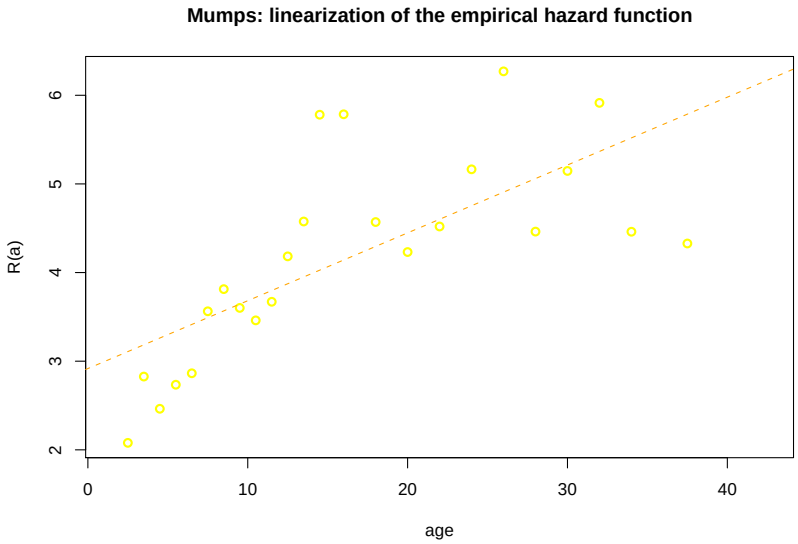
Mumps and rubella in Fig. 3.4 show a broadly linear relationships in spite of some clear outliers and a degree of convexity for the mumps data. Instead, parvovirus shows a more caotic pattern, however in this case also it is possible to find out a broadly linear relationship.

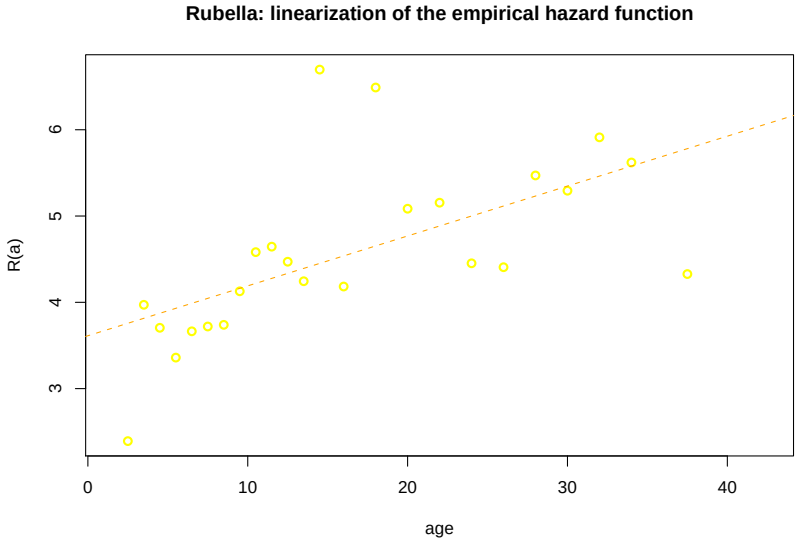
If we regress these points on the age, we can estimate the starting values for parameters b_1 and b_2 in Eq. 4.88, taking in account the assumption $b_3 = 0$:

$$\begin{aligned} R(a) &= -\ln\left(\frac{|L(a)|}{a}\right) \\ &\simeq -\ln\left(\frac{b_1 a e^{-b_2 a}}{a}\right) \\ &\simeq -[\ln(b_1 a) - b_2 a - \ln(a)] \\ &\simeq \ln(a) - \ln(b_1 a) + b_2 a \\ &\simeq \ln\left(\frac{a}{b_1 a}\right) + b_2 a \\ &\simeq -\ln(b_1) + b_2 a. \end{aligned} \tag{3.17}$$

As we have previously told, the function $R(a)$ is linear in the variable "age". If we denote with r_1 and r_2 the parameters of $R(a)$, we have that:

1. $r_1 = -\ln(b_1)$ is the known term, from which the initial guess for b_1 is e^{-r_1};
2. $r_2 = b_2$ is the coefficient of regression of the age and is the initial guess for the parameter b_2.





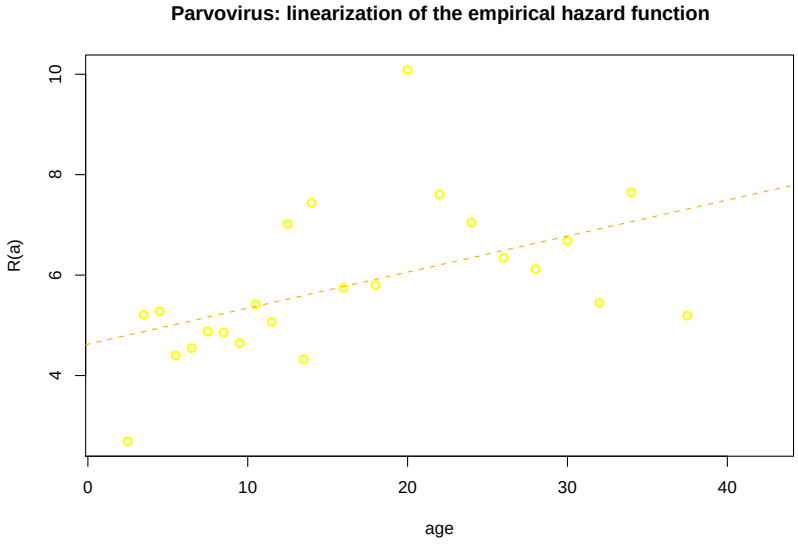


Figure 3.4: Linearization of the empirical hazard function for mumps, rubella and parvovirus; elaboration from Farrington *et al.* [3]

Chapter 4

Nonlinear Estimation Methods

4.1 Least-Squares Estimation

4.1.1 Nonlinear Least Squares

Suppose we have n observations (x_i, y_i) with $i = 1, 2, \ldots, n$, from a fixed-regressor nonlinear model with a known functional relationship f. Thus

$$y_i = f(\mathbf{x}_i; \theta^*) + \varepsilon_i \qquad (i = 1, 2, \ldots, n), \tag{4.1}$$

where $E[\varepsilon_i] = 0$, $\mathbf{x}_i$ is a p x 1 vector (where p is the number of parameters θ) and the true value θ^* of θ, denoted by $\hat{\theta}$, minimizes the error sum of squares

$$S(\theta) = \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i; \theta)]^2. \tag{4.2}$$

It should be noted that, unlike the linear least-squares situation, $S(\theta)$ may have several relative minima in addition to the absolute minimum $\hat{\theta}$.

Assuming that $\varepsilon_i \sim IID(0, \sigma^2)$, it has been shown that, under certain regularity assumptions, $\hat{\theta}$ and $s^2 = S(\hat{\theta})/(n-p)$ are consistent estimates of θ^* and σ^2 respectively.

With further regularity conditions, $\hat{\theta}$ is also asymptotically normally distributed as $n \to \infty$.

If, in addition, we assume that $\varepsilon_i \sim N(0, \sigma^2)$, then $\hat{\theta}$ is also the maximum-likelihood estimator.

When each $f(\mathbf{x}_i; \theta)$ is differentiable with respect to θ, $\hat{\theta}$ will satisfy the following condition:

$$\left.\frac{\partial S(\theta)}{\partial \theta_r}\right|_{\hat{\theta}} = 0 \qquad (r = 1, 2, \ldots, p). \tag{4.3}$$

We shall use the notation $f_i(\theta) = f(\mathbf{x}_i; \theta)$. So the n x 1 vector of the covariates is

$$\mathbf{f}(\theta) = [f_1(\theta), f_2(\theta), \ldots, f_n(\theta)]', \tag{4.4}$$

the gradient vector of $\mathbf{f}(\theta)$ is

$$q(\theta) = \frac{\partial \mathbf{f}(\theta)}{\partial \theta'} = \left[\left(\frac{\partial f_i(\theta)}{\partial \theta_j}\right)\right] \tag{4.5}$$

and the Hessian matrix of $\mathbf{f}(\theta)$ is

$$H(\theta) = \begin{pmatrix} \frac{\partial^2 f_1(\theta)}{\partial^2 \theta_1} & \frac{\partial^2 f_1(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f_1(\theta)}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 f_2(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f_2(\theta)}{\partial^2 \theta_2} & \cdots & \frac{\partial^2 f_2(\theta)}{\partial \theta_p \partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_p(\theta)}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 f_p(\theta)}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 f_p(\theta)}{\partial \theta_2 \partial \theta_n} \end{pmatrix}. \tag{4.6}$$

Also, for brevity, let $q = q(\theta)$ and $\hat{q} = \hat{q}(\theta)$, $H = H(\theta)$ and $\hat{H} = \hat{H}(\theta)$.

Eq. 4.3 leads to

$$\sum_i \{y_i - f_i(\theta)\} \left.\frac{\partial f_i(\theta)}{\partial \theta_r}\right|_{\theta=\hat{\theta}} = 0 \qquad (r = 1, 2, \ldots, p), \tag{4.7}$$

or, using matrices,

$$\begin{aligned} 0 &= \hat{q}'\{\mathbf{y} - \hat{\mathbf{f}}\} \\ &= \hat{q}'\hat{\varepsilon}, \end{aligned} \tag{4.8}$$

The equations 4.8 are called the *normal equations* for the nonlinear model. For most nonlinear models, these equations cannot be solved analytically, so that iterative methods are necessary.

We now have the following theorem:

Theorem 2 *Given $\varepsilon \sim N(0, \sigma^2 I_n)$ and appropriate regularity conditions, then, for large n, we have approximately:*

1. *$(\hat{\theta} - \theta^*) \sim N_p(0, C^{-1})$, where $C = q'q$;*

2. $(n-p)s^2/\sigma^2 \approx \varepsilon'(I_n - P_F)\varepsilon/\sigma^2 \sim \chi^2_{n-p}$, *where* $\varepsilon = \mathbf{y} - q(\theta)$ *and* $P_F = q(q'q)^{-1}q'$;

3. $\hat{\theta}$ *is statistically independent of* s^2;

4.

$$\frac{[S(\theta^*) - S(\hat{\theta})]/p}{S(\hat{\theta})/(n-p)} \approx \frac{\varepsilon' P_F \varepsilon}{\varepsilon'(I_n - P_F)\varepsilon} \cdot \frac{n-p}{p} \sim F_{p,n-p}. \tag{4.9}$$

As we can see, the gradient vector q plays, in nonlinear regression, the same role as the X-matrix in linear regression. This idea is taken further when we shall develop approximate confidence intervals for θ^*.

4.1.2 Generalized Least Squares

We mention now a generalization of the least-squares procedure called *weighted* or *generalized least squares* (GLS). The function to be minimized is

$$S(\theta) = [\mathbf{y} - \mathbf{f}(\theta)]' W^{-1} [\mathbf{y} - \mathbf{f}(\theta)], \tag{4.10}$$

where W is a known positive definite matrix (and in many applications a diagonal matrix). This minimization criterion usually arises from the generalized least-squares model $\mathbf{y} = \mathbf{f}(\theta) + \varepsilon$, where $E[\varepsilon] = 0$ and $V[\varepsilon] = \sigma^2 W$.

Thus the ordinary least squares (OLS) procedure is a special case in which $W = I_n$. Denote by $\hat{\theta_G}$ the generalized least-squares estimate which minimizes $S(\theta)$ above.

Cholesky decomposition

Now, we have to introduce the Cholesky Decomposition of the matrix W to transform our GLS model in a OLS model.

Theorem 3 (Cholesky decomposition) *If* A *is an* n *x* n *positive definite matrix, then there exists an* n *x* n *upper triangular matrix* $U = [(u_{ij})]$ *such that*

$$A = U'U. \tag{4.11}$$

The matrix U is unique if its diagonal elements are all positive or all negative. Let $D_1 = \text{diag}(u_{11}, u_{22}, \ldots, u_{nn})$, and let

$$U_1 = D_1^{-1}U = \begin{bmatrix} 1 & \tilde{u_{12}} & \tilde{u_{13}} & \dots & \tilde{u_1} \\ 0 & 1 & \tilde{u_{23}} & \dots & \tilde{u_{2n}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix}. \tag{4.12}$$

Then

$$A = U_1' D_1^2 U_1 = U_1' D U_1, \tag{4.13}$$

where D is a diagonal matrix with positive elements (and so D is a positive definite matrix similar to A).

Logically, $U'AU = I_n$.

Let $W = U'U$ be the Cholesky decomposition of W, where U is an upper triangular matrix. Multiplying the nonlinear model through by $R = (U')^{-1}$, we obtain

$$\mathbf{z} = \mathbf{k}(\theta) + \eta, \tag{4.14}$$

where $\mathbf{z} = R\mathbf{y}$, $\mathbf{k}(\theta) = R\mathbf{f}(\theta)$ and $\eta = R\varepsilon$.

Then $E[\eta] = 0$ and $V[\eta] = \sigma^2 RWR' = \sigma^2 I_n$. Thus our original GLS model has now been transformed to an OLS model.

Furthermore,

$$\begin{aligned} S(\theta) &= [\mathbf{y} - \mathbf{f}(\theta)]' W^{-1} [\mathbf{y} - \mathbf{f}(\theta)] \\ &= [\mathbf{y} - \mathbf{f}(\theta)]' R' R [\mathbf{y} - \mathbf{f}(\theta)] \\ &= [\mathbf{z} - \mathbf{k}(\theta)]' [\mathbf{z} - \mathbf{k}(\theta)]. \end{aligned} \tag{4.15}$$

Hence the GLS sum of squares is the same as the OLS sum of squares for the transformed model and $\hat{\theta}_G$ is the OLS estimate from the transformed model.

Besides, if $\hat{\theta}_G$ is the OLS estimate from the transformed model, it has for large n a variance-covariance matrice, whose estimate is given by

$$\hat{V}[\hat{\theta}_G] = \hat{\sigma}^2 [(\hat{K}.(\theta^*))'(\hat{K}.(\theta^*))]^{-1} = \hat{\sigma}^2 [q(\theta^*)' W^{-1} q(\theta^*)]^{-1}, \tag{4.16}$$

where

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-p}[\mathbf{z}-\mathbf{k}(\hat{\theta}_{\mathbf{G}})]'[\mathbf{z}-\mathbf{k}(\hat{\theta}_{\mathbf{G}})] \\ &= \frac{1}{n-p}[\mathbf{y}-\mathbf{f}(\hat{\theta}_{\mathbf{G}})]'W^{-1}[\mathbf{y}-\mathbf{f}(\hat{\theta}_{\mathbf{G}})]. \end{aligned} \tag{4.17}$$

However, in practice we would not compute $R = (U')^{-1}$ and multiply out $R\mathbf{y}$. Instead it is better to solve the lower triangular system $U'\mathbf{z} = \mathbf{y}$ for $\mathbf{z}$ directly by forward substitution.

4.2 Maximum-Likelihood Estimation

If the joint distribution of the ε_i in the model 4.20 is assumed known, then the maximum-likelihood estimate of θ is obtained by maximizing the likelihood function. We shall discuss normally distributed errors below. In this case we find that the maximum-likelihood estimator of θ can be found using least-squares methods.

4.2.1 Normal Errors

If $\varepsilon_i \sim N(0, \sigma^2)$, then the joint distribution of ε_i is

$$p(\mathbf{y}|\theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{[y_i - f(\mathbf{x}_i;\theta)]^2}{\sigma^2}\right). \tag{4.18}$$

Ignoring constants, we denote the logarithm of the above likelihood by $L(\theta, \sigma^2)$ and obtain

$$\begin{aligned} L(\theta, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i;\theta)]^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\theta). \end{aligned} \tag{4.19}$$

This trasformation of $p(\mathbf{y}|\theta, \sigma^2)$, taking its logarithm, is possible because the logarithm is a monotonic increasing function and so the maximum of $\log f(\cdot)$ is the same of $f(\cdot)$. The aim of this transformation is to linearize the function $p(\mathbf{y}|\theta, \sigma^2)$, to obtain a function which is easier to manage.

Given σ^2, Eq. 4.19 is maximized with respect to θ when $S(\theta)$ is minimized, that is, when $\theta = \hat{\theta}$ (the least-squares estimate).

Furthermore, the first derivative of $L(\theta, \sigma^2)$ with respect to σ^2, that is $\partial L(\theta, \sigma^2)/\partial\sigma^2 = 0$ has solution $\sigma^2 = S(\theta)/n$, which gives a maximum (for given θ) as the second derivative is negative. This suggests that $\hat{\theta}$ and $\sigma^2 = S(\theta)/n$ are the maximum-likelihood estimates.

4.3 Asymptotic Confidence Intervals

Let

$$y_i = f(\mathbf{x}_i; \theta^*) + \varepsilon_i \qquad (i = 1, 2, \ldots, n), \tag{4.20}$$

where $\varepsilon_i \sim N(0, \sigma^2)$. For notational convenenience we now omit the star from θ^* and denote the true value of this parameter vector by θ.

Confidence interval for a linear combination

If we are interested in constructing a confidence interval for a given linear combination of parameters $\mathbf{a}'\theta$, then we can apply Theorem 2 in Section 4.1.1. In particular we have the asymptotic (linearization) result

$$\hat{\theta} \sim N_p(\theta, C^{-1}) \qquad C = q'q, \tag{4.21}$$

which holds under appropriate regularity conditions.

From Eq. 4.21 we have, asymptotically, $\mathbf{a}'\hat{\theta} \sim N(\mathbf{a}'\theta, \mathbf{a}'C^{-1}\mathbf{a})$ independently of $s^2 = \|\mathbf{y} - \mathbf{f}(\hat{\theta})\|^2/(n-p)$, the latter being an unbiased estimate of σ^2 to order n^{-1}. Hence, for large n we have, approximately,

$$T = \frac{\mathbf{a}'\hat{\theta} - \mathbf{a}'\theta}{s(\mathbf{a}'C^{-1}\mathbf{a})^{1/2}} \sim t_{n-p}, \tag{4.22}$$

where t_{n-p} is the t-distribution with $n-p$ degrees of freedom. So, an approximate $100(1-\alpha)\%$ confidence interval for $\mathbf{a}'\theta$ is then

$$\mathbf{a}'\hat{\theta} \pm t_{n-p}^{\alpha/2} s(\mathbf{a}'C^{-1}\mathbf{a})^{1/2}, \tag{4.23}$$

where C can be estimated by $\hat{C} = \hat{q}'\hat{q}$.

Confidence interval for a single parameter

Setting $\mathbf{a}' = (0, 0, \ldots, 1, 0, \ldots, 0)$, where the rth element of $\mathbf{a}$ is one and the remaining elements are zero, and defining $[\hat{c}^{rs}] = \hat{C}^{-1}$, a confidence interval for the rth elements of θ, that is θ_r is

$$\theta_r \pm t_{n-p}^{\alpha/2} s \sqrt{\hat{c}^{rr}}. \tag{4.24}$$

4.4 Computation of the Estimates

In this section, we introduce the problem of *unconstrained optimization*, where the unknown parameters are free to assume any values at all. In other cases, only values satisfying certain inequalities and/or equations are admissible. The problem is to find θ such that $\Phi(\theta)$ is maximum (or minimum) subject to:

$$\begin{cases} h(\theta) \geq 0 \\ g(\theta) = 0 \end{cases}$$

where h and g are vectors of given functions.

Typical functions to be minimized are the sum of squares and the weighted sum of squares; a function to be maximized is the likelihood. We restrict our attention to minimization, for maximizing a function can be accomplished by minimizing its negative.

So, if our aim is to maximize the log likelihood, we reach the same result minimizing the negative of log likelihood.

4.4.1 Iterative Scheme

The methods we shall discuss are *iterative* in nature. We start with a given point θ_0, known as the *starting value*, and proceed to generate a sequence of points $\theta_1, \theta_2, \ldots$ which we hope converges to the point θ^* at which the function $\Phi(\theta)$ is minimum. The computation of θ_{i+1} is called the ith *iteration* and the point θ_i the ith *iterate*. In practice, one *terminates* the sequence after a finite number N of iterations and one accepts θ_N as an approximation to θ^*. The vector

$$\sigma_i \equiv \theta_{i+1} - \theta_i \tag{4.25}$$

is called the ith *step*. We wish each step to bring us closer to minimum. Since we do not know where the minimum is, we cannot test for this condition directly. However we

may consider the ith step to have "improved" out situation if

$$\Phi(\theta_{i-1}) < \Phi(\theta_i). \tag{4.26}$$

If the previous condition is verified, the ith step is *acceptable*. An iterative method is *acceptable* if all the steps are acceptable. All the methods we shall discuss represent in detail the following scheme:

1. Set $i = 0$. A starting value θ_0 must be provided externally.

2. Determine a vector d_i in the direction of the proposed ith step.

3. Determine a scalar t_i such that the step

$$\sigma_i = t_i d_i \tag{4.27}$$

   is acceptable. That is, we take

$$\theta_{i+1} = \theta_i + \sigma_i = \theta_i + t_i d_i \tag{4.28}$$

   and require that t_i be chosen so that Eq. 4.26 holds.

4. Test whether the termination criterion is met. If not, increase i by one and return to step 2. If yes, accept θ_{i+1} as the value of θ^*.

The various methods to be described below differ only in the manner of choosing d_i and t_i. We refer to these quantities as *step direction* and *step size* respectively. Since d_i is not required to be a unit vector, t_i is only proportional, but not necessarily equal, to the step length in the usual sense.

4.4.2 Acceptability

Consider the ith iteration of a minimization procedure. Suppose we move from θ_i along some direction d, generating the ray

$$\theta(t) \equiv \theta_i + td \qquad (t \geq 0). \tag{4.29}$$

Along this ray, the objective function to be minimized varies as t is changed, thus becoming a function of t alone. We designate this function

$$\Psi_{id}(t) \equiv \Phi(\theta(t)) = \Phi(\theta_i + td), \tag{4.30}$$

where $\Psi_{id}(t)$ indicates that the function $\Phi(\theta)$ to be minimized is dependent on the step size t. The derivative of $\Psi_{id}(t)$ is given by

$$\frac{d\Psi_{id}}{dt} = \left(\frac{\partial\Phi}{\partial\theta}\right)^T \left(\frac{\partial\theta}{\partial t}\right) = \left(\frac{\partial\Phi}{\partial\theta}\right)^T d, \tag{4.31}$$

that is to say the vector $(\partial\theta/\partial t)$ is the step direction d.
The *gradient vector* of $\Phi(\theta)$ is $(\partial\Phi/\partial\theta)$, which we designate as $q(\theta)$. Denoting by q_i the gradient vector evaluated at $\theta = \theta_i$, we have

$$\Psi'_{id} \equiv \left.\frac{d\Psi_{id}}{dt}\right|_{t\to 0} = q_i^T d. \tag{4.32}$$

In the sequel we assume $q_i \neq 0$. The quantity Ψ'_{id} is called the *directional derivative* of Φ relative to d at θ_i.

Let us see the sign of Ψ'_{id}. On the one hand, if $\Psi'_{id} < 0$, then $\Phi(\theta)$ decreases in value when one starts moving away from θ_i in the direction of d. Therefore, if t is a sufficiently small positive number, the step td is acceptable.

On the other hand, if $\Psi'_{id} \geq 0$, there may not exist any positive value of t for which td is an acceptable step.

Thus, we call d an *acceptable direction* if $\Psi'_{id} < 0$.

Let us see now the following theorem:

Theorem 4 *A direction d is acceptable if and only if there exist a positive definite matrix R such that*

$$d = -Rq_i. \tag{4.33}$$

Proof

Let R be a positive definite matrix and let d be given by Eq. 4.33. Then, from Eq. 4.32 and the definition of a positive definiteness

$$\Psi'_{id} = q_i^T d = -q_i^T R q_i < 0, \tag{4.34}$$

as a result of $q_i^T R q_i > 0$.

The requirement $\Psi'_{id} = q_i^T d < 0$ says that the direction d leads downhill if it forms a greater than 90° angle with the gradient q_i. The theorem states that this condition can be insured if the direction is determined by operating on the negative gradient q_i with a positive definite matrix according to the condition $d = -Rq_i$. A minimization method in which the directions are obteined in this manner is called an *acceptable gradient method*.

The basic equation of the ith iteration in any gradient method is

$$\theta_{i+1} = \theta_i - t_i R_i q_i. \tag{4.35}$$

Various gradient methods differ in the manner of choosing the R_i and t_i.

In planning or choosing an optimization method, one attempts to minimize the total computation time required for convergence to the minimum. This time is composed primarily of the following two factors:

- function and derivative evaluations;
- algebraic manipulations such as matrix inversions or eigenvalue determinations.

4.4.3 Steepest Descent

The simplest gradient method employs the following conditions:

- $R_i = I$, where I is the identity matrix;
- $d_i = -q_i$ in all iterations;
- t_i is the solution of the problem of minimization

$$\min_t(\Psi_{id}) = \Phi(\theta_i + t d_i). \tag{4.36}$$

So, the equation of the ith iteration is

$$\theta_{i+1} = \theta_i + t_i q_i. \tag{4.37}$$

The direction $-q_i$ is the one in which the objective function decreases most rapidly, at least initially. Hence this method is called *steepest descent*. Unfortunately this method is often very inefficient, requiring a large number of steps which tend to zigzag.

This method is not recommended for practical applications, but it is important because a lot of algorithms used in practice have this method as the theoretical starting point.

4.4.4 The Newton-Raphson method

Suppose the function $\Phi(\theta)$ we want to minimize is the negative of the log-likelihood of a sample. The Newton-Raphson method employs the following conditions:

- $R_i = H_i^{-1}$, where H_i is the Hessian matrix of the function $\Phi(\theta)$;
- $t_i = 1$.

So, the equation of the ith step is

$$\theta_{i+1} = \theta_i - H_i^{-1} q_i. \tag{4.38}$$

Let us see why we apply these conditions. The Hessian matrix of the function $\Phi(\theta)$ is the matrix of second partial derivatives. For example, if the parameters to be estimated are only two, the Hessian matrix will be:

$$H(\theta) = \begin{pmatrix} \frac{\partial^2 \Phi}{\partial^2 \theta_1} & \frac{\partial^2 \Phi}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \Phi}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \Phi}{\partial^2 \theta_2} \end{pmatrix}$$

Let H_i be the Hessian matrix of Φ evaluated at $\theta = \theta_i$. We define the function

$$Q_i(\theta) = \Phi_i + q_i^T(\theta - \theta_i) + \frac{1}{2}(\theta - \theta_i)^T H_i(\theta - \theta_i). \tag{4.39}$$

If $\Phi(\theta) = -L(\theta)$, that is the negative of the log likelihood, the function $Q_i(\theta)$ will be:

$$Q_i(\theta) = -L(\theta_i) + L'(\theta_i)^T(\theta - \theta_i) + \frac{1}{2}(\theta - \theta_i)^T L''(\theta_i)(\theta - \theta_i). \tag{4.40}$$

This function consists of the terms up to second order in the Taylor expansion of Φ around the point θ_i. In a sense, $Q_i(\theta)$ shows the behaviour of $\Phi(\theta)$ at $\theta = \theta_i$ more closely than does any other second order surface.

Suppose we wish to find the point at which $Q_i(\theta)$ is stationary. We set to 0 the gradient of Q_i:

$$\frac{\partial Q_i}{\partial \theta} = q_i + H_i(\theta - \theta_i) = 0; \tag{4.41}$$

if H_i is nonsingular, we have the following solution:

$$\theta_{i+1} = \theta_i - H_i^{-1} q_i \tag{4.42}$$

And so, we have verified the conditions we have introduced at the beginning of this section.

$\Phi(\theta)$ as a quadratic function

If $\Phi(\theta)$ is a quadratic function, then θ_{i+1} is a stationary point of Φ. This point is a minimum if H_i is positive definite, for the second order sufficient condition: in this case $R_i = H_i^{-1}$ is positive definite, the method is acceptable and it converges in a single iteration. In this case the Newton step $-H^{-1}q$ is the only one that take us to the minimum in a single iteration. Any other step $-Rq$ with $R \neq H^{-1}$ will miss the minimum. If H_i is negative definite, θ_{i+1} is a maximum, and if H_i is indefinite, θ_{i+1} is a saddle point: in both cases, the method is not acceptable.

$\Phi(\theta)$ as a non quadratic function

When $\Phi(\theta)$ is not quadratic, θ_{i+1} does not generally coincide with the stationary point and the method does not converge in a single iteration. However the method is acceptable as long as H_i is positive definite, as it should be at least in some neighborhood of the minimum. In this neighborhood, convergence is quadratic. This means that the number of correct digits in θ is approximately doubled by each iteration, until further improvement is barred by the rounding errors in the calculation.

Outside the neighborhood, the convergence cannot be guaranteed.

In spite of its very good performance in those cases where it works, the Newton-Raphson method is not a practical one, for the following reasons:

1. It does not converge in many cases, because $H(\theta)$ is not necessarily positive definite, except near the minimum.

2. It requires the evaluation of second derivatives. This places a heavy burden on the user, particularly where the objective functions are as complicated as those to be found in parameter estimation problems.

Various modifications have been proposed for overcoming these difficulties, while retaining the advantage of the method:

1. To overcome the problem of indefiniteness, it has been designed the Levenberg-Marquardt method;

2. To overcome the difficulties derived from the evaluation of second derivatives, the proposed method is the Gauss-Newton's.

We note that of the two deficiencies, the first one (nonconvergence) *must* be overcome if the method is to be useful. The second difficulty (second derivatives required) is merely a matter of convenience.
One may raise the question of whether the Newton-Raphson method is not so much more efficient than methods that do not require second derivatives, as to make the evaluation of these derivatives worthwhile. Authors have no definitive answers to this question, but a limited amount of experience has led to the following tentative conclusions:

1. If the model fits the data well, the Gauss-Newton method often requires no more iterations than the Newton-Raphson one [27].

2. If the model does not fit well, the Newton-Raphson method may require fewer iterations than the Gauss-Newton method, but the computing times for the two methods are roughly the same.

4.4.5 The Levenberg-Marquardt Method

In the Newton-Raphson method we have a problem if the Hessian matrix H_i is not positive definite. The Levenberg-Marquardt method allows to convert an arbitrary matrix into a positive definite one.
The method is based on the observation that if P is any positive definite matrix, then $H_i + \kappa P$ is positive definite for sufficiently large κ, no matter what H_i. The authors suggest the following choice:

$$P_i \equiv \begin{cases} |H_{ii}| & (H_{ii} \neq 0) \\ 1 & (H_{ii} = 0) \end{cases}$$

That is, P is a sort of diagonal matrix whose elements coincide with the absolute values of the diagonal elements of H_i (with say zero elements replaces by ones).

The Levenberg-Marquardt method employs the following conditions:

- $R_i = (H_i + \kappa_i P_i)^{-1}$, where H_i is the Hessian matrix of the function $\Phi(\theta)$;
- $t_i = 1$.

So, the equation of the ith step is

$$\theta_{i+1} = \theta_i - (H_i + \kappa_i P_i)^{-1} q_i. \tag{4.43}$$

Observe that as $\kappa_i \to \infty$, the term $\kappa_i P_i$ dominates H_i. In this case the step σ_i becomes

$$\sigma_i \to -\kappa_i^{-1} P_i^{-1} q_i. \tag{4.44}$$

This is an extremely short step in a downhill direction, P_i being positive definite. A sufficiently large κ_i always produces an acceptable step. On the other hand, when κ_i is very small, σ_i approaches the Newton-Raphson direction $-H_i^{-1} q_i$. Marquardt suggests the following algorithm for the selection of κ_i:

1. When $i = 0$, start with $\kappa_0 = 0.01$.

2. At the start of the ith iteration, compute

   - $d_i = -(H_i + \kappa_i P_i)^{-1} q_i$,
   - $\theta^{(1)} = \theta_i + d_i$,
   - $\Phi_{(1)} \equiv \Phi(\theta_{(1)})$.

3. If $\Phi_{(1)} < \Phi_i$, accept $\theta_{i+1} = \theta^{(1)}$, and replace κ_{i+1} with $\max(0.1\kappa, \varepsilon)$ where ε is a small positive number, say 10^{-7}.

4. Otherwise, if $\Phi_{(1)} \geq \Phi_i$, find a value t_i sufficiently small so that

   $$\Phi(\theta_i + t_i d_i) < \Phi(\theta_i)$$

   Accept $\theta_{i+1} = \theta_i + t_i d_i$. Replace κ_{i+1} with $10\kappa_i$.

It is worth remarking that Marquardt's method finds the step d which minimizes the quadratic approximation to Φ given by

$$Q_i(d) \equiv \Phi_i + d^T q_i + \frac{1}{2} d^T H_i d \tag{4.45}$$

subject to the restriction that

$$d^T P_i d = c \tag{4.46}$$

That is, the step d takes us to the point on the ellipsoid defined by Eq. 4.46 at which the function $Q(d)$ reach its minimum. To prove this, by the Lagrangian multipliers we get:

$$\Lambda(\mathbf{d}) = \Phi_i + d^T q_i + \frac{1}{2} d^T H_i d + \frac{1}{2} \lambda_i (d^T P_i d - c) \tag{4.47}$$

Differentiating with respect to d and equating to zero we have

$$q_i + H_i d + \lambda_i P_i d = 0 \tag{4.48}$$

Then solving for d we obtain

$$d = -(H_i + \lambda_i P_i)^{-1} q_i \tag{4.49}$$

in agreement with $R_i = (H_i + \kappa_i P_i)^{-1}$.

The particular ellipsoid chosen depends on λ_i, since by substituting Eq. 4.49 into Eq. 4.46 we find

$$c = q_i^T (H_i + \lambda_i P_i)^{-1} P_i (H_i + \lambda_i P_i)^{-1} q_i \tag{4.50}$$

The larger λ_i is, the smaller is c, and the smaller is the ellipsoid of a certain size, determined through Eq. 4.50 by the initial choice of λ_i. If the corresponding step d fails to decrease the objective function, this is an indication that the chosen ellipsoid is larger than the region within which the quadratic approximation (Eq. 4.45) holds. By increasing λ, we shrink the ellipsoid and we go on.

4.4.6 The Gauss-Newton Method

In most parameter estimation problems, the unknown parameters appear only indirectly in the objective function. This depends explicitly on the model equations, which in turn depend on the parameters. To compute derivatives of the object function, we first differentiate it with respect to the model equation, and then differentiate those with respect to the parameters. The Gauss Method consists of simply omitting the second derivatives of the model equation when the Hessian is been computed. We illustrate it better with a practical example. We want to minimize:

$$\begin{aligned} \Phi(\theta) &= \sum_{\mu=1}^{n} [y_\mu - f(x_\mu, \theta)]^2 \\ &= \sum_{\mu=1}^{n} (y_\mu - f_\mu)^2 \\ &= \sum_{\mu=1}^{n} e_\mu^2. \end{aligned} \tag{4.51}$$

We now compute q_α and $H_{\alpha,\beta}$ simply differentiating:

$$q_\alpha = \partial\Phi/\partial\theta_\alpha = 2\sum_{\mu=1}^{n} e_\mu \partial e_\mu/\partial\theta_\alpha \tag{4.52}$$

$$= -2\sum_{\mu=1}^{n} e_\mu \partial f_\mu/\partial\theta_\alpha. \tag{4.53}$$

and

$$H_{\alpha,\beta} = \partial^2\Phi/\partial\theta_\alpha\partial\theta_\beta = -2\sum_{\mu=1}^{n} e_\mu(\partial^2 f_\mu/\partial\theta_\alpha\partial\theta_\beta) + 2\sum_{\mu=1}^{n}(\partial f_\mu/\partial\theta_\alpha)(\partial f_\mu/\partial\theta_\beta) \tag{4.54}$$

In the Gauss method, we neglect the first term, and use N in place of H, where N is defined by

$$N_{\alpha,\beta} = 2\sum_{\mu=1}^{n}(\partial f_\mu/\partial\theta_\alpha)(\partial f_\mu/\partial\theta_\beta) \tag{4.55}$$

So we derive the matrix N as an approximation to H and the Gauss method as an approximation to the Newton method.

But there is an alternative interpretation: suppose to replace the model equation with their tangents, that is the *nonlinear model (in θ)* is approximated by one that is linear. If we solve the corresponding linear least squares problem we find the solution to be

$$\tilde{\theta} = \theta_i - N^{-1}q$$

It is worth remarking that $\tilde{\theta}$ is not the correct solution to the *non-linear* problem.

The term neglected in Eq. 4.54 contained the residual e_μ as a factor. Since the residual are, hopefully, small, this provides some justification to consider N as a good approximation to H, particularly near the minimum. The same justification applies to all of the more general cases in which the objective function depends on the parameter only through the elements of the matrix of the residuals

$$M(\theta) = \sum_{\mu} e_\mu(\theta) e_\mu^T(\theta) \tag{4.56}$$

In this case we have

$$\Phi(\theta) = \Psi(M(\theta)) \tag{4.57}$$

where Ψ is a suitable function. Differentiating Eq. 4.57 we obtain:

$$\partial M/\partial\theta_\alpha = \sum_\mu (e_\alpha \partial e_\beta/\partial\theta_\alpha + e_\beta \partial e_\alpha/\partial\theta_\alpha) \tag{4.58}$$

where the subscript μ has been dropped for convenience. Therefore from Eq. 4.54 and because of the symmetry of M we obtain the expression for the gradient:

$$\begin{aligned} q_\alpha = \partial\Phi/\partial\theta_\alpha &= \sum (\partial\Psi/\partial M)(\partial M/\partial\theta_\alpha) && (4.59)\\ &= \sum (\partial\Psi/\partial M)(e_\alpha \partial e_\beta/\partial\theta_\alpha + e_\beta \partial e_\alpha/\partial\theta_\alpha) && (4.60)\\ &= 2\sum (\partial\Psi/\partial M) e_\alpha (\partial e_\beta \partial\theta_\alpha) && (4.61) \end{aligned}$$

and the Hessian:

$$H_{\alpha,\beta} = \partial^2\Phi/\partial\theta_\alpha\partial\theta\beta = 2\sum (\partial\Psi/\partial M)(\partial e_\alpha/\partial\theta\alpha)(\partial e_\beta/\partial\theta\beta) + \xi \tag{4.62}$$

where ξ contains second derivatives terms of the model equation and terms involving residual as factors. As we noted earlier, these terms are dropped in the Gauss method. This leaves us with the approximate Hessian:

$$N_{\alpha,\beta} \equiv \partial^2\Phi/\partial\theta_\alpha\partial\theta\beta = 2\sum (\partial\Psi/\partial M)(\partial e_\alpha/\partial\theta\alpha)(\partial e_\beta/\partial\theta\beta) \tag{4.63}$$

or, in matrix notation

$$N \equiv 2\sum_\mu B^T \Gamma B \tag{4.64}$$

where B and Γ are defined as $B \equiv -\partial e/\partial\theta = \partial f/\partial\theta$ and $\Gamma \equiv \partial\Psi/\partial M$.

Using the same notation the gradient in Eq. 4.59 come out:

$$q = -2\sum_\mu B^T \Gamma e \tag{4.65}$$

It is significant that in all these cases Γ turns out to be positive definite (or at least semidefinite).

Resuming, the Gauss-Newton method employs the following conditions:

- $R_i = N^{-1}$, where N_i is the approximation to the Hessian matrix of the function $\Phi(\theta)$;

- $t_i = 1$.

So, the equation of the ith step is

$$\theta_{i+1} = \theta_i - d_i = \theta_i - N^{-1} q_i, \tag{4.66}$$

with d_i is the solution of the following linear system:

$$N_i d_i = -q_i. \tag{4.67}$$

From Eq. 4.65 and Eq 4.64 it can be written as

$$\sum_{\mu} B^T \Gamma B d = \sum_{\mu} B^T \Gamma e \tag{4.68}$$

Pre-multiplying for $(B^T\Gamma)^{-1}$ we obtain

$$Bd = e$$

so we can easily solve for d and determine the step direction.

Gauss Method Implementation

There are several ways in which the direction d_i given by $d_i = -N_i^{-1} q_i$ may be computed. Any method suitable for the solution of multiple linear regression can be used. For linear problems we expect to obtain the correct answer in a single step and to compute N^{-1} very precisely. A non linear problem, on the other hand, requires several iterations; slight errors in each iteration can be tolerated, as long as the chosen directions are acceptable.

In other words, N_i^{-1} need in principle to be positive definite only for nonlinear problems. However, substantial errors in the computation of N_i^{-1} may greatly increase the number of iterations required.

Numerical techniques for computing the direction d_i fall into two classes:

1. Methods for solving the normal equations, without taking account of their particular structure. These methods are obviously applicable whether or not the equations have a linear regression structure. The simplest method solve $Nd_i = -q_i$ for d_i using simultaneous equations techniques. The fastest method is the Cholesky de-

composition, but is not recommended unless N_i is known to be positive definite[1]. In general the Levenberg-Marquardt method is recommended.

2. Methods that rely on the linear regression structure. For example Jennrich and Sampson proposed a stepwise regression technique: once the normal equation were formed, all the d_i components which cannot significantly reduce the value of the objective function are set to zero. This is a *Directional Discrimination Method*

Methods which do not require formation of the normal equations show greater numerical accuracy and are particularly suitable when precise solutions are required. We just mention as a detailed description is beyond our aims: in particular the Golub method and the Longley one are found to be considerably more accurate than solution of normal equation.

4.4.7 The Variable Metric Method

The Gauss method is the best available for the solutions of those problems to which it applies. The bug is that it can't be applied to any objective function. For those cases in which Gauss method doesn't work, a *variable metric* method is recommended.

The *variable metric* term was coined by Davidon to designate schemes in which the matrix R is systematically adjusted from iteration to iteration to make it behave like H^{-1}. These methods may be viewed as sophisticated finite difference schemes for computing the second derivatives of Φ. The Davidon scheme, slightly modified, has been widely used, gaining a reputation of being one of the most efficient general unconstrained optimization method available. Starting from this implementation, an effort to improve the outcomes is represented by ROC method (see below).

The main idea behind the variable metric method is the following: from the definition of gradient q and the Hessian H we have

$$H_i = \frac{\partial q}{\partial \theta}|_{\theta=\theta_i} \approx \frac{q_{i+1} - q_i}{\theta_{i+1} - \theta_i}$$

If $\sigma_i = \theta_{i+1} - \theta_i$ and $\eta = q_{i+1} - q_i$ we can write

$$H_i \sigma_i = \eta_i$$

[1]This method can be adapted to the singular or near-singular case, but this adaptation has performed poorly. In fact, although the Cholesky method gives a precise solution to the normal equation even when they are nearly singular, the step direction thus generated is so far from the negative gradient to be almost unacceptable.

or alternatively

$$\sigma_i = H_i^{-1}\eta_i$$

Suppose that before the ith iteration we have a matrix A which is an approximation to H^{-1}. We wish to add to it a correction ΔA_i such that the resulting matrix A_{i+1} satisfies $\sigma_i = H_i^{-1}\eta_i$ when replacing H^{-1}. In this way,

$$A_{i+1} \equiv A_i + \Delta A_i \tag{4.69}$$

We require that

$$\sigma_i = A_{i+1}\,\eta_i = A_i\eta_i + \Delta A_i\eta_i \tag{4.70}$$

Hence

$$\Delta A_i\eta_i = p_i \tag{4.71}$$

where $p_i = \sigma_i - A_i\eta_i$. Eq. 4.71 does not determine ΔA_i uniquely, since it contains only l conditions for the *l(l+1)/2* independent elements of the symmetric matrix ΔA_i. The simplest possible matrix ΔA_i is of rank one and it has the form

$$\Delta A_i = r_i r_i^T \tag{4.72}$$

where r_i is a suitable vector. Substituting in $\Delta A_i\eta_i = p_i$ we obtain

$$r_i r_i^T \eta_i = p_i \tag{4.73}$$

that is $r_i = (1/r_i^T\eta_i)p_i = \alpha p_i$ where $\alpha \equiv (r_i^T\eta_i)^{-1}$ is an unknown constant. Substituting in Eq. 4.73 and rearranging we find out

$$\alpha^2 = 1/p_i^T\eta_i \tag{4.74}$$

Finally

$$\Delta A_i = r_i r_i^T = \alpha^2 p_i p_i^T = (1/p_i^T\eta_i)p_i p^T \tag{4.75}$$

Eq. 4.75 define the *Rank One Correction* Method (ROC). Broyden, Davidon, Fiacco and McCormick have proved the following theorem:

Theorem 5 *Suppose $\Phi(\theta)$ is a quadratic function with a constant nonsingular Hessian matrix H. Let $\theta_1, \theta_2, \ldots, \theta_{l+1}$ be a set of points such that the vectors $\sigma_i \equiv \theta_{i+1} - \theta_i (i = 1, 2, \ldots, l)$ are linearly independent. Let A_1 be an arbitrary symmetric matrix, and let $A_i (i = 2, 3, \ldots, l+1)$ be defined recursively by $A_{i+1} \equiv A_i + \Delta A_i$ and Eq 4.75. Then,*

provided $p_i^T \eta_i \neq 0$ *for* $i = 1, 2, \ldots, l$, *we have*

$$A_{l+1} = H^{-1} \tag{4.76}$$

The theorem says that if Φ is quadratic, the ROC method produces the exact inverse Hessian in l steps. Once the inverse Hessian is known, a single Newton step converges to the minimum.

When Φ is not quadratic, one expects $A_i (i \geqslant 1)$ to represent an approximation to H^{-1} evaluated somewhere in the region of the last l iterates. This is particularly true near the minimum. We expect the matrices A_i to converge to the value of H^{-1} at the minimum.

Although the theorem in principle holds for arbitrary A_1, Bard suggests to chose a diagonal matrix with

$$A_{1\alpha\alpha} = -\theta_{1\alpha}/q_{1\alpha}. \tag{4.77}$$

Since A_i is an approximation to H^{-1}, we would like to take A_i for R_i. There is no guarantee, however, that A_i is positive definite. One could apply some correction method (as Greenstadt method or Farris-Law one) to render A_i positive definite. But these types of correction do not appear to work very well when applied to a matrix that is an approximation to the inverse, rather than to the Hessian itself. In this case we can use a procedure, entirely analogous to the ROC method, to construct an approximation to the Hessian directly. We call this method *Inverse Rank One Correction, IROC* (Bard, 1970). Instead of $H_i \sigma_i = \eta_i$, first-order approximation of the ROC method, here we wish to satisfy the condition

$$(A_i + \Delta A_i)\sigma_i = \eta_i \tag{4.78}$$

Rearranging,

$$A_i = (1/s_i^T d_i) s_i s_i^T \tag{4.79}$$

where $s_i \equiv \eta_i - A_i \sigma_i$. We initialize A_i as the inverse of the matrix defined by Eq. 4.77. The matrices A_i converge to H in the quadratic case. Since A_i is an approximation to H, we can use the Levenberg-Marquardt method to compute d_i efficiently.

In the Davidon-Fletcher-Powell (DPF) method, the matrix ΔA_i is of rank two instead of rank one. The simplest choice to satisfy $\Delta A_i \eta_i = p_i$ (Eq 4.71) is

$$\Delta A_i = (1/\sigma_i^T \eta_i)\sigma_i \sigma_i^T - (1/\eta_i^T A_i \eta_i) A_i \eta_i \eta_i^T A_i \tag{4.80}$$

We choose $\sigma_i = -\mu_i A_i q_i$ where μ_i is a positive value of t at which $\Phi(\theta_i - tA_i q_i)$ reaches a minimum. DPF have shown that under these conditions $A_{i+1} = A_i + \Delta A_i$ is positive definite, provided A_i was so. Therefore using $R_i = A_i$ always produces an acceptable step.

4.4.8 The Initial Guess

All the optimization method that we have described require that one supply an initial guess (or starting value) θ_0 for the values of the parameters. The choice of a good initial guess can spell the difference between success and failure in locating the optimum, or between rapid and slow convergence to the solution. Unfortunately while we can prescribe algorithms for proceeding from the initial guess, we must rely heavily on intuition and prior knowledge in selecting the initial guess. Nevertheless, we can provide some suggestions which may be helpful in many cases.

The most obvious method for making the initial guesses is by the use of prior information. Estimates calculate from previous experiments, known values from similar systems, values computed from theoretical considerations: all these form ideal initial guesses.

On the opposite end of the *spectrum* stand problems in which our only information concerning the parameter values is given in the form of upper and lower bounds of their values. If we do not have bounds, we can transform our variables into bounded ones; e.g. a positive variable θ can be replaced by the bounded variable $\phi = 1/(1 + \theta)$.

Once we have all our parameters confined to a rectangular region in θ space, we can conduct a *grid search*: compute the value of the objective function at every point on a regular rectangular grid, and choose with the best value as the initial guess.

An alternative to the grid search is *random search*. Here a number of points within the feasible region are chosen at random, and the one giving the best value of the objective function is used as the initial guess. Random search permits the use of termination criterion: one can stop sampling as soon as a function value is significantly better than the average that has been found.

It is not always necessary to provide initial guesses for all the parameters in a model. If some of the parameters enter the model equations linearly, and an initial guess is provided for the other parameters, then the linear parameters can be estimated by linear multiple regression.
Suppose, for instance, that the model has the form:

$$\hat{y}_\mu = \theta_\alpha e^{-\theta_\beta x}$$

If we have the initial guess $\theta_\beta = 6$, and let $z_\mu = e^{-6x}$, then the initial guess for θ_α can be found by solving the linear least squares problem:

$$\min \sum_\mu (y_\mu - \theta_1 z_\mu)^2$$

The most useful approach to find an initial guess is to substitute a simpler problem for the original estimation problem. The answer to the simpler problem can be used as initial guesses for the original problem.

There is no systematic way of applying this idea to all problems, but the following is a partial list of what may be attemped.

- **Linearization**. We try, by means of trasformation of variable, to change the model equations, into ones that are *linear in the parameters*. The linear problem can be solved by multiple linear regression with no need for an initial guess.

- **Multi-stage Estimation**. By breaking up data into groups, we may estimate certain auxiliary parameters for each groups; then we estimate the original parameters as function of the auxiliary parameters.
  For instance, if the model has the form:

  $$y = kxe^{-\frac{E}{T}}$$

  where y is the response variable, x the explication variable, T a controlled parameter and k, E the parameter to be estimated. So y is measured as a function of x at several values of T. We use the data taken at T_i to estimate the coefficient ξ_i in the equation

  $$y = \xi_i x$$

  The estimated ξ_i can then be used as data for estimating $\log k$ and E in the linearized model:

  $$\log \xi_i = \log k - E/T_i$$

- **Model Simplification**. Sometimes is possible to approach the final model through a sequence of simpler ones, in which various effects are neglected and the corresponding parameters suppressed. After the parameters have been estimated for the simpler model, analysis of the residuals can provide an indication to what terms should be added to the next model of the sequence.

4.4.9 Step Size

So far we have concerned primarily with choosing the direction of the step taken in the ith iteration, that is, in the choice of R. We now shall see how to determine the *step size* t_i. The methods proposed should be divided into three categories:

1. $\mathbf{t_i = 1}$. Required by Newton method, to guarantee quadratic convergence to the minimum, and by Marquardt method. In the latter case the step size is determined indirectly through the choice of κ_i.

2. $\mathbf{t_i = \mu_i}$. We proceed along the chosen direction to the point at which Φ stops to decrease, as required by DFP method.

3. **Interpolation-Extrapolation**, employed in conjunction with the Gauss and ROC methods. Here the effort is spent to find a good, acceptable value of t_i, without bothering to locate μ_i precisely.

In general, the closer t_i is to μ_i, the smaller is the total number of required iterations. On the other hand, the more precisely μ_i is determined, the larger is the number of times that we must evaluate the objective function at each iteration. The difference between point 2 and point 3 is that in the former the best outcome is achieved when μ_i is determined with much greater precision than is required in the latter.
In the succeeding section we report an algorithm to compute t_i in case number 3.

An Algorithm for the "Interpolation-Extrapolation" Case

Although the following algorithm has worked with a reasonable degree of success, there is no evidence that it is the most efficient possible.

The search for t_i proceeds without computation of derivatives, because it would be wasteful to compute at each point $l+1$ functions (Φ and l components of its gradient). In this algorithm it is assumed that at each iteration we are given an upper bound, $t_{i,max}$, on the feasible values of t_i, which can be chosen as an arbitrarily large number. We are also given a lover bound, $t_{i,min}$. If no acceptable $t > t_{i,min}$ can be found, the search stops.

BASIC IDEA: Assuming that we have chosen an acceptable direction, there always exists a number λ_i such that if $0 < t < \lambda_i$, then $\Psi_i(t) \equiv \Phi(\theta_i - tR_iq_i) < \Phi_i$.
The basic idea of the

- *Interpolation Method* is that if we have initially picked a value $t = t^0$ such that $\Psi_i(t^0) \geqslant \Phi_i$, we next try a smaller value of t and repeating the process until an

acceptable value is found.

- *Extrapolation Method* is that if our initial choice $t = t^0$ is acceptable, it pays to try at least one other value of t to see whether we can do better.

In both cases, the new trial value of t is chosen to minimize a quadratic approximation to $\Psi_i(t)$. We know that $\Psi_i(0) = \Phi_i$ and $d\Psi_i/dt|_{t=0} = -q_i^T R_i q_i$. Suppose we have computed $\Psi_i(t^0)$. Defining $\alpha \equiv \Phi_i, \beta \equiv \Psi_i(t^0), \gamma \equiv -q_i^T R_i q_i$, we try to find a quadratic function $a + bt + ct^2$ whose values match those of $\Psi_i(t)$ at $t = 0$ and $t = t^0$ and whose slope matches that of $\Psi_i(t)$ at $t = 0$. So we have:

$$a = \alpha \equiv \Psi_i \tag{4.81}$$

$$b = \gamma \equiv -q_i^T R_i q_i \tag{4.82}$$

$$a + bt^{(0)} + ct^{(0)^2} = \beta \equiv \Psi_i(t^0) \tag{4.83}$$

Rearranging we find:

$$c = (\beta - \alpha - \gamma t^{(0)})/t^{(0)^2}$$

The quadratic $a + bt + ct^2$ has a stationary point at

$$t^* = -b/2c = \gamma t^{(0)^2}/2(\gamma t(0) + \alpha - \beta)$$

A detailed implementation of this idea is given in the Fig. 4.1 and 4.2.

4.4.10 Termination Rules

It is necessary to devise a criterion to stop the iterative search for the minimum $\Phi(\theta)$. The best situation that should occur is convergence to a stationary point of Φ. It may seem natural therefore to adopt the vanishing of the gradient as the termination criterion. Unfortunately, rounding errors often make the goal of a vanishing gradient unattainable, even approximately.

In many cases the computer comes up with parameter values very close to the minimum, yet the gradient is still sizable. In addition, if the algorithm fails to converge at all, a termination rule based entirely on the gradient leaves the program to iterate endlessly.

A more practical criterion dictates to stop as soon as further iterations fail to match the parameter values significantly. That is, given a set of small numbers $\varepsilon_\alpha (\alpha = 1, 2, \ldots, l)$, we accept θ_{i+1} as the solution θ^* provided

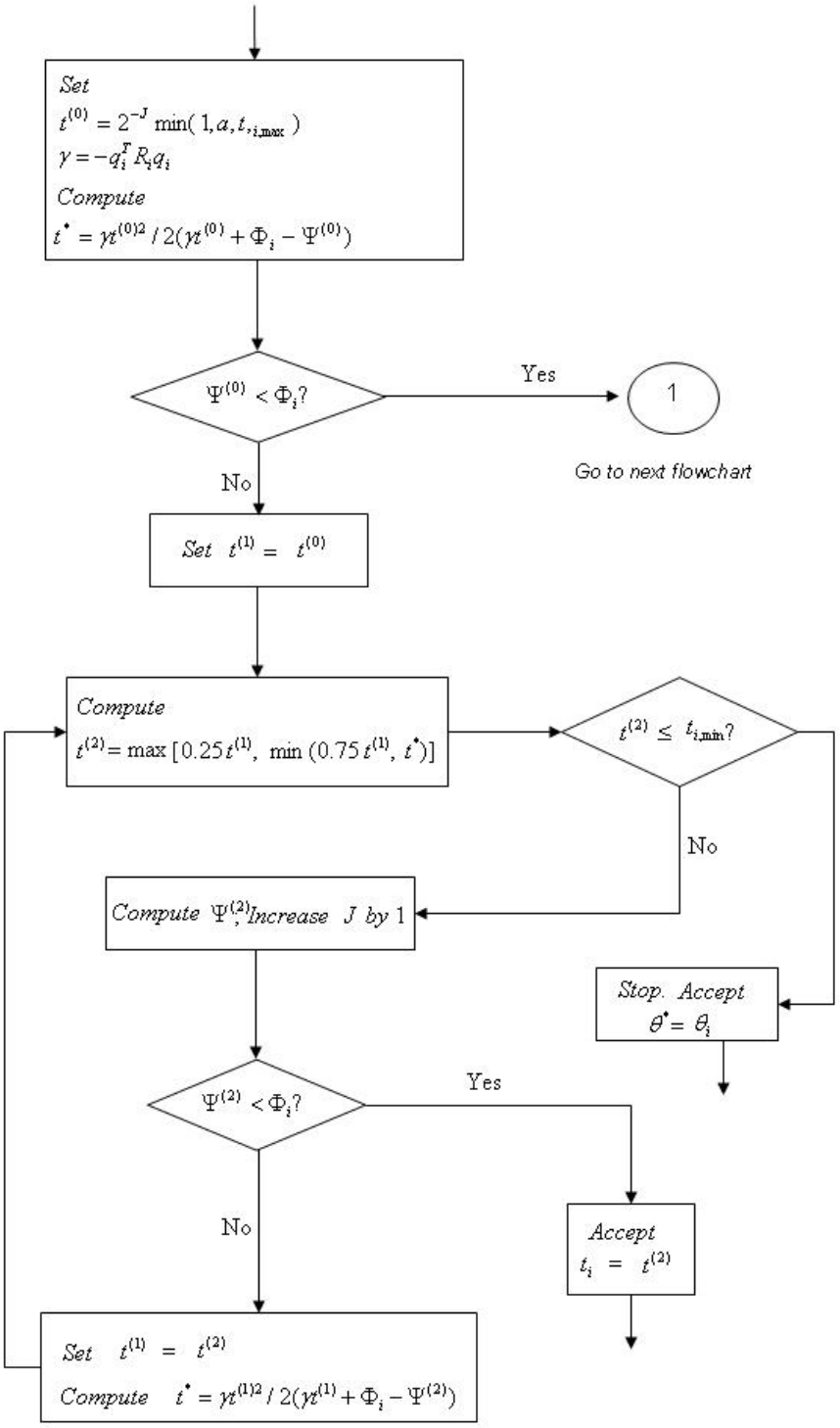


Figure 4.1: Interpolation method

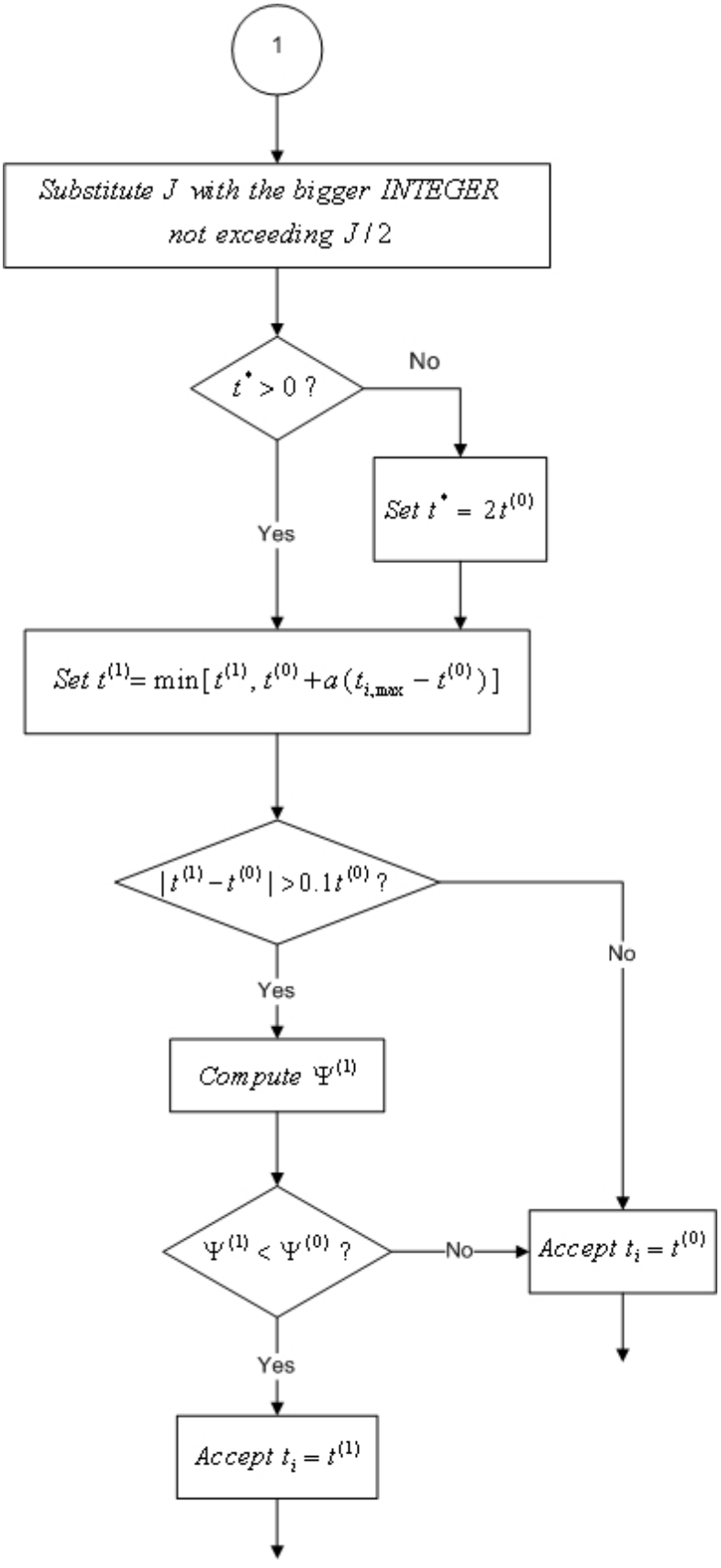


Figure 4.2: Extrapolation method

$$|\theta_{i+1,\alpha} - \theta_{i,\alpha}| \leqslant \varepsilon_\alpha \quad (\alpha = 1, 2, \ldots, l) \tag{4.84}$$

where $\theta_{i,\alpha}$ is the αth component of θ_i. The numbers ε_α may either be prescribed in advance, or they may be computed by the program. In the latter case, following Marquardt, we raccomended

$$\varepsilon_\alpha = 10^{-4}(\theta_{i,\alpha} + 10^{-3}) \tag{4.85}$$

The additive term 10^{-3} is designed to avoid the situation where θ_α equate to zero. This criterion has worked very well in practice, also if sometimes it allows a few more iterations than are strictly necessary. Suppose in the ith iteration a step direction d_i has been determined. Then Eq. 4.84 is satisfied if for each $\alpha, t|d_{i,\alpha}| \leqslant \varepsilon_\alpha$. Hence the minimum admissible t for the ith iteration is $t_{i,min} = \min_\alpha[\varepsilon_\alpha/|d_{i,\alpha}|]$.

Termination occurs if the algorithm is forced to choose $t_i \leqslant t_{i,min}$. The above criterion does not offer a strong guarantee that the process will terminate in a finite number of steps. If the objective function is known to have a finite minimum, then the termination can be guaranteed if we stop whenever $\Phi_{i-1} - \Phi_i < \varepsilon$ for some small specified positive number ε. That is, we stop as soon as no significant progress is made in reducing the value of the objective function. Finally an upper bound may be placed on the number of iterations allowed.

Once the iterative process is terminated at $\theta = \theta^*$, one would like to know whether or not one has arrived at a minimum. We assume that we know the gradient $q^* = q(\theta^*)$ and at least some approximation H^* to the Hessian $H(\theta^*)$. If we cut a cross section of the Φ surface along the θ_α axis, we have a curve whose approximate equation near θ^* is given by

$$\Psi(\theta_\alpha) = \Phi^* + q^*_\alpha(\theta_\alpha - \theta^*_\alpha) + \frac{1}{2}H^*_{\alpha\alpha}(\theta_\alpha - \theta^*_\alpha)^2 \tag{4.86}$$

which has a stationary point at

$$\theta_\alpha = \theta^*_\alpha - q^*_\alpha/H^*_{\alpha\alpha} \tag{4.87}$$

The quantity $\delta_\alpha = |q^*_\alpha/H^*_{\alpha\alpha}|$ is therefore a measure of the error in the determination of θ^*. If each δ_α is small on the scale by which θ_α is measured, then it is likely that θ^* is very close to a stationary point of Φ.

If H^* is indeed the Hessian of Φ at θ^*, then we may easily determine whether θ^* (already known to be a stationary point) is really a minimum. All that is required is

that H^* be positive definite, that is all its eigenvalues be positive.

- When using the Gauss method, our approximation N^* is constructed to be automatically positive definite, regardless of whether H^* is positive or not. In these case N^* does not contain any information on the nature of the point θ^*. So we have to explore directly the behavior of Φ around θ^*.

- On the other hand, in the ROC method, the matrix R_i from the last iteration may be a true approximation to $[H(\theta^*)]^{-1}$. We cannot prove that R_i is or not positive definite when θ^* is or not a minimum; however if R_i is not positive definite we suspect that θ^* is not a minimum and *vice versa*.

So, if one doubts that θ^* is a minimum, one should restart the iterative procedure from a point close but not identical to θ^*. If it converges to the same θ^*, this is likely to be at least a local minimum.

4.5 Estimation of the Farrington's seroprevalence model

In Chapter 3 we have seen Farrington's model for seroprevalence:

$$F(a) = 1 - \exp\left\{\frac{b_1}{b_2} a e^{-b_2 a} + \frac{1}{b_2}\left(\frac{b_1}{b_2} - b_3\right)(e^{-b_2 a} - 1) - b_3 a\right\}. \tag{4.88}$$

This model comes from an hypothesis of exponentially damped linear model for the force of infection of measles, mumps and rubella:

$$\ell(a) = (b_1 a - b_3) e^{-b_2 a} + b_3 \qquad b_1, b_2, b_3 \geq 0. \tag{4.89}$$

In that chapter we have also seen by graphical representations that this model appears broadly consistent with the observations.

In this chapter we fit model 4.88 by nonlinear least squares (NLS) methods (see Section 4.1.1). The function we want to minimize is:

$$S(a; b_1, b_2, b_3) = \sum_{i=1}^{k} [F(a) - F(a; b_1, b_2, b_3)]^2, \tag{4.90}$$

where k are the number of age groups, $F(a)$ are the observed seropositive proportions and $F(a; b_1, b_2, b_3)$ are the estimated ones. For example, in Fig. 4.3 we show the surface

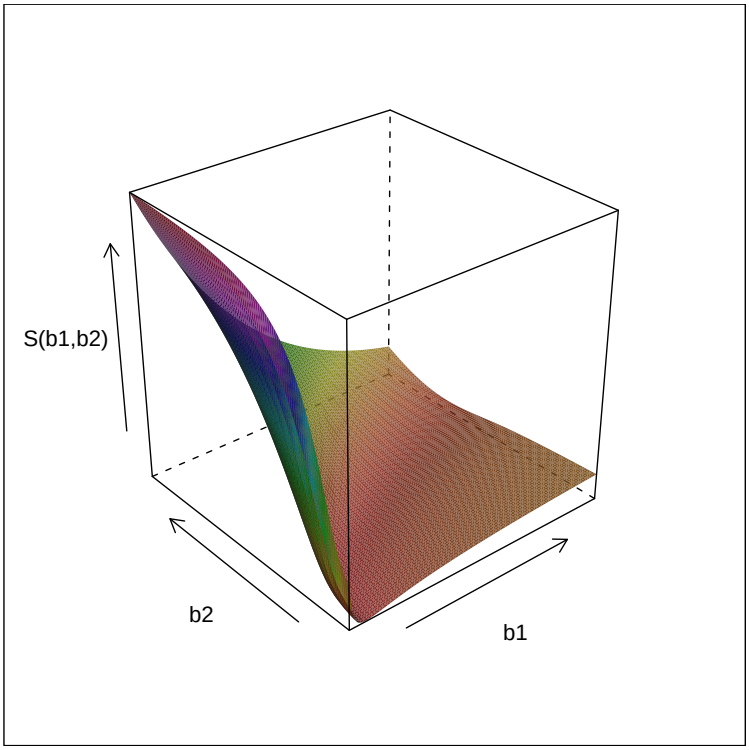


Figure 4.3: Mumps: surface of the sum of squares $S(b_1, b_2, b_3)$ with b_3 constrained to 0

of $S(a; b_1, b_2, b_3)$ for mumps, with b_3 constrained to 0 (as we see from the estimation of the models, the parameter b_3 will never be significantly diffent from 0).

The gradient vector, that is the 3x1 vector of the first derivatives of $S(a; b_1, b_2, b_3)$ with respect to the parameters, is:

$$q(a; b_1, b_2, b_3) = \sum_{i=1}^{k} [F(a) - F(a; b_1, b_2, b_3)] \left[\frac{\partial F(a; b_i)}{\partial b_i} \right]. \tag{4.91}$$

To minimize $S(a; b_1, b_2, b_3)$, we use four different iterative algorithms:

1. the Newton-Raphson algorithm;
2. the Levenberg-Marquardt algorithm;
3. the Gauss-Newton algorithm;
4. the Variable Metric algorithm.

All these algorithms are implemented with the statistical software R 2.4.1 [28]. Now we present the results for the three datasets previously introduced in a series of comparative tables. These tables show the following values:

1. The starting values for the algorithm, derived from the linearization of the empirical hazard function (see Section 3.2.2).
2. The parameter estimates b_i (the values with the asterisk are the significative parameters).
3. The standard errors for parameter estimates $se(b_i)$, necessary to test the significance of the parameters by the ratio $b_i/se(b_i)$.
4. The estimated minimum of the function $S(a; b_1, b_2, b_3)$ to be minimized. Being this function the sum of squared difference between the observed proportions and the estimated proportions, the minimum represents the residual sum of squares (RSS) of the model.
5. The number of iterations required by the algorithm to converge to the minimum of the function $S(a; b_1, b_2, b_3)$.

MUMPS			
		Newton-Raphson	Levenberg-Marquardt
Starting Values			
	b_1	0.054105	0.054105
	b_2	0.076544	0.076544
	b_3	0.000000	0.000000
Parameter Estimates			
	b_1	0.132012*	0.132012*
	b_2	0.163132*	0.163135*
	b_3	−0.041836	−0.41831
Parameter Standard Errors			
	b_1	0.006680	0.006774
	b_2	0.019004	0.020145
	b_3	0.039834	0.043966
Minimum of $S(b_1, b_2, b_3)$		0.016060	0.016060
Iterations		29	12

Table 4.1: Comparative table between Newton-Raphson and Levenberg-Marquardt algorithms for mumps data

MUMPS			
		Gauss-Newton	Variable Metric
Starting Values			
	b_1	0.054105	0.054105
	b_2	0.114098	0.076544
	b_3	0.000000	0.000000
Parameter Estimates			
	b_1	0.132012*	0.132008*
	b_2	0.163134*	0.163007*
	b_3	−0.041833	−0.042118
Parameter Standard Errors			
	b_1	0.006774	0.006676
	b_2	0.020146	0.018770
	b_3	0.043966	0.038985
Minimum of $S(b_1, b_2, b_3)$		0.016060	0.016060
Iterations		15	20

Table 4.2: Comparative table between Gauss-Newton and Variable Metric algorithms for mumps data

4.5.1 Mumps: estimation of the seroprevalence

In Tab. 4.1 and Tab. 4.2 we present the comparated results for mumps.

From the two tables, it emerges that the more efficient algorithms, that is the algorithms which converge at the solution with the minor number of iterations, are the Levenberg-Marquardt (12 iterations) and the Gauss-Newton (15 iterations).

All the algorithms, except the Gauss-Newton, converge to the minimum of $S(a; b_1, b_2, b_3)$ starting from the initial guesses furnished by the linear fitting of the $R(a)$ function (0.054105 and 0.076544 respectively for b_1 and b_2), even though the initial guesses are significatively far from the parameter estimates. Instead, in this case the Gauss-Newton algorithm is not able to converge to the minimum using these starting values: in particular, the initial guess for the parameter b_2 is too far from the estimate and this fact causes an error ("singular gradient") in the procedure. So, trying different values, every time closer to the estimate, we have found the minimum initial guess (0.114098) required from the algorith to converge.

Besides, whatever algorithm we use, we found that the parameter b_3 is not significantly different from 0.

Fig. 4.4 shows the observed seropositive proportions and the estimated curve determined with the Gauss-Newton algorithm.

Now we report some measures of goodness of fit, whose meaning will be fully explicated in Chapter 5. These measures are:

1. *Degrees of freedom* (d.f.): the residual degrees of freedom of the model, which are $N - k$, where N is the number of covariate patterns and k is the number of parameters in the model.

2. *Deviance* (D): see Section 5.4.1. When comparing fitted models, the smaller the deviance, the better the fit.

3. *Pearson's chi-squared statistic* (X^2): see Section 5.4.2. When comparing fitted models, the smaller X^2, the better the fit.

4. *Likelihood ratio chi-squared statistic* (C): see Section 5.4.3. When comparing fitted models, the higher C, the better the fit.

5. *Pseudo* R^2: see Section 5.4.4. When comparing fitted models, the higher the pseudo R^2, the better the fit.

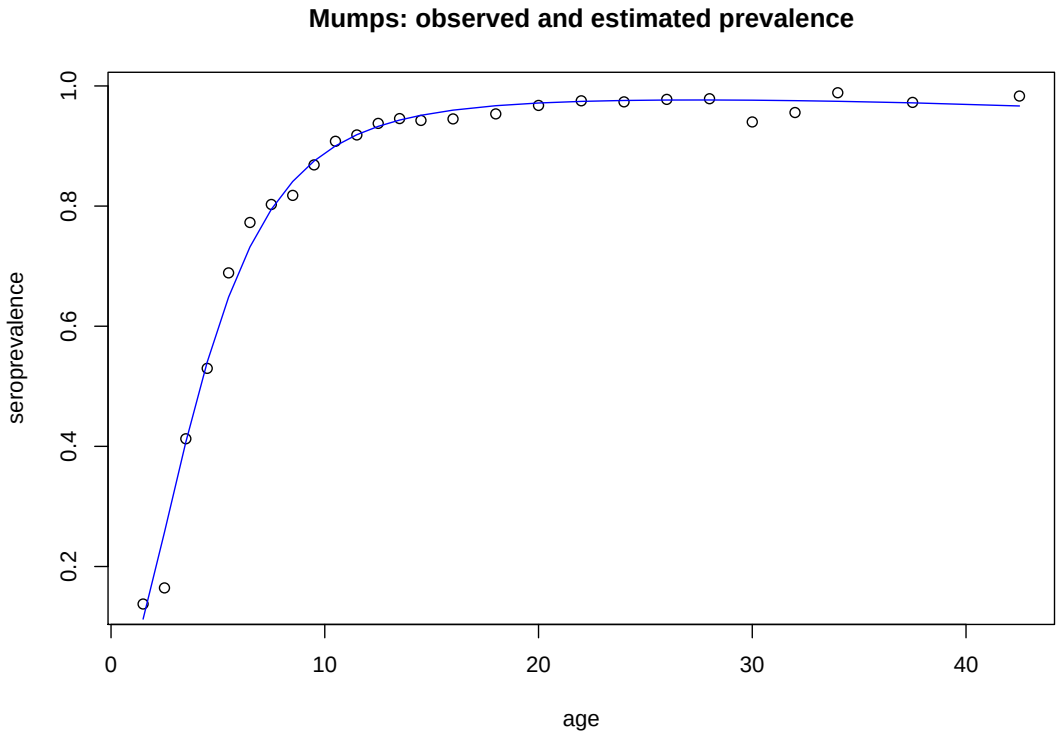


Figure 4.4: Mumps: observed and estimated prevalence by Gauss-Newton algorithm

6. R^2 *based on the Kullback - Leibler divergence* (R^2_{KL}): see Section 5.4.5. When comparing fitted models, the higher this statistic, the better the fit.

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
23.00	46.48	49.89	2877.06	0.3637	0.9841

Table 4.3: Mumps: measures of goodness of fit for the estimated non-linear least squares model for prevalence

The measures presented in Tab. 4.3 tell us that Farrington's model is a good model: for this fitted model the deviance is 46.48 on 23 degrees of freedom and the chi-squared goodness of fit statistics is about 50 (23 d.f.). The R^2 constructed from the Kullback-Leibler divergence is 0.98, that is about 98% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.3637, while the maximum achievable is 0.3696.

4.5.2 Rubella: estimation of the seroprevalence

In Tab. 4.4 and Tab. 4.5 we present the comparated results for rubella.

In this case also the more efficient algorithms are the Levenberg-Marquardt (9 iterations) and the Gauss-Newton (14 iterations).

Differently from mumps data, for rubella all the algorithms achieve the convergence starting from the initial guesses provided by the linear plot of $R(a)$, that is 0.026984 for b_1 and 0.057823 for b_2.

Even for rubella, the parameter b_3 is never significant.

Fig. 4.5 shows the observed seropositive proportions and the estimated curve determined with the Gauss-Newton algorithm.

As we have done with mumps, now we report some measures of goodness of fit in Tab. 4.6.

The measures presented in Tab. 4.6 tell us that Farrington's model for rubella is also a good model: for this fitted model the deviance is 47.40 on 23 degrees of freedom and the chi-squared goodness of fit statistics is about 54 (23 d.f.). The R^2 constructed from the Kullback-Leibler divergence is 0.97, that is about 97% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.2516, while the maximum achievable is 0.2607.

RUBELLA			
		Newton-Raphson	Levenberg-Marquardt
Starting Values			
	b_1	0.026984	0.026984
	b_2	0.057823	0.057823
	b_3	0.000000	0.000000
Parameter Estimates			
	b_1	0.064746*	0.064746*
	b_2	0.175518*	0.175515*
	b_3	0.023834	0.023832
Parameter Standard Errors			
	b_1	0.005328	0.005258
	b_2	0.030396	0.030205
	b_3	0.026386	0.026697
Minimum of $S(b_1, b_2, b_3)$		0.031598	0.031598
Iterations		28	9

Table 4.4: Comparative table between Newton-Raphson and Levenberg-Marquardt algorithms for rubella data

RUBELLA			
		Gauss-Newton	Variable Metric
Starting Values			
	b_1	0.026984	0.026984
	b_2	0.057823	0.057823
	b_3	0.000000	0.000000
Parameter Estimates			
	b_1	0.064746*	0.064805*
	b_2	0.175515*	0.175863*
	b_3	0.023832	0.024034
Parameter Standard Errors			
	b_1	0.005258	0.005269
	b_2	0.030204	0.029218
	b_3	0.026697	0.025370
Minimum of $S(b_1, b_2, b_3)$		0.031598	0.031599
Iterations		14	23

Table 4.5: Comparative table between Gauss-Newton and Variable Metric algorithms for rubella data

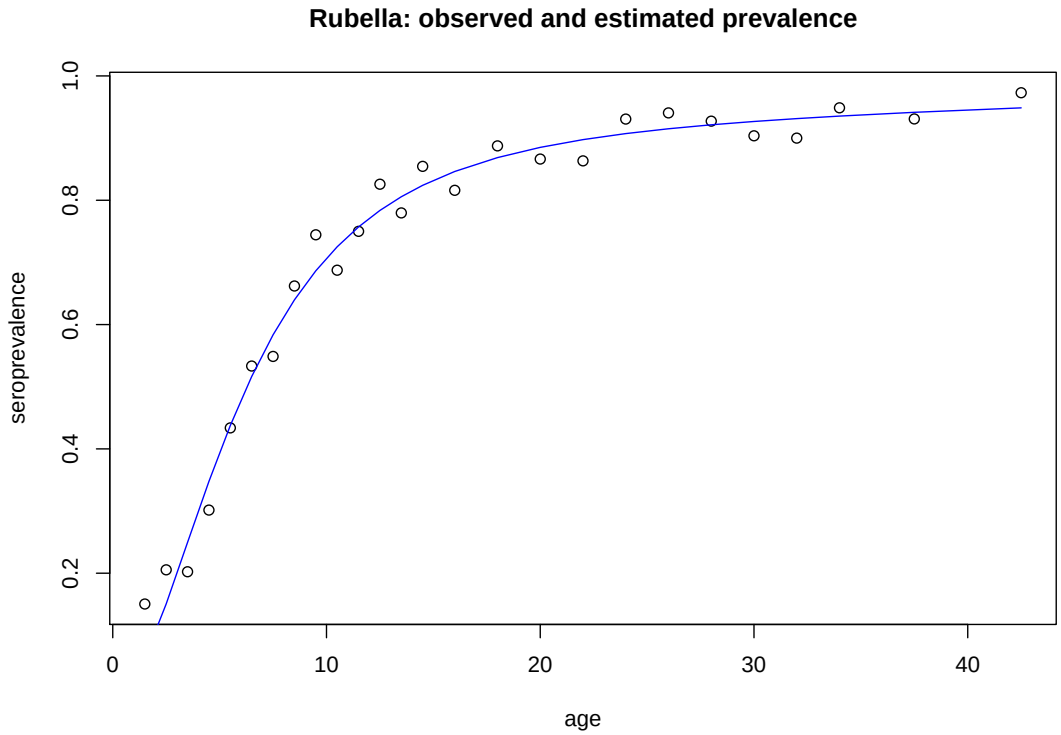


Figure 4.5: Rubella: observed and estimated prevalence by Gauss-Newton algorithm

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
23.00	47.40	54.18	1309.49	0.2516	0.9651

Table 4.6: Rubella: measures of goodness of fit for the estimated non-linear least squares model for prevalence

PARVOVIRUS			
		Newton-Raphson	Levenberg-Marquardt
Starting Values			
	b_1	0.009861	0.009861
	b_2	0.072097	0.072097
	b_3	0.000000	0.000000
Parameter Estimates			
	b_1	0.045824*	0.045821*
	b_2	0.252131*	0.252111*
	b_3	0.006502	0.006500
Parameter Standard Errors			
	b_1	0.010373	0.010381
	b_2	0.051183	0.052943
	b_3	0.006052	0.006431
Minimum of $S(b_1, b_2, b_3)$		0.088321	0.088321
Iterations		27	13

Table 4.7: Comparative table between Newton-Raphson and Levenberg-Marquardt algorithms for parvovirus data

4.5.3 Parvovirus: estimation of the seroprevalence

In Tab. 4.7 and Tab. 4.8 we present the comparated results for parvovirus.

The algorithms applied to parvovirus data confirm that the best ones are the Levenberg-Marquardt (13 iterations) and the Gauss-Newton (13 iterations). All the algorithms are able to converge starting from the initial guesses (0.009861 for b_1 and 0.072097 for b_2) provided by the linear plot of $R(a)$.

These data also confirm that the parameter b_3 is not significantly different from 0.

Fig. 4.6 shows the observed seropositive proportions and the estimated curve determined with the Gauss-Newton algorithm. As we can see, the model is not able to fit well the behaviour of data after 20 years old, when the observed seropositive proportions

PARVOVIRUS			
		Gauss-Newton	Variable Metric
Starting Values			
	b_1	0.009861	0.009861
	b_2	0.072097	0.072097
	b_3	0.000000	0.000000
Parameter Estimates			
	b_1	0.045821*	0.046102*
	b_2	0.252113*	0.253827*
	b_3	0.006500	0.006700
Parameter Standard Errors			
	b_1	0.010380	0.009538
	b_2	0.052941	0.046587
	b_3	0.006432	0.005655
Minimum of $S(b_1, b_2, b_3)$		0.088321	0.088326
Iterations		13	34

Table 4.8: Comparative table between Gauss-Newton and Variable Metric algorithms for parvovirus data

before decrease and then increase again.

As we have done with mumps and rubella, now we report some measures of goodness of fit in Tab. 4.9.

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
23.00	49.34	54.27	256.75	0.0576	0.8388

Table 4.9: Parvovirus: measures of goodness of fit for the estimated non-linear least squares model for prevalence

The measures presented in Tab. 4.6 tell us that Farrington's model for parvovirus is not a bad model, but neither a very good one: for this fitted model the deviance is 49.34 on 23 degrees of freedom and the chi-squared goodness of fit statistics is about 54 (23 d.f.). However, it may be that these small values of D and X^2 are due more to the small values of the fitted probabilities $\hat{\pi}$ rather than to the adequacy of the model. The R^2 constructed from the Kullback-Leibler divergence is 0.84, that is about 84% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.0576, while the maximum achievable is 0.0687.

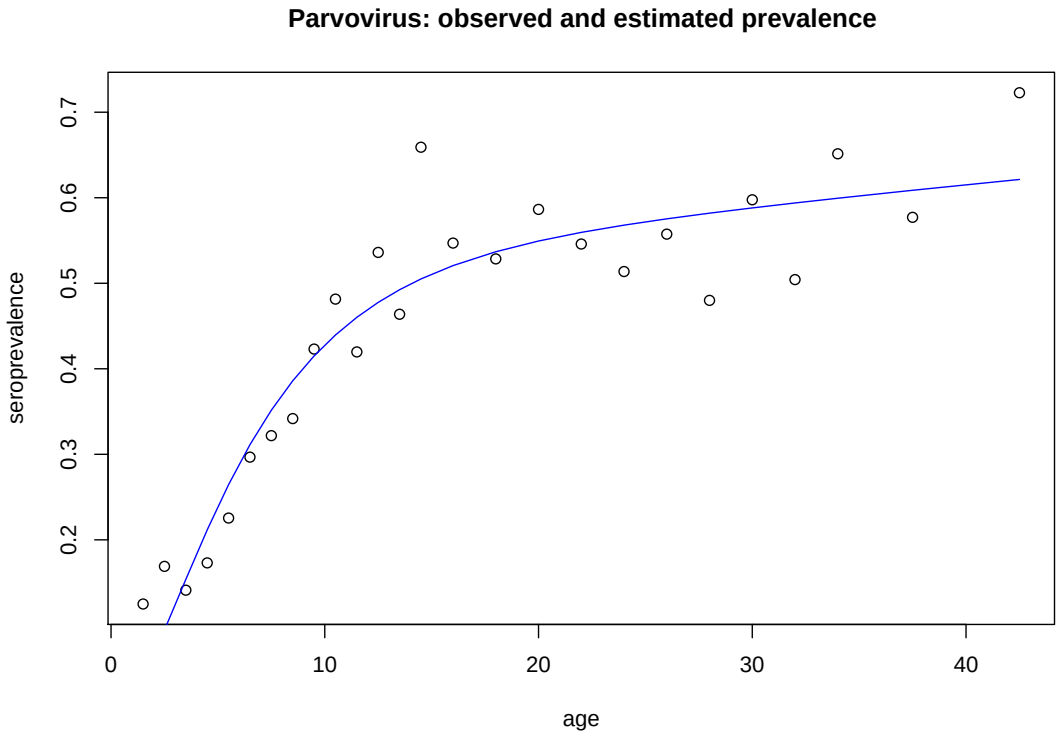


Figure 4.6: Parvovirus: observed and estimated prevalence by Gauss-Newton algorithm

4.6 Estimation of the Farrington's force of infection model

At this point, after the estimation of the parameters of seroprevalence function 4.88, we can finally estimate the force of infection, in accordance with the following function, proposed by Farrington [1]:

$$\ell(a) = (b_1 a - b_3)e^{-b_2 a} + b_3. \tag{4.92}$$

However, it is important to remember that in all the dataset presented, the parameter b_3 is resulted not significant.

Besides, for every dataset, we have also determined the average age at infection A, proposed by Griffiths [13], which is the expected value of the age at infection in the population:

$$A = \int_0^L a f(a) da = \int_0^L P(a) da, \tag{4.93}$$

where L is the average life expectancy (taken to be 75 years).

Since $F(L) < 1$, there is a finite atom of probability that an individual will remain uninfected throughout his or her lifetime. Let f denote this atom:

$$f = 1 - F(L). \tag{4.94}$$

But, in general, if we have data available to some upper age limit $U < L$ (e.g. in our case $U = 44$), then

$$f = 1 - F(U). \tag{4.95}$$

In this way, if we want to determine the average age at infection, we have to add to the formula proposed by Griffiths [13] a correction factor:

$$A = \int_0^U P(a) da + f(L - U). \tag{4.96}$$

4.6.1 Mumps: estimation of the force of infection

In Fig. 4.7 we present the graph of the estimated force of infection for mumps data.

The force of infection for mumps presents a steep rise until 6.5 years old, which is the modal age at infection, that is the age at which the number of new cases is maximum:

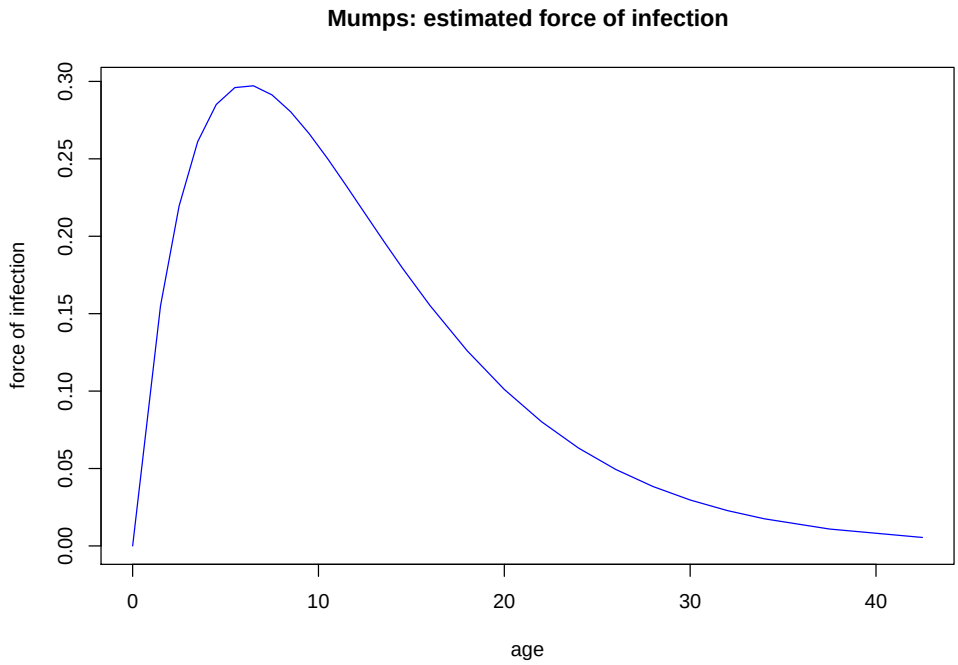


Figure 4.7: Mumps: estimated force of infection in accordance with Farringtons's model

here the force of infection is $\ell(6.5) = 0.2972$, that is to say people aged 6.5 years old have a risk of 29.72% to acquire the infection.

Afterwards, the force of infection declines in a less rapid way.

The average age at infection A for mumps, evaluated using Eq. 4.96, is 5.2 years old. This means that mumps is a typical disease of preschool children and young schoolchildren.

4.6.2 Rubella: estimation of the force of infection

In Fig. 4.8 we present the graph of the estimated force of infection for rubella data.

The force of infection for rubella has a similar pattern to mumps, but less steep. It presents a rise until 5.5 years old: here the force of infection is 0.1356, that is to say people aged 5.5 years old have a risk of 13.56% to acquire the infection. As we can see, the force of infection for rubella is weaker than that for mumps: a low peak is followed by a much slower decline in the force of infection with age.

The average age at infection A for rubella, evaluated using Eq. 4.96, is 12.1 years old. This means that rubella is typical of older schoolchildren.

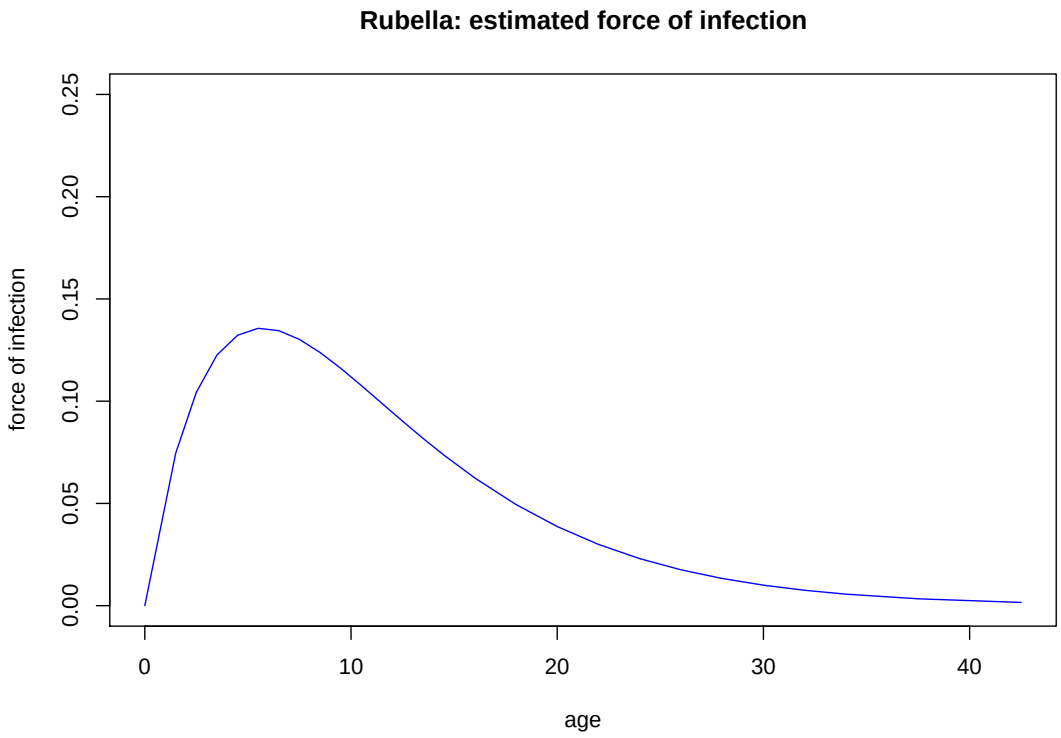


Figure 4.8: Rubella: estimated force of infection in accordance with Farringtons's model

4.6.3 Parvovirus: estimation of the force of infection

In Fig. 4.9 we present the graph of the estimated force of infection for parvovirus data.

The force of infection for parvovirus has a similar pattern to mumps and rubella, but less steep than rubella. It presents a rise until 3.5 years old: here the force of infection is 0.0664, that is to say people aged 3.5 years old have a risk of 6.64% to acquire the infection. As we can see, the force of infection for parvovirus is weaker than that for rubella and much lower than that for mumps: a very low peak is followed by a much slower decline in the force of infection with age.

The weaker the force of infection, the higher the average age at infection. So, in parvovirus case the average age at infection A, evaluated using Eq. 4.96, is 25.4 years old. This means that parvovirus infections are typical of young men and young women.

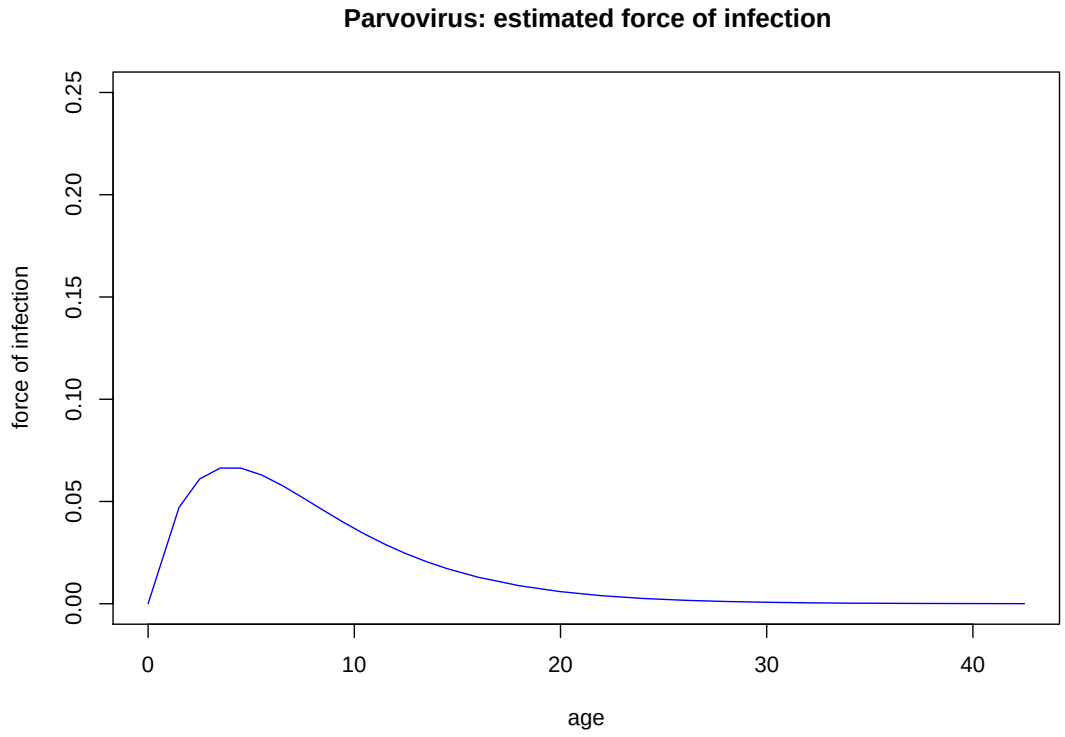


Figure 4.9: Parvovirus: estimated force of infection in accordance with Farringtons's model

Chapter 5

Generalized linear models

5.1 Generalized linear models

A generalized linear model (GLM) is composed by three components:

1. The random component Y, which is identically and independently distributed with constant variance and $E[Y] = \mu$.

2. The systematic component η, which is the following linear predictor:

$$\eta = \sum_i \beta_i x_i. \tag{5.1}$$

3. The *link* between the random and the systematic components. This link is a function, $g(\cdot)$, that puts together the two previous components:

$$g(E[Y]) = \eta, \tag{5.2}$$

or, that is the same,

$$E[Y] = g^{-1}(\eta). \tag{5.3}$$

Generalized linear models allow two extensions:

1. The distribution of Y may come from an exponential family

2. The link function $g(\cdot)$ may become any monotonic differentiable function.

The subject of generalized linear models was formulated by Nelder and Wedderburn [29] as a way of putting under one framework various previous models, and finding their commonalities. See also McCullagh and Nelder [30] and Dobson [31].

5.2 Exponential familiy of distributions

Consider a single random variable Y whose probability distribution depends on a single parameter θ. The distribution belongs to the exponential family if it can be written in the form

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}. \tag{5.4}$$

The functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ define the distribution.

The parameter ϕ is called *dispersion parameter*. If it is known, then the distribution of Y belongs to the exponential family; otherwise, we cannot state that the distribution belongs to the exponential family.

The parameter θ is called *canonical* or *natural* parameter of the exponential family.

Many well-known distributions belong to the exponential family. For example, the Poisson, Normal, binomial and gamma distributions can all be written in the canonical form. Thus for the binomial distribution

$$f_Y(y;\theta,\phi) \quad = \quad \left(\begin{array}{c} n \\ y \end{array}\right) \pi^y (1-\pi)^{n-y} \tag{5.5}$$

We write $L(\theta,\phi;y) = lnf_Y(y;\theta,\phi)$ for the log-likelihood function considered as a function of θ and ϕ, y being given. The mean and the variance of Y can be derived from the relations

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0 \tag{5.6}$$

and

$$E\left(\frac{\partial^2 L}{\partial \theta^2}\right) + E\left(\frac{\partial L}{\partial \theta}\right)^2 = 0. \tag{5.7}$$

The log-likelihood of $f_Y(y;\theta)$ is

$$L(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi), \tag{5.8}$$

from what

$$E\left(\frac{\partial L}{\partial \theta}\right) = \frac{y - b'(\theta)}{a(\phi)}, \tag{5.9}$$

and

$$E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -\frac{b''(\theta)}{a(\phi)}, \tag{5.10}$$

where primes denote differentiation with respect to θ.
Setting Eq. 5.9 to 0, we have

$$0 = E\left(\frac{\partial L}{\partial \theta}\right) = \frac{\mu - b'(\theta)}{a(\phi)}, \tag{5.11}$$

so that

$$E(Y) = \mu = b'(\theta). \tag{5.12}$$

Similarly from Eq. 5.7, Eq. 5.9 and Eq. 5.10 we have

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{[\mu - b'(\theta)]^2}{a^2(\phi)} = -\frac{b''(\theta)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)}, \tag{5.13}$$

so that

$$Var(Y) = b''(\theta)a(\phi). \tag{5.14}$$

Thus the variance of Y is the product of two functions:

1. $b''(\theta)$, which only depends on the canonical parameter θ (and hence on the mean μ) and is the so-called *variance function*.

2. $a(\phi)$ is independent of θ and depends only on ϕ. This function is commonly of the form

$$a(\phi) = \frac{\phi}{w}, \tag{5.15}$$

where the dispersion parameter ϕ is constant over observations and w is a known *prior weight* that varies from observation to observation.

	Normal	Poisson	Binomial
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$Bin(n, \pi)/n$
Range of y	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)n}{n}$
Dispersion parameter: ϕ	σ^2	1	$1/n$
Cumulant function: $b(\theta)$	$\theta^2/2$	e^θ	$\ln(1+e^\theta)$
$c(y;\phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$	$-\ln y!$	$\ln\begin{pmatrix} n \\ ny \end{pmatrix}$
$\mu(\theta) = E(Y;\theta)$	θ	e^θ	$e^\theta/(1+e^\theta)$
Canonical link: $\theta(\mu)$	identity	log	logit
Variance function	1	μ	$\pi(1-\pi)$

Table 5.1: Characteristics of some common univariate distributions in the exponential family

The most important distributions of the form 5.4 are summarized in Tab. 5.1.

5.3 The link function

The link function relates the linear predictor η to the expected value μ of a datum y.

For the binomial distribution we have $o < \pi < 1$ and a link should satisfy the condition that it maps the interval (0,1) onto the whole real line. We consider three possible functions:

1. Logit:
$$\ln\left(\frac{\pi}{1-\pi}\right) = \eta; \tag{5.16}$$

2. Probit:
$$\Phi^{-1}(\pi) = \eta, \tag{5.17}$$

where Φ^{-1} is the Normal cumulative distribution function;

3. Complementary log-log:

$$\ln[-\ln(1-\pi)] = \eta. \tag{5.18}$$

5.3.1 Sufficient statistics and canonical links

Each of the distributions in Tab 5.1 has a special link function for which there exists a sufficient statistic equal in dimension to β in the linear predictor $\eta = \sum x_j \beta_j$. These canonical links, as they are called, occur when

$$\theta = \eta, \tag{5.19}$$

where θ is the canonical parameter.

The canonical link are reported in Tab. 5.1; for the binomial distribution, the canonical link is the logit link:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \eta. \tag{5.20}$$

For the canonical links, the sufficient statistic is $X'Y$ in matrix notation, with components

$$\sum_i x_{ij} Y_i \qquad j = 1, \ldots, N, \tag{5.21}$$

summation being over the units.

Although canonical links lead to desirable statistical properties of the model, particularly in small samples, there is in general no *a priori* reason why the systematic effects in a model should be additive on the scale given by that link. It is convenient if they are, but convenience alone must no replace quality of fit as a model selection criterion.

5.4 Measuring the goodness of fit

Fitting a model to data may be regarded as a way of replacing a set of data values y by a set of fitted values $\hat{\mu}$ derived from a model involving (usually) a relatively small number of parameters. In general the μs will not equal the ys exactly, and the question then arises of how discrepant they are, because while a small discrepancy may be tolerable, a large discrepancy is not. Measures of discrepancy (or goodness of fit) may be formed in various ways, but we shall be primarily concerned with that formed from the logarithm of a ratio of likelihoods, to be called the *deviance*.

Given N observations we can fit models to them containing up to N parameters. The two extreme models are the *null model* and the *full model*:

- The null model f_{null} is one in which only one parameter μ is used so that $\eta(E[Y_i]) = \mu$, that is all responses have the same predicted outcome. This model consigns all the variation between the ys to the random component.
- The full model f_{max} is the other extreme where the maximum number of parameters are used in the model so that the observed response values equal to the predicted response values exactly, $\eta(E[Y_i]) = y_i$. With this model we have a perfect fitting to observed data: it consigns all the variation in the ys to the systematic component leaving none for the random component.

In practice the null model is usually too simple and the full model is uninformative because it does not summarize the data but merely repeats them in full. However, the full model gives us a baseline for measuring the discrepancy for an internediate model with k parameters.

5.4.1 The deviance

It is convenient to express the log-likelihood in terms of the mean-value parameter μ rather than the canonical parameter θ. Let $L(\hat{\mu}, \phi; y)$ be the log likelihood maximized over β for a fixed value of the dispersion parameter ϕ. The maximum likelihood achievable in a full model with N parameters is $L(y, \phi; y)$, which is ordinarily finite. The discrepancy of a fit is proportional to twice the difference between the maximum log-likelihood achievable and that achieved by the model under investigation. If we denote by $\hat{\theta} = \theta(\hat{\mu})$ and $\tilde{\theta} = \theta(y)$ the estimates of the canonical parameters under the two models, the discrepancy, assuming $a_i(\phi) = \phi/w_i$, can be written

$$\frac{D(y;\hat{\mu})}{\phi} = \frac{\sum_i 2w_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right\}}{\phi}, \tag{5.22}$$

where $D(y;\hat{\mu})$ is known as the *deviance* (or the *log-likelihood ratio statistic*) for the current model and is a function of the data only. Note that

$$D^*(y;\hat{\mu}) = \frac{D(y;\hat{\mu})}{\phi}, \tag{5.23}$$

so that the *scaled deviance* $D^*(y;\hat{\mu})$ is the deviance expressed as a multiple of the dispersion parameter.

Let us see now the deviance for binomial data. The log-likelihood is

$$L(\hat{\pi}; y) = \sum_i \{y_i \ln \hat{\pi}_i + (n_i - y_i) \ln(1 - \hat{\pi}_i)\}. \tag{5.24}$$

The maximum achievable log-likelihood is attained at the point

$$\tilde{\pi}_i = \frac{y_i}{n_i}. \tag{5.25}$$

The deviance function is therefore

$$\begin{aligned} D(y; \hat{\pi}) &= 2[L_{max} - L] \\ &= 2[L(\tilde{\pi}; y) - L(\hat{\pi}; y)] \\ &= 2 \sum_i \left\{ y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \ln \left[\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right] \right\}. \end{aligned} \tag{5.26}$$

This function behaves in much the same way as the residual sum of squares or weighted residual sum of squares in ordinary linear models. The addition of further covariates has the effect of reducing D.

It is often claimed that the random variable $D(Y; \hat{\pi})$ is asymptotically or approximately distributed as χ^2_{N-k}, where k is the number of fitted parameters. This claim is then used to justify the use of D as a goodness of fit statistic for testing the adequacy of the fitted model. Proofs of the limiting χ^2_{N-k} distribution are based on the following assumptions:

1. The observations are distributed independently according to the binomial distribution. In other words, the possibility of over-dispersion is not considered.

2. The approximation is based on a limiting operation in which N is fixed, $n_i \to \infty$ for each i, and in fact $n_i \pi_i (1 - \pi_i) \to \infty$.

In the limit given by the assumption 2, D is approximately independent of the estimated parameters $\hat{\beta}$ and hence approximately independent of the fitted probabilities $\hat{\pi}$. Approximate independence is essential for D to be considered as a goodness of fit statistic, but this property alone does not guarantee good power.

If N is large and $n_i\pi_i(1-\pi_i)$ remains bounded, the whole theory breaks down in two ways. First, the limiting χ^2 approximation no longer holds. Second, and more importantly, D is not independent of $\hat{\pi}$ even approximately. As a consequence, a large value of D could be obtained with high probability by judicious choice of β and π. In other words, a large value of D cannot necessarily be considered to be evidence of a poor fit.

The deviance function is most directly useful not as an absolute measure of goodness of fit but for comparing two nested models. For istance, we may wish to test wheter the addition of a further covariate significantly improves the fit. Let f_0 denote the model under test and f_1 the extended model containing an additional covariate. The corresponding fitted values are denoted by $\hat{\mu}_0$ and $\hat{\mu}_1$ respectively. The reduction in deviance

$$D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1) = 2L(\hat{\mu}_1;y) - 2L(\hat{\mu}_0;y) \tag{5.27}$$

is identical to the likelihood-ratio statistic for testing f_0 against f_1. This statistic is distributed approximately like χ^2_1 independently of $\hat{\mu}$ under assumption 1 above provided that either N is large or that assumption 2 is satisfied. In particular, $D(Y;\hat{\mu}_0)$ need not have an approximate χ^2 distribution nor need it be distributed independently of $\hat{\mu}_0$. The χ^2 approximation is usually quite accurate for differences of deviances even though it is inaccurate for the deviances themselves.

5.4.2 The Pearson's chi-squared statistic

Consider the *Pearson residual*, defined by

$$X = \frac{y-\mu}{\sqrt{V(\mu)}}. \tag{5.28}$$

It is just the raw residual scaled by the estimated standard deviation of Y.

The Pearson's chi-squared statistic is

$$X^2 = \sum X^2 = \sum_{i=1}^{N} \frac{(y_i-\mu_i)^2}{V(\mu_i)}. \tag{5.29}$$

For binomial case, the Pearson residual is

$$X = \frac{y-n_i\pi_i}{\sqrt{n_i\pi_i(1-\pi_i)}}, \tag{5.30}$$

so the X^2 statistic is

$$X^2 = \sum X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}. \tag{5.31}$$

This statistic is equivalent to the weighted residual sum of squares.

When X^2 is evaluated at the estimated expected frequencies, the statistic is

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}. \tag{5.32}$$

which is asymptotically equivalent to the deviance

$$D = 2\sum_{i=1}^{N} \left\{ y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\ln\left[\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right] \right\}. \tag{5.33}$$

The proof of the relatioship between X^2 and D uses the Taylor series expansion of $s\ln(s/t)$ about $s = t$. The asymptotic distribution of D, under the hypothesis that the model is correct, is $D \sim \chi^2_{N-k}$, therefore approximately $X^2 \sim \chi^2_{N-k}$. The choice between D and X^2 depends on the adequacy of the approximation to the χ^2_{N-k}. There is some evidence to suggest that X^2 is often better than D because D is excessively influenced by very small frequencies. Both the approximations are likely to be poor, however, if the expected frequencies are too small (e.g. less than 1).

5.4.3 The likelihood ratio chi-squared statistic

Sometimes the log-likelihood function for the fitted model is compared with the log-likelihood function for the null model, in which the values π_i are all equal (in contrast to the full model which is used to define the deviance). Under the null model

$$\tilde{\pi} = \frac{\sum y_i}{\sum n_i}, \tag{5.34}$$

so the log-likelihood for this model is

$$L(\tilde{\pi}; y) = \sum_i \left\{ y_i \ln\frac{\sum y_i}{\sum n_i} + (n_i - y_i)\ln\left(1 - \frac{\sum y_i}{\sum n_i}\right) \right\}. \tag{5.35}$$

Let $\hat{\pi}_i$ denote the estimated probability for Y_i under the model of interest (so the fitted value is $\hat{y}_i = n_i\hat{\pi}_i$). The statistic is defined by

$$C = 2[L - L_{min}] = 2[L(\hat{\pi}; y) - L(\tilde{\pi}; y)], \tag{5.36}$$

where $L(\tilde{\pi}; y)$ denotes the maximum value of the log-likelihood function for the null model with linear predictor $\eta = \beta_0$ and $L(\hat{\pi}; y)$ is the corresponding value for a more general model $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. Thus

$$C = 2 \sum_{i=1}^{N} \left\{ y_i \ln \left(\frac{\hat{y}_i}{n_i \tilde{\pi}_i} \right) + (n_i - y_i) \ln \left[\frac{n_i - \hat{y}_i}{n_i - n_i \tilde{\pi}_i} \right] \right\}. \tag{5.37}$$

The approximate sampling distribution for C is χ^2_{k-1} if all the k parameters except the intercept term β_0 are zero. Otherwise C will have a non-central distribution. Thus C is a test statistic for the hyppthesis that none of the explanatory variables is needed for a parsimonious model.

5.4.4 The pseudo R^2

By analogy with the index of determination R^2 for multiple linear regression another statistic sometimes used is

$$\text{pseudo } R^2 = 1 - \frac{L(\hat{\pi}; y)}{L(\tilde{\pi}; y)} = 1 - \frac{L}{L_{min}}, \tag{5.38}$$

which represents the proportional improvement in the log-likelihood function due to the terms in the model of interest, compared to the minimal model.

The maximum value achievable from the pseudo R^2 is given by 1 minus the ration between the log-likelihood of the full model and the log-likelihood of the null model: this value represents the proportional improvement in the log-likelihood function from the model without covariates and the model with the maximum number of covariates.

$$\text{pseudo } R^2 = 1 - \frac{L(y; y)}{L(\tilde{\pi}; y)} = 1 - \frac{L_{max}}{L_{min}}. \tag{5.39}$$

5.4.5 R^2 based on the Kullback - Leibler divergence

A standard measure of the information content from observations in a density $f(y)$ is the expected information, or Shannon's entropy, $E[\ln f(y)]$.

This is the basis for the standard measure of discrepancy between two densities, the *Kullback - Leibler divergence* (or information divergence, or information gain, or relative entropy) (see Kullback [32]). The KL divergence is a natural distance measure from a

"true" probability distribution P to an arbitrary probability distribution Q. Typically P represents data, observations, or a precise calculated probability distribution. The measure Q typically represents a theory, a model, a description or an approximation of P.

For probability distributions P and Q of a discrete variable the KL divergence of Q from P is defined to be

$$K(P||Q) = 2E_P \ln \frac{P(i)}{Q(i)} = \sum_i P(i) \ln \frac{P(i)}{Q(i)}. \tag{5.40}$$

For distributions P and Q of a continuous random variable the summations give way to integrals, so that

$$K(P||Q) = 2E_P \ln \frac{P(i)}{Q(i)} = \int_{-\infty}^{\infty} p(y) \ln \frac{p(y)}{q(y)} dy, \tag{5.41}$$

where $p(y)$ and $q(y)$ denote the densities of P and Q and E_P denotes expectation with respect to the true density P. The term "divergence" rather than "distance" is used because it does not in general satisfy the simmetry and triangular properties of a distance measure. However, $K(P||Q) \geq 0$ with equality if $P \equiv Q$.

We consider now the density $f(y; y)$, for which the mean is set equal to the realized y. Then the KL divergence $K(y||\mu)$ can be defined as

$$K(y||\mu) = 2E_y \ln \left[\frac{f(y; y)}{f(y; \mu)} \right] = \int f(y; y) \ln \left[\frac{f(y; y)}{f(y; \mu)} \right] dy. \tag{5.42}$$

The random variable $K(y||\mu)$ is a measure of the deviation of y from the mean μ. For the exponential family, Hastie [33] shows that the expectation in (5.44) drops out and so

$$K(y||\mu) = 2 \ln \left[\frac{f(y; y)}{f(y; \mu)} \right]. \tag{5.43}$$

In the estimated model, with N estimated means $\mu_i = x_i' \hat{\beta}$, the estimated KL divergence between the N-vectors y and μ is equal to twice the difference between the maximum log-likelihood achievable, i.e. the log-likelihood in a full model with as many parameters as observations, $L(y; y)$, and the log-likelihood achieved by the model under investigation, $L(\mu; y)$:

$$K(y||\mu) = 2[L(y; y) - L(\mu; y)]. \tag{5.44}$$

Let μ_0 denote the N-vector with entries μ_0, the fitted mean from ML estimation of the

null model. We interpret $K(y||\mu_0)$ as the estimate of the information in the sample data on y potentially recoverable by inclusion by inclusion of regressors. It is the difference between the information in the sample data on y and the estimated information using μ_0, the best point estimate when data on regressors are not utilized, where information is measured by takinng expectation with respect to the observed value y. By choosing μ_0 to be the ML estimate, $K(y||\mu_0)$ is maximized. The proposed R^2 is the proportionate reduction in this potentially recoverable information achieved by the fitted GLM:

$$R^2_{KL} = 1 - \frac{K(y||\hat{\mu})}{K(y||\hat{\mu}_0)} = 1 - \frac{2[L_{max} - L]}{2[L_{max} - L_{min}]} = 1 - \frac{D}{D_{min}}. \tag{5.45}$$

This measure can be used for fitted means obtained by any estimation method. For maximum-likelihood estimates of generalized linear models, based on the exponential density function (5.4), R^2_{KL} has the following properties:

1. R^2_{KL} is nondecreasing as regressors are added;
2. $0 \leq R^2_{KL} \leq 1$;
3. R^2_{KL} is a scalar multiple of the likelihood ratio test (deviance) for the joint significance of the explanatory variables;
4. R^2_{KL} equals the likelihood ratio index $1 - L(\hat{\mu}; y)/L(\hat{\mu}_0; y)$ if and only if $L(y; y) = 0$;
5. R^2_{KL} measures the proportionate reduction in recoverable information due to the inclusion of regressors, where information is measured by the Kullback-Leibler divergence.

Property 4 is of interest as the likelihood ratio index, which measures the proportionate reduction in the log-likelihood due to inclusion of regressors, is sometimes proposed as a general pseudo R^2 measure. Equality occurs for the Bernoulli model, but in general the likelihood ratio index differs and, for other discrete dependent variable models, is more pessimistic regarding the contribution of regressors, as $L(y; y) \leq 0$. In the continuous case, large values (positive or negative) of the likelihood ratio index can arise if $L(\hat{\mu}_0; y)$ is close to zero (positive or negative). By contrast, R^2_{KL} will always be bounded by zero and one.

5.4.6 Residuals

For generalized linear models applied to binomial data, there are two main forms of residuals corresponding to the goodness of fit measures D and X^2. If there are N

different *covariate patterns* (i.e. observations with the same values of all the explanatory variables), then N residuals can be calculated. Let y_i denote the number of successes, n_i the number of trials and $\hat{\pi}_i$ the estimated probability of success for the ith covariate pattern.

The Pearson residuals

These residuals, already presented in Section 5.4.2, are

$$X_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \qquad i = 1, \ldots, N. \tag{5.46}$$

From Eq. 5.32, $\sum_i^N X_i^2 = X^2$, the Pearson's chi-squared goodness of fit statistic.

The deviance residuals

These residuals are

$$d_i = \text{ sign}(y_i - n_i\hat{\pi}_i)\sqrt{\left\{2\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]\right\}}, \tag{5.47}$$

where the term $\text{sign}(y_i - n_i\hat{\pi}_i)$ [1] ensures that d_i has the same sign as X_i.
From Eq. 5.26, $\sum_i^N d_i^2 = D$, the deviance.

These two kinds of residuals can be used for checking the adequacy of a model. For example, they should be plotted against each continuous explanatory variable in the model to check if the assumption of linearity is appropriate. Normal probability plots can also be used because the standardized residuals should have, approximately, the standard Normal distribution N(0,1), provided the numbers of observations for each covariate pattern are not too small.

5.5 An algorithm for fitting GLM

The maximum-likelihood estimates of the parameters β in the linear predictor η can be obtained by *iterative weighted least squares* (IWLS). In this regression the dependent

[1] The function *sign* takes a vector as argument and returns a vector with the signs of the corresponding elements of its argument (the sign of a real number is 1, 0, or -1 if the number is positive, zero, or negative, respectively).

variable is not y but z, a linearized form of the link function applied to y, and the weights are functions of the fitted values $\hat{\mu}$. The process is iterative because both the adjusted dependent variable z and the weight W depend on the fitted values, for which only current estimates are available. The procedure underlying the iteration is as follows.

Let $\hat{\eta}_0$ be the current estimate of the linear predictor, with corresponding fitted value $\hat{\mu}_0$ derived from the link function $\eta = g(\mu)$. Form the adjusted dependent variate with typical value

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0) \left(\frac{\partial \eta}{\partial \mu} \right)_0, \tag{5.48}$$

where the derivative of the link is evaluated at η_0 and the quadratic weight defined by

$$W_0^{-1} = \left(\frac{\partial \eta}{\partial \mu} \right)_0^2 V_0, \tag{5.49}$$

where V is the variance function of y.

Now regress z_0 on the covariate matrix X with weight W_0 to give new estimates $\mathbf{b}_1$ of the parameters; from these form a new estimate $\hat{\eta}_1$ of the linear predictor. Repeat until changes are sufficiently small.

Note that z is just a linearized form of the link function applied to the data, for to the first order

$$g(y) \simeq g(\mu) + (y - \mu) g'(\mu) \tag{5.50}$$

and the right-hand side is

$$\eta + (y - \mu) \frac{\partial \eta}{\partial \mu}. \tag{5.51}$$

The variance of Z is just W^{-1} (ignoring the dispersion parameter), assuming that η and μ are fixed and known.

A convenient feature of generalized linear models is that they have a simple starting procedure necessary to allow the iteration to get under way. This consists of using the data themselves as the first estimate of $\hat{\mu}_0$ and from this deriving $\hat{\eta}_0$, $(\partial \eta / \partial \mu)_0$ and V_0.

5.5.1 Justification of the fitting procedure

From the log-likelihood L, written in canonical form

$$L(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi), \tag{5.52}$$

we have

$$\frac{\partial L}{\partial \theta} = \frac{[y - b'(\theta)]}{a(\phi)} = \frac{(y - \mu)}{a(\phi)}. \tag{5.53}$$

Hence

$$\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \theta} \Big/ \frac{\partial \mu}{\partial \theta} = \frac{(y - \mu)}{V}, \tag{5.54}$$

since

$$\frac{\partial \mu}{\partial \theta} = b''(\theta) = \frac{V}{a(\phi)}. \tag{5.55}$$

Now

$$\frac{\partial L}{\partial \eta} = \frac{\partial L}{\partial \mu} \frac{\partial \mu}{\partial \eta}, \tag{5.56}$$

and finally

$$\frac{\partial L}{\partial \beta_i} = \frac{\partial L}{\partial \eta} \frac{\partial \eta}{\partial \beta_i} = \frac{\partial L}{\partial \mu} \frac{\partial \mu}{\partial \eta} x_i = \frac{(y - \mu)}{V} \frac{\partial \mu}{\partial \eta} x_i. \tag{5.57}$$

The maximum-likelihood equations for β_i are therefore given by

$$\sum \left(\frac{y - \mu}{V} \right) \frac{\partial \mu}{\partial \eta} x_i = \sum W (y - \mu) \frac{\partial \eta}{\partial \mu} x_i = 0 \tag{5.58}$$

for each variate x_i, with summation over the units.
Fisher's scoring method uses the expected value of the Hessian matrix, i.e.

$$E\left(\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} \right) \quad = \quad E\left(\frac{\partial}{\partial \beta_j} \left[\sum (Y - \mu) V^{-1} \frac{\partial \mu}{\partial \eta} x_i \right] \right) \tag{5.59}$$

which is equal to

$$E\left(\sum (Y - \mu) \frac{\partial}{\partial \beta_j} \left[V^{-1} \frac{\partial \mu}{\partial \eta} x_i \right] + \sum \frac{\partial}{\partial \beta_j} \sum (Y - \mu) V^{-1} \frac{\partial \mu}{\partial \eta} x_i \right). \tag{5.60}$$

Eq. 5.60 is the equal to

$$-\sum V^{-1}\left(\frac{\partial\mu}{\partial\eta}\right)^2 x_i x_j = -\sum W x_i x_j. \tag{5.61}$$

Thus given current estimates $\mathbf{b}$ of β, the method gives adjustements $\delta\mathbf{b}$ defined by

$$A\delta\mathbf{b} = \mathbf{c}, \tag{5.62}$$

where A is a kxk matrix given by

$$A_{ij} = \sum_t W_t x_{it} x_{jt} \tag{5.63}$$

and $\mathbf{c}$ is a kx1 vector given by

$$c_i = \sum_t W_t x_{it}(y_t - \mu_t)\frac{\partial\eta_t}{\partial\mu_t}. \tag{5.64}$$

Now

$$(A\mathbf{b})_i = \sum_j A_{ij} b_j = \sum_t W_t x_{it}\eta_t, \tag{5.65}$$

and therefore new estimates $\mathbf{b}^* = \mathbf{b} + \delta\mathbf{b}$ satisfy the equations

$$(A\mathbf{b}^*)_i = [A(\mathbf{b} + \delta\mathbf{b})]_i = \sum W_t x_{it}\left[\eta_t + (y_t - \mu_t)\frac{\partial\eta_t}{\partial\mu_t}\right], \tag{5.66}$$

and these have the form of linear weighted least-squares equations with weight

$$W = V^{-1}\left(\frac{\partial\mu}{\partial\eta}\right)^2 \tag{5.67}$$

and dependent variate

$$z = \eta + (y - \mu)\frac{\partial\eta}{\partial\mu}. \tag{5.68}$$

Note that simplification occurs for the canonical links: for these the expected value and the actual value of the Hessian matrix coincide, so that the Fisher's scoring method and the Newton-Raphson method reduce to the same algorithm. This comes about because the linear weight function $V^{-1}\partial\mu/\partial\eta$ in the maximum-likelihood equations is a constant and the first term in the expansion of the Hessian (5.60) is identically 0. Note also that $W = V$ for this case.

Finally, if the model is linear on the scale on which Fisher's information is constant,

i.e. $g'(\mu) = 1/\sqrt{V(\mu)}$, the vector of weights is constant and need not to be recomputed at each iteration.

5.6 Log-likelihood for binomial data

Let the responses $y_1, \ldots, y_N$ correspond to independent random variables $Y_1, \ldots, Y_N$ where Y_i is assumed to be binomially distributed with index n_i and parameter π_i. The log-likelihood considered as a function of the vector $\pi = (\pi_1, \ldots, \pi_N)$ is

$$L(pi; y) = \sum_{i=1}^{N} \left[y_i \ln \left(\frac{\pi_i}{1 - \pi_i} \right) + n_i \ln(1 - \pi_i) \right]. \tag{5.69}$$

The systematic part of the model specifies the relation between π and the experimental or observational conditions as summarized by the model matrix X of order Nxk. For generalized linear models this relationship takes the form

$$g(\pi_i) = \eta_i = \sum_j x_{ij}\beta_j; \qquad i = 1, \ldots, N, \tag{5.70}$$

so that Eq. 5.69 can be expressed as a function of the unknown parameters $\beta_1, \ldots, \beta_k$. In particular, if $g(\pi)$ is the logit function

$$g(\pi) = \ln \left(\frac{\pi}{1 - \pi} \right), \tag{5.71}$$

Eq. 5.69 becomes

$$L(\beta; y) = \sum_i \sum_j y_i x_{ij} \beta_j - [\sum_i n_i \ln(1 + \exp \sum_j x_{ij}\beta_j)], \tag{5.72}$$

where we have written $L(\beta; y)$ instead of $L(\pi(\beta); y)$.

The usual asymptotic results associated with statistics derived from Eq. 5.69 or Eq. 5.72 depend only on second-moment assumptions. Thus it is not essential to assume binomial variation and independence. It is sufficient to assume simply that

$$E(Y_i) = n_i \pi_i; \qquad i = 1, \ldots, N, \tag{5.73}$$

and that

$$Cov(Y_i, \ldots, Y_k) = \text{ diag}[n_i \pi_i (1 - \pi_i)]. \tag{5.74}$$

In particular, the observations need not to be integer-valued but we must have $0 \leq Y_i \leq n_i$.

The method of maximizing $L(\beta; y)$ is applicable with adjusted dependent variable

$$z = \eta + \left(\frac{y}{n} - \pi\right) \frac{\partial \eta}{\partial \pi}, \tag{5.75}$$

and quadratic weight function

$$W = \frac{n}{\pi(1-\pi)} \left(\frac{\partial \eta}{\partial \pi}\right)^2. \tag{5.76}$$

For the logit function, which is the canonical link, these simplify to

$$z = \eta + \frac{y - n\pi}{n\pi(1-\pi)}, \tag{5.77}$$

and

$$W = n\pi(1-\pi). \tag{5.78}$$

The approximate covariance matrix of $\hat{\beta}$ is $(X'WX)^{-1}$.

5.6.1 Parameter estimation

Parameter estimates denoted by $\hat{\beta}$ are obtained by maximizing the log-likelihood over the space specified by the systematic part of the GLM.

A simple method of wider applicability is to use the Newton-Raphson method with the second-derivative matrix replaced by its expected value. The diagonal matrix of weights for the linear logistic model is given by

$$W^{-1} = \text{ diag} \left[\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}\right], \tag{5.79}$$

and the vector of adjusted dependent variates has elements

$$z_i = \hat{\eta}_i + \frac{y_i - n_i\hat{\pi}_i}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{5.80}$$

where $\eta_i = \ln[\pi_i/(1-\pi_i)]$. The iterative process can therefore be written in terms of the model matrix X as

$$X'WX\hat{\beta} = X'WZ \tag{5.81}$$

where W and Z are computed as functions of the current estimate of β.

If the systematic part of the model is linear on the probit or other scale, the weighting matrix W and adjusted dependent variate z must be changed accordingly, while the form of the iterative equations (5.81) remains unaltered.

Convergence of the iterative process

The convergence of process (5.81) is rarely a problem unless one or more elements of $\hat{\beta}$ are infinite. This can occur, for example, when the data are sparse and, for some observations, $y_i = 0$ or $y_i = n_i$. Although the iterative process will not converge under these circumstances, nevertheless generally the jth iterate $\hat{\pi}^{(j)}$ tends quite rapdily towards $\hat{\pi}$ and the deviance tends towards its limiting value. Thus the fitted values $n_i\hat{\pi}$ will be accurate bu the parameter estimates $\hat{\beta}$ may not be. The criterion used for deciding wheter the process has converged should be based on $\hat{\pi}^{(j+1)} - \hat{\pi}^{(j)}$ (e.g. by using the deviance) rather than on $\hat{\beta}^{(j+1)} - \hat{\beta}^{(j)}$.

Some results concerning the existence and uniqueness of the parameter estimates $\hat{\beta}$ have been given by Wedderburn [34]. These results show that if the link function $g(\pi)$ is log concave, as it is for the three functions logit, probit and complementary log-log, and if $0 < y_i < n_i$ for each i, then $\hat{\beta}$ is finite and $L(\pi; y)$ has a unique maximum at $\hat{\beta}$.

Starting values $\hat{\beta}_0$ can be obtained beginning with fitted values defined by $\hat{\mu} = (y + 0.5)/(n + 1)$. A good choice of starting value usually reduces the number of cycles in (5.81) by about one or perhaps two. Consequently, the choice of initial estimate is usually not critical.

5.6.2 Asymptotic theory for grouped data

We are concerned here with the asymptotic distribution of the parameter estimates and likelihood-ratio statistics. A very careful analysis would require the consideration of a hypothetical sequence of problems as the elements of the binomial index vector $\mathbf{n}$ tend to infinity. However, we take the scalar n to represent a typical binomial index, say $n = \min(n_1, \ldots, n_N)$ such that as $n \to \infty$ each element $n_i \to \infty$ in constant proportion. The number of distinct binomial observations, N, is considered fixed.

The principal results given below refer to generalized linear models where X, of order Nxk, is the model matrix and the weighting matrix W is diagonal with elements

$$W = \text{ diag}\left[\frac{1}{n\pi(1-\pi)}\left(\frac{\partial\pi}{\partial\eta}\right)^2\right]. \tag{5.82}$$

The dispersion parameter $\phi = \sigma^2$, included for generality, should be equal to 1 for binomial data. The asymptotic distribution of $\hat{\beta}$ is given by

$$\sqrt{n}(\hat{\beta} - \beta) \sim N[0, n\sigma^2(X'WX)^{-1}] + O_k(1/\sqrt{n}), \tag{5.83}$$

assuming, as usual, the adequacy of the model. The error term in (5.83) means that the cumulative distribution of $\sqrt{n}(\hat{\beta}-\beta)$ differs from the cumulative normal distribution by a term of order $1/\sqrt{n}$. In other words probability calculations based on the Normal approximation 5.83 have an error of order $1/\sqrt{n}$. It is also possible to show that the bias in $\hat{\beta}$ is $O(1/n)$.

An important requirement in (5.83) is that k, the number of parameters, should remain fixed as $n \to \infty$. In practice this means that k should not be a large fraction of N, particularly if some of the binomial denominators are small.

The second major asymptotic result concerns the distribution of the deviance, $D(Y; \hat{\pi})$. The null distribution is given by

$$D(Y; \hat{\pi}) \sim \sigma^2 \chi^2_{N-k} + O_k(1/\sqrt{n}), \tag{5.84}$$

where as before $\sigma^2 = 1$ for binomial data.

For over-dispersed data is necessary to find an estimate of σ^2 in order to use (5.83) For this, there are several possibilities, including the estimator $D(Y; \hat{\pi})/(N-k)$ based on the deviance; consistency as $n \to \infty$ follows from (5.84). However, it is preferable the estimator

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{N-k} \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{[n_i\hat{\pi}_i(1-\hat{\pi}_i)]} \\ &= \frac{X^2}{N-k}, \end{aligned} \tag{5.85}$$

where X^2 is *Pearson's chi-squared statistic.* The main reason for preferring $\tilde{\sigma}^2$ is that it is consistent in the limit as $N \to \infty$ with n fixed and its aymptotic distribution is known to be

$$\frac{\sigma^2 \chi^2_{N-k}}{N-k} + O_k(1/\sqrt{n}). \tag{5.86}$$

Thus $\tilde{\sigma}^2$ is a satisfactory estimator in either limit. Furthermore it can be shown that

as $n \to \infty$, $\hat{\beta}$ and $\tilde{\sigma}^2$ are asymptotically independent.

Approximate confidence intervals for an element, β_1 say, of β are formed in the usual way, namely

$$\hat{\beta}_1 \pm z_{\alpha/2}\sqrt{i^{(11)}} \tag{5.87}$$

if $\sigma^2 = 1$ is known and

$$\hat{\beta}_1 \pm \tilde{\sigma} t_{\alpha/2,N-k}\sqrt{i^{(11)}} \tag{5.88}$$

if σ^2 is unknown. Here $i^{(11)}$ is the [1,1] element of $(X'WX)^{-1}$, $\Phi(z_\alpha) = 1-\alpha$ and $t_{\alpha/2,N-k}$ is the $100(1-\alpha/2)\%$ quantile of the t-distribution with $N-k$ degrees of freedom. The intervals given above have coverage probability $1-\alpha+O(1/\sqrt{n})$, the probability in each tail being $\alpha/2+O(1/\sqrt{n})$.

5.7 Age-dependent prevalence and force of infection

As we have already seen previously, in a serological survey we consider an age-specific cross-sectional prevalence sample of size N (the covariate patterns) and let a_i be the age of the ith subject. Instead of observing the age at infection, we compose a set of current status data by observing binary variables Y_i such that

$$Y_i = \begin{cases} 0 & \text{if subject } i \text{ had experienced infection before age } a_i \\ 1 & \text{otherwise.} \end{cases} \tag{5.89}$$

With $F(a_i)$ be the probability to be infected before age a_i, $F(a_i) = 1 - P(a_i)$, the log-likelihood is given by

$$L(\beta; Y) = \sum_{i=1}^{N} Y_i \ln[F(a_i)] + (1-Y_i)\ln[1-F(a_i)]. \tag{5.90}$$

Here, $F(a) = g^{-1}(\eta(a))$, where g is the link function and $\eta(a)$ is the linear predictor.

To estimate the force of infection $\ell(a)$, we use the definition for the hazard rate (see Section 1.3.1):

$$\ell(a) = -\frac{\partial P(a)}{\partial a}\frac{1}{P(a)} = \frac{\partial F(a)}{\partial a}\frac{1}{1-F(a)}. \tag{5.91}$$

The logit link

If we use the logit link, we have the catalytic model

$$F(a) = \frac{e^{\eta(a)}}{1 + e^{\eta(a)}}. \tag{5.92}$$

The force of infection will be determined in the following way:

$$\begin{aligned} \ell(a) &= \frac{F'(a)}{1 - F(a)} \\ &= \frac{\eta'(a)e^{\eta(a)}(1 + e^{\eta(a)}) - e^{\eta(a)}(\eta'(a)e^{\eta(a)})}{(1 + e^{\eta(a)})^2}(1 + e^{\eta(a)}) \\ &= \frac{\eta'(a)e^{\eta(a)}(1 + e^{\eta(a)} - e^{\eta(a)})}{1 + e^{\eta(a)}} \\ &= \eta'(a)\frac{e^{\eta(a)}}{1 + e^{\eta(a)}}. \end{aligned} \tag{5.93}$$

The complementary log log link

Using the complementary log log link instead, the catalytic model of the prevalence is

$$F(a) = 1 - e^{-e^{\eta(a)}}. \tag{5.94}$$

Now, the force of infection will be

$$\begin{aligned} \ell(a) &= \frac{F'(a)}{1 - F(a)} \\ &= \frac{-e^{-e^{\eta(a)}}(-e^{\eta(a)})(\eta'(a))}{e^{-e^{\eta(a)}}} \\ &= \eta'(a)e^{\eta(a)}. \end{aligned} \tag{5.95}$$

The probit link

Finally, using the probit link, the catalytic model of the prevalence is

$$F(a) = \Phi[\eta(a)] = \int_{-\infty}^{\eta(a)} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \tag{5.96}$$

and the force of infection is

$$\begin{aligned} \ell(a) &= \frac{\partial \Phi[\eta(a)]}{\partial a_i} \frac{1}{1-\Phi[\eta(a)]} \\ &= \eta'(a) \frac{\phi[\eta(a)]}{1-\Phi[\eta(a)]}, \end{aligned} \tag{5.97}$$

where $\phi(\cdot)$ is the normal probability density function.

It is easy to see that for the binomial distribution, the force of infection can be expressed as a product of two functions:

$$\ell(a) = \eta'(a)\delta(\eta(a)), \tag{5.98}$$

where the form of $\delta(\cdot)$ is determined by the link function. Tab. 5.2 summaries the three link functions presented above with their corresponding structure for the force of infection.

Link function	$F(a)$	$\ell(a)$	$\delta(\eta(a))$
Logit	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\eta'(a)\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$
Complementary log log	$1-e^{-e^{\eta(a)}}$	$\eta'(a)e^{\eta(a)}$	$e^{\eta(a)}$
Probit	$\Phi[\eta(a)]$	$\eta'(a)\frac{\phi[\eta(a)]}{1-\Phi[\eta(a)]}$	$\frac{\phi[\eta(a)]}{1-\Phi[\eta(a)]}$

Table 5.2: General forms for the force of infection

For every dataset (mumps, rubella and parvovirus), we will show now the results of the fitting of a generalized linear model for prevalence. For every dataset we have fitted three models in accordance to a different link function $g(\cdot)$: the logit, the probit and the complementary log log link.

In the following subsections there are the results for the best model: for every dataset, the first table reports the estimated parameters and their significance test; the second one shows some measures of goodness of fit. These measures are listed below:

1. *Degrees of freedom* (d.f.): the first column of these tables reports the residual degrees of freedom of the model, which are $N-k$, where N is the number of covariate patterns and k is the number of parameters in the model.

2. *Deviance* (D): see Section 5.4.1. When comparing fitted models, the smaller the deviance, the better the fit.

3. *Pearson's chi-squared statistic* (X^2): see Section 5.4.2. When comparing fitted models, the smaller X^2, the better the fit.

4. *Likelihood ratio chi-squared statistic* (C): see Section 5.4.3. When comparing fitted models, the higher C, the better the fit.

5. *Pseudo* R^2: see Section 5.4.4. When comparing fitted models, the higher the pseudo R^2, the better the fit.

6. R^2 *based on the Kullback - Leibler divergence* (R^2_{KL}): see Section 5.4.5. When comparing fitted models, the higher this statistic, the better the fit.

Afterwards, we will plot the estimated force of infection for every dataset in accordance with the function determined at the beginning of this section.

5.7.1 Mumps: seroprevalence and force of infection

The seroprevalence estimated function

The best GLM for mumps has a logit link function:

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i, \tag{5.99}$$

where a_i is the age of the ith subject.
The prevalence $F(a)$ is given by

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i)}. \tag{5.100}$$

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	-0.8684	0.0603	-14.39	0.0000
$\hat{\beta}_1$	0.2247	0.0069	32.49	0.0000

Table 5.3: Mumps: summary of the estimated GLM-logit for prevalence

Fig. 5.1 shows the observed seropositive proportions and the plot of the estimated prevalence function.

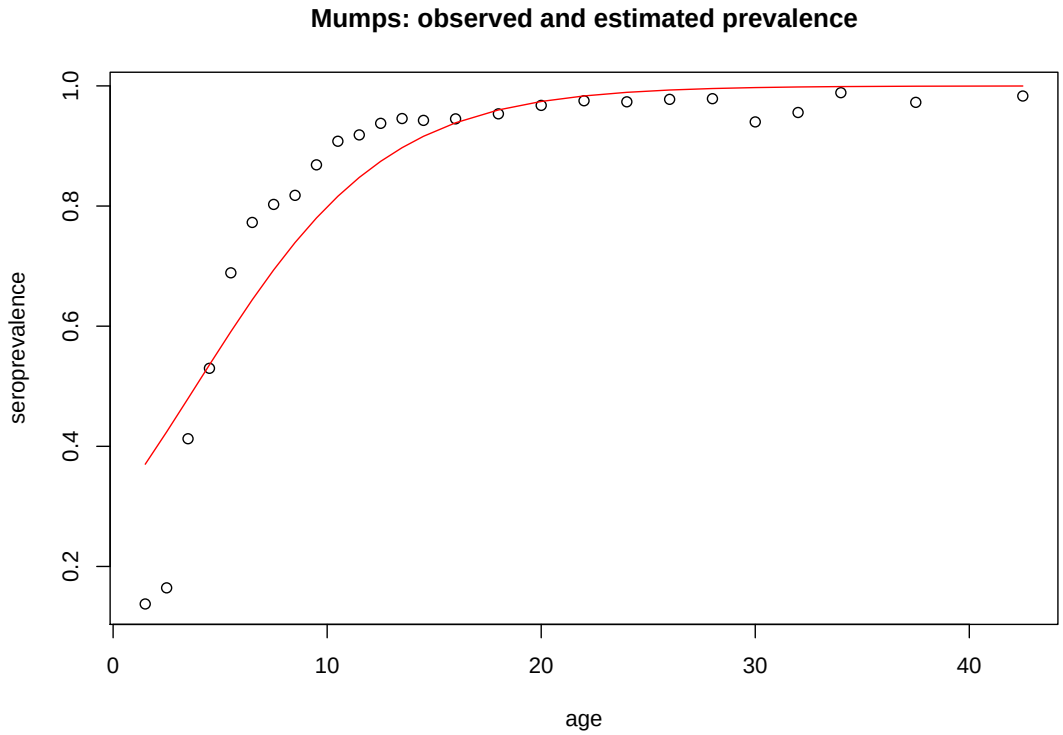


Figure 5.1: Mumps: observed and estimated prevalence with GLM-logit

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
24.00	581.37	1755.63	2342.17	0.2961	0.8011

Table 5.4: Mumps: measures of goodness of fit for the estimated GLM-logit for prevalence

As we can see from the measures of goodness of fit, but also from the graph, this logit model does not fit very well the observed seropositive proportion, above all in the first years of life until 15 years old: for example, the observed proportion in the first age class [0,1] is 0.14, while the respective estimated proportion is 0.37! For this fitted model the deviance is very high, i.e. about 581 on 24 degrees of freedom and the chi-squared goodness of fit statistics is about 1756 (24 d.f.). The R^2 constructed from the Kullback-Leibler divergence is 0.80, that is about 80% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.2961, while the maximum achievable is 0.3696.

The force of infection estimated function

Now we plot the estimated force of infection, whose function under a logit link model is

$$\ell(a_i) = \eta'(a_i)\frac{e^{\eta(a_i)}}{1+e^{\eta(a_i)}} = \hat{\beta}_1\frac{\exp(\hat{\beta}_0+\hat{\beta}_1 a_i)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 a_i)} = \hat{\beta}_1 F(a_i). \tag{5.101}$$

We assume that $\ell(0) = 0$, in accordance to the protection of maternal antibodies for the first year of life.

As we can see from Eq. 5.101, the force of infection for a linear logistic model is simply a multiple of the seroprevalence, so the model $\ell(a_i) = \hat{\beta}_1 F(a_i)$ predicts an upward trend for the force of infection.

5.7.2 Rubella: seroprevalence and force of infection

The seroprevalence estimated function

The best generalized linear model for rubella is under a logit link function, as it happens for mumps. That is the model function:

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i, \tag{5.102}$$

where a_i is the age of the ith subject, while the prevalence $F(a)$ is given by

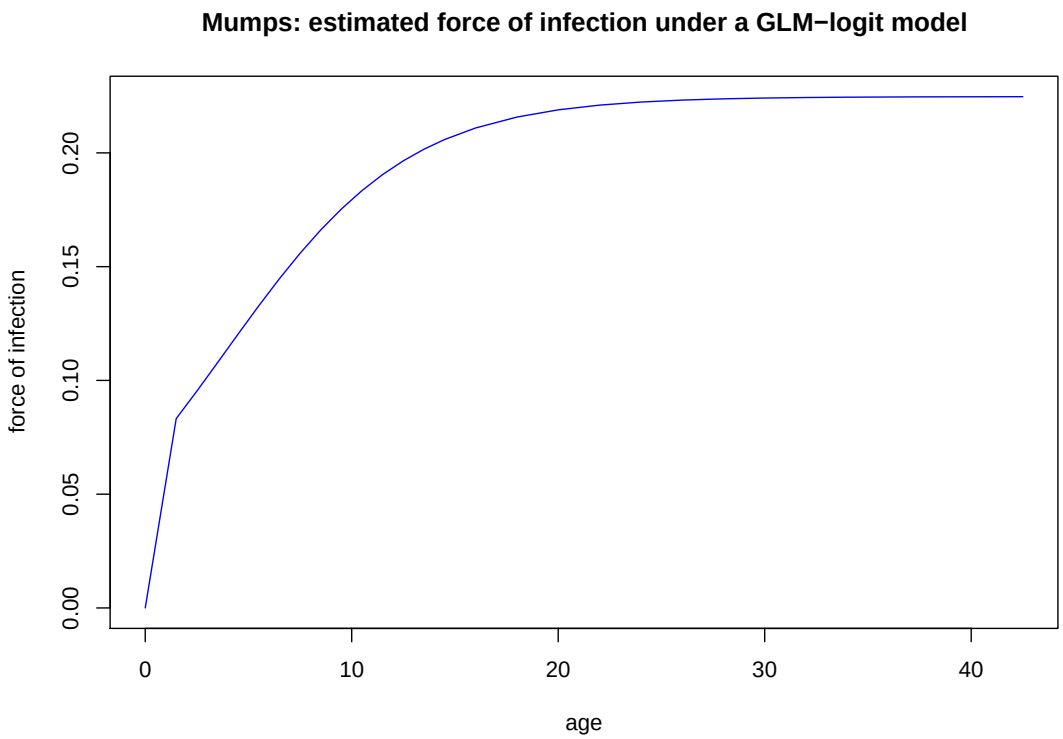


Figure 5.2: Mumps: estimated force of infection under a GLM-logit model

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i)}. \tag{5.103}$$

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	-1.0311	0.0692	-14.90	0.0000
$\hat{\beta}_1$	0.1468	0.0057	25.97	0.0000

Table 5.5: Rubella: summary of the estimated GLM-logit for prevalence

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
24.00	208.84	249.59	1148.05	0.2206	0.8461

Table 5.6: Rubella: measures of goodness of fit for the estimated GLM-logit for prevalence

Fig. 5.3 shows the observed seropositive proportions and the plot of the estimated prevalence function.

As we can see from the measures of goodness of fit, but also from the graph, this logistic model does not fit very well the observed seropositive proportion, above all in the first years of life until 20 years old: for example, the observed proportion in the first age class [0,1] is 0.14, while the respective estimated proportion is 0.31! For this fitted model the deviance is about 209 on 24 degrees of freedom and the chi-squared goodness of fit statistics is about 250 (24 d.f.). The R^2 constructed from the Kullback-Leibler divergence is 0.85, that is about 85% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.2206, while the maximum achievable is 0.2607.

The force of infection estimate function

Now we plot the estimated force of infection in Fig. 5.4, whose function under a logit link model is the same for mumps (see Eq. 5.101). We assume again that $\ell(0) = 0$, in accordance to the protection of maternal antibodies for the first year of life.

The force of infection is that of a linear logistic model again, so the model $\ell(a_i) = \hat{\beta}_1 F(a_i)$ predicts an upward trend for the force of infection for rubella too: indeed it reaches a peak at age 42.5 ($\ell(42.5) = 0.1460$).

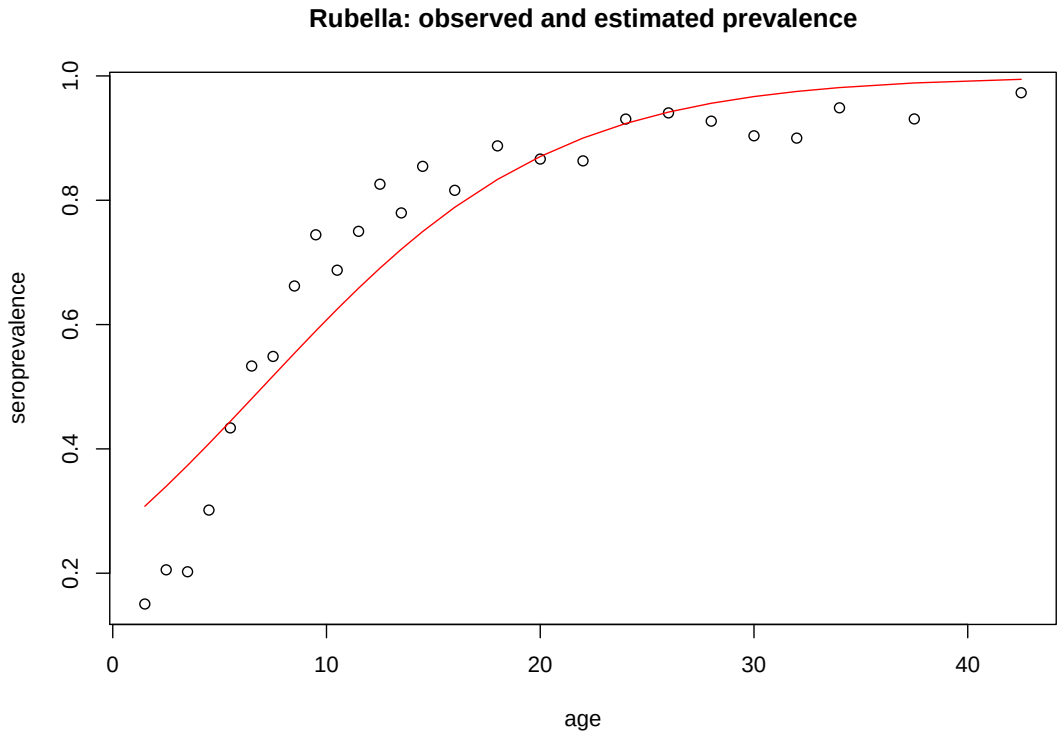


Figure 5.3: Rubella: observed and estimated prevalence with GLM-logit

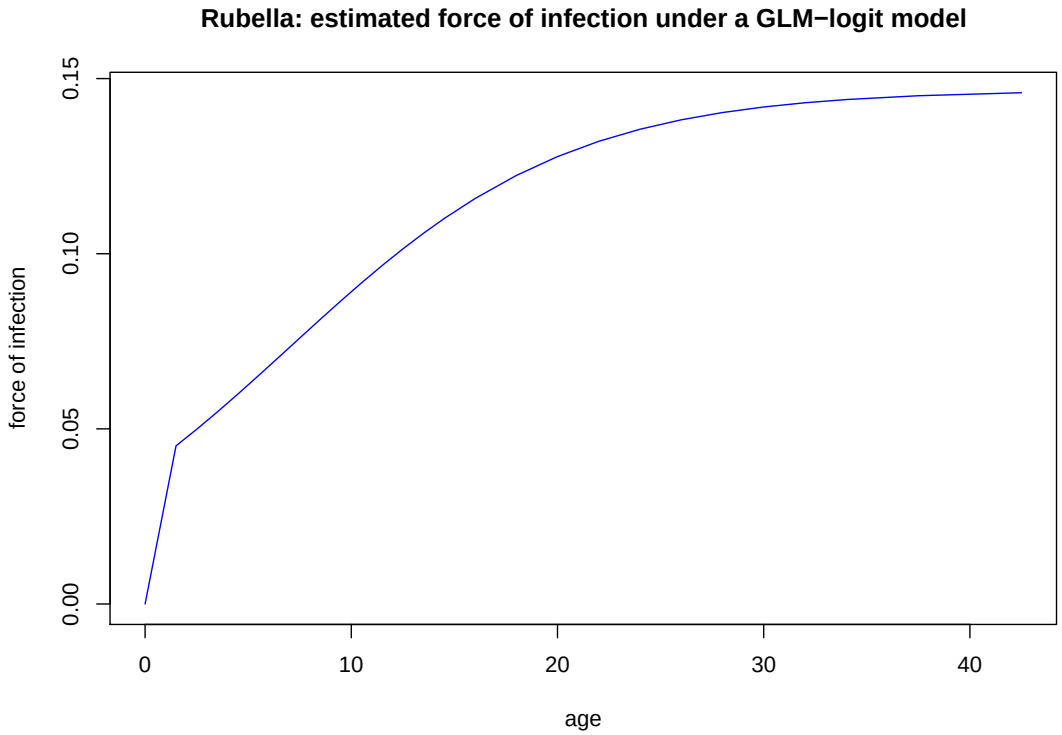


Figure 5.4: Rubella: estimated force of infection under a GLM-logit model

5.7.3 Parvovirus: seroprevalence and force of infection

The seroprevalence estimated function

The best GLM for parvovirus has a probit link function, differently from what happens for mumps and rubella:

$$\Phi^{-1}[F(a_i)] = \hat{\beta}_0 + \hat{\beta}_1 a_i, \tag{5.104}$$

where a_i is the age of the ith subject.

The prevalence $F(a)$ is given by

$$F(a_i) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 a_i) = \int_{-\infty}^{\hat{\beta}_0 + \hat{\beta}_1 a_i} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz. \tag{5.105}$$

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	-0.6175	0.0471	-13.11	0.0000
$\hat{\beta}_1$	0.0284	0.0021	13.53	0.0000

Table 5.7: Parvovirus: summary of the estimated GLM-probit for prevalence

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
24.00	118.97	115.17	187.12	0.0420	0.6113

Table 5.8: Parvovirus: measures of goodness of fit for the estimated GLM-probit for prevalence

Fig. 5.5 shows the observed seropositive proportions and the plot of the estimated prevalence function.

As we can see from the measures of goodness of fit, but most of all from the graph, this probit model fit very badly the observed seropositive proportion: the probit link function gives us a series of estimated proportions which lie all on a straight line and so the model is not able to explicate the non-linearity of the data. For this fitted model the deviance is about 119 on 24 degrees of freedom and the chi-squared goodness of fit statistics is about 115 (24 d.f.). As we have already told speaking about Farrington's prevalence fitted model, the measures D and X^2 are small because of the low values of the fitted probabilities. The R^2 constructed from the Kullback-Leibler divergence is 0.61, that is only about 61% of the information provided by the full model with respect

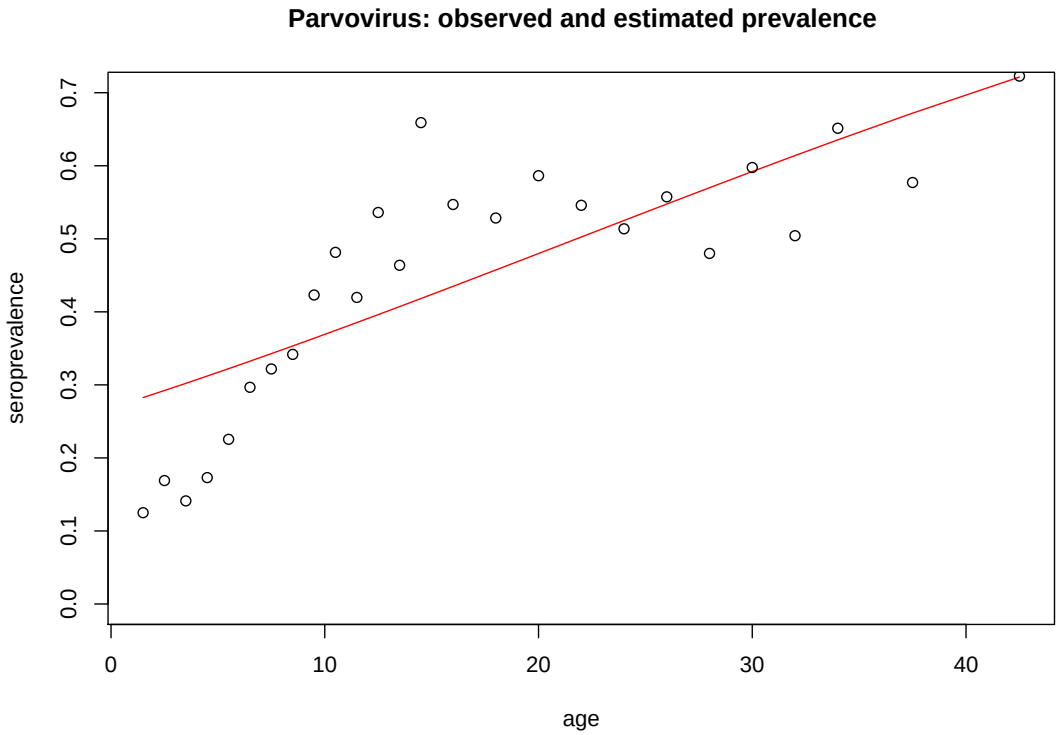


Figure 5.5: Parvovirus: observed and estimated prevalence with GLM-probit

to the null model is explicated by the fitted model. The pseudo R^2 is 0.0420, while the maximum achievable is 0.0687.

The force of infection estimated function

Now we plot the estimated force of infection, whose function under a probit link model is

$$\ell(a_i) = \hat{\beta}_1 \frac{\phi(\hat{\beta}_0 + \hat{\beta}_1 a_i)}{1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 a_i)}. \tag{5.106}$$

We assume again that $\ell(0) = 0$, in accordance to the protection of maternal antibodies for the first year of life.

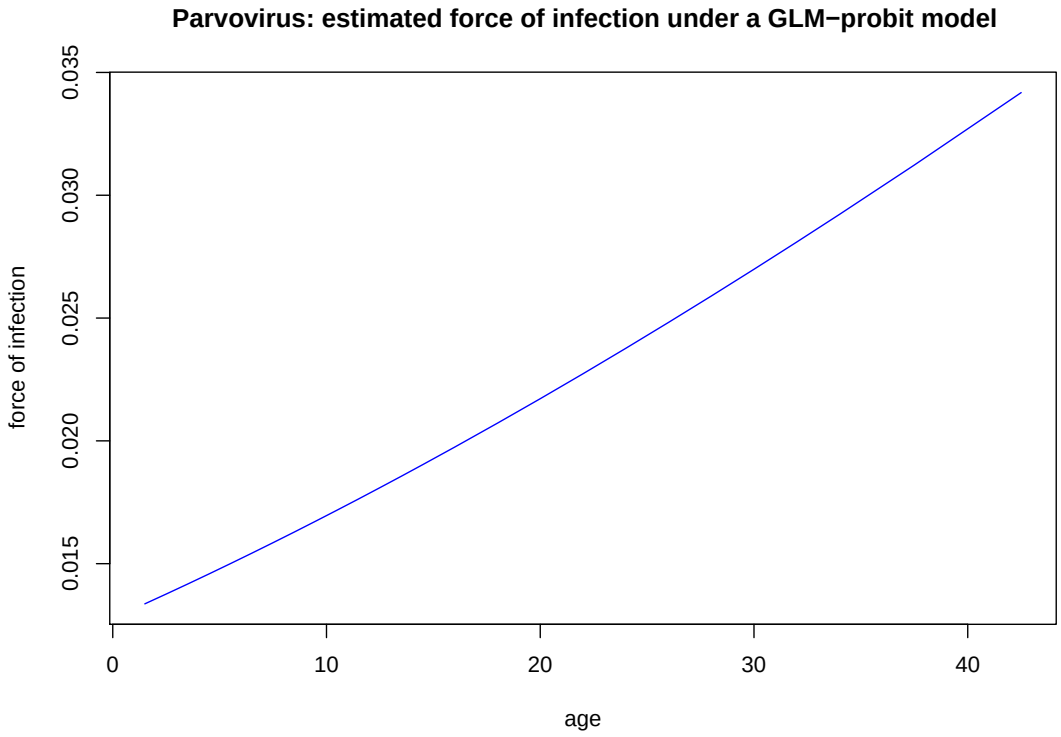


Figure 5.6: Parvovirus: estimated force of infection under a GLM-probit model

Fig. 5.6 show the plot of the estimated force of infection for parvovirus. It lies on a straight line, predicting an upward trend for it: the peak is reached at age 42.5 ($\ell(42.5) = 0.0342$).

5.8 Conclusions

The generalized linear models have been introduced in our analysis primarily because they furnish the theoretical basis for more complex models, as the fractional polynomials, which will be presented in the next chapter. Indeed the measures of the goodness of fit introduced in this chapter can be utilized for classes of models other than GLM: in this work, for example, they have been used for non-linear least squares models and for fractional polynomials too. This happens because these measures of discrepancy are based on the log-likelihood function. Given the responses $y_1, \ldots, y_N$ from a sample, corresponding to independent random variable $Y_1, \ldots, Y_N$, where Y_i is assumed to be distributed in accordance to a known density function belonging to the exponential family, it is always possible to construct the log-likelihood function. So in all these cases it is always possible to construct the goodness-of-fit measures presented in this chapter.

Tab. 5.9 compares all the measures of discrepancy for the three preceding prevalence models.

GLM Model	d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
Mumps	24.00	581.37	1755.63	2342.17	0.2961	0.8011
Rubella	24.00	208.84	249.59	1148.05	0.2206	0.8461
Parvovirus	24.00	118.97	115.17	187.12	0.0420	0.6113

Table 5.9: Comparing the measures of goodness of fit for mumps, rubella and parvovirus GLM models for prevalence

These measures seem to behave in an ambiguous way. For example, keeping fixed the degrees of freedom of the models, the deviance and the Pearson's chi-squared statistic for the parvovirus model are the lowest and this would mean that the model is a good one; however, the C statistic, the pseudo R^2 and R^2_{KL} give us the opposite message, because C says that the estimated log-likelihood is not so far from the log-likelihood of the null model, the pseudo R^2 tells us that the covariate "age" does not improve a lot the log-likelihood function compared to the null model and then R^2_{KL} says that the model explicates only 61% of the total information, measured by the Kullback-Leibler divergence.

The problem is that the deviance and the X^2 statistic are not independent of the fitted probabilities $\hat{\pi}$: in effect, the fitted seropositive proportions for parvovirus are not very high (the maximum is 0.72, while the maximum for mumps and rubella is very next to 1) and thus the deviance and the Pearson statistic have low values, but this does not necessarily mean that the model is better than the one for mumps or rubella.

Therefore, other measures as the pseudo R^2 and the R^2 derived from the Kullback-Leibler divergence seem to test better the adequacy of a fitted model.

Chapter 6

Fractional Polynomials

The relationship between a response variable and one or more covariates is often non-linear. Attempts to represent curvature with regression models are usually made using polynomials of the covariates, typically quadratics. However low order polynomials offer a limited family of shapes, and high order polynomials may fit poorly at the extreme values of the covariates. A further disadvantage is that polynomials don't have asymptotes and cannot fit data where limiting behavior is expected. Royston and Altman [35] proposed a family of curves, called *Fractional Polynomials (FP)* whose power terms are restricted to a small predefined set of number. The powers are selected so that conventional polynomials are a subset of the family. A great advantage of FP is that they are shown to have flexibility and are straightforward to fit using standard method.

The link between conventional polynomials and the modern methods of nonparametric smoothing is represented by *cubic spline*. Splines were originally developed in the 1920s. It was largely used as a method for fitting curves to data. The basic idea is that a knot is placed at each data point and a parameter is used to control the degree of of smoothing. Nonparametric scatterplot smoothers are an attempt to "*let the data show us the appropriate function form*" rather than imposing a limited range of forms on the data. Typically the smoothers are constructed at each data point by weighted regression within a neighbourhood of the corresponding covariate value. The best known smoothers is Cleveland's **lowess** (**lo**cally **w**eighted **s**catterplot **s**moother).

Nonparametric and spline smoothers are powerful and flexible tools which impose few limitations on the functional form.

6.1 The Model

6.1.1 Fractional Polynomials

We aim to model a trend in a response variable Y in terms of covariate(s) $X = (X_1, X_2, \ldots, X_k)$; we restrict our analysis to *Generalized Linear Models* (GLM), which include a random variable Y with mean μ, a model function $\eta = \eta(X, \beta)$ and a link function g such that $g(\mu) = \eta$.

A flexible model function is the additive predictor $\eta = f_0 + \sum f_j(X_j)$ where f_0 is a constant and f_j $(j > 0)$ is a function of X_j and a set of parameters. The linear predictor in a GLM is an additive predictor with $f_j(X_j) = \beta_j X_j \ \forall j$.

For example, a model incorporating a quadratic polynomial in X_j has the following form:

$$f_j(X_j) = \beta_{j1} X_j + \beta_{j2} X_j^2.$$

If each of the components $f_j(X_j)$ can be written in the form $\sum_i \beta_{ij} h_i(X_j)$, the model function η is then a linear predictor over the set of covariates $h_i(X_j)$.

We now describe a family of model functions of a single covariate X, subject to the restriction $X > 0$:

Definition 1 *A Fractional Polynomial of degree m is the function*

$$\phi_m(X; \xi, \mathbf{p}) = \xi_0 + \sum_{j=1}^{m} \xi_j X^{(p_j)} \tag{6.1}$$

where m is a positive integer, $\mathbf{p} = (p_1, \ldots, p_m)$ is a real vector of powers with $p_1 < \cdots < p_m$, $\xi = (\xi_0, \xi_1, \ldots, \xi_m)$ are real coefficients and $X^{(p_j)}$ is defined as the Box-Tidwell trasformation:

$$X^{(p_j)} = \begin{cases} X^{p_j} & p_j \neq 0 \\ \ln X & p_j = 0 \end{cases}$$

A conventional polynomial of degree m has $p_j = j$ for $j = 1, \ldots, m$ and $\xi_m \neq 0$. Definition 1 can be extended to the cases of equal power, i.e. $m > 1$ and $\mathbf{p} = (p_i, p_j)$. For $m = 2$ and $\mathbf{p} = (p_1, p_1)$, we have

$$\phi_m(X; \xi, \mathbf{p}) = \xi_0 + (\xi_1 + \xi_2) X^{(p_1)} \tag{6.2}$$

a fractional polynomial of degree 1, not 2. Now, let consider the standard equation

$$\phi_m(X;\xi^*,\mathbf{p}) = \xi_0^* + \xi_1^* X^{(p_1)} + \xi_2^* X^{(p_2)}. \tag{6.3}$$

Writing $\xi_0 = \xi_0^*$, $\xi_1 = \xi_1^* + \xi_2^*$, $\xi_2 = (p_2 - p_1)\xi_2^*$ and rearranging, we obtain $\xi_1^* = \xi_1 - \xi_2/(p_2 - p_1)$.

Substituting in Eq 6.3 we find

$$\xi_0 + \xi_1 X^{(p_1)} + \xi_2 X^{(p_1)}(X^{(p_2-p_1)} - 1)/(p_2 - p_1) \tag{6.4}$$

When p_2 tends to p_1, the limit of Eq 6.4 is

$$\xi_0 + \xi_1 X^{(p_1)} + \xi_2 X^{(p_1)} \log X \tag{6.5}$$

For $m > 2$ and $p_1 = \cdots = p_m$ expression 6.5 may be generalized in

$$\xi_0 + \xi_1 X^{(p_1)} + \sum_{j=2}^{m} \xi_j X^{(p_1)} (\ln X)^{j-1} \tag{6.6}$$

For arbitrary powers $p_1 \leq \ldots \leq p_m$, we set $H_0(X) = 1$, $p = 0$ and combine definition 6.1 and Eq. 6.6 to obtain an extended definition

$$\phi_m(X;\xi,\mathbf{p}) = \sum_{j=0}^{m} \xi_j H_j(X), \tag{6.7}$$

where for $j = 1, \ldots, m$

$$H_j(X) = \begin{cases} X^{(p_j)} & p_j \neq p_{j-1} \\ H_{j-1}(X) \ln X & p_j = p_{j-1}. \end{cases} \tag{6.8}$$

As an example of Eq. 6.7, $\phi_5(X;0,1,2,2,2)$ has component functions $H_0 = 1$, $H_1 = \ln X$, $H_2 = X$, $H_3 = X^2$, $H_4 = X^2 \ln X$ and $H_5 = X^2(\ln X)^2$; so, ϕ_5 has the following form:

$$\phi_5(X;0,1,2,2,2) = \xi_0 + \xi_1 \ln X + \xi_2 X + \xi_3 X^2 + \xi_4 X^2 \ln X + \xi_5 X^2 (\ln X)^2. \tag{6.9}$$

If non-positive values of X can occur, a preliminary transformation of X to ensure positivity is needed. Onw solution is to choose a non-zero origin $\zeta < X$ and to rewrite Eq. 6.7 as

$$\phi_m(X;\xi,\mathbf{p}) = \sum_{j=0}^{m} \xi_j H_j(X-\zeta). \tag{6.10}$$

6.1.2 Fractional Polynomials of Degree 1 and Degree 2

It is worth considering the families $\phi_1(X;\mathbf{p})$ and $\phi_2(X;\mathbf{p})$, for we have so far found that models with degree higher than 2 are rarely required in practice. Fractional polynomials with $m \geq 2$ offer many potential improvements in fit compared with conventional polynomials (see Fig. 6.1)

6.2 Termination Rules

6.2.1 Fractional Polynomials as Model Functions

Conditional on given values of m and $\mathbf{p}$, $\phi_m(X;\mathbf{p})$ in Eq. 6.7 has the form of a linear predictor in terms of the covariate vector $\mathbf{H}(X)$ and the parameter vector ξ.

For modelling a data set of size n using fractional polynomials, the Authors propose to determine the "best" value of m and of the power vector $\mathbf{p}$ by criteria to be discussed in the following section.

Candidate values of $\mathbf{p}$ are all possible m-tuples selected with replacement from a fixed set $\mathcal{P}$.

Experience so far suggests that $\mathcal{P} = \{-2,-1,-0.5,0,0.5,1,2,3\}$, which includes all conventional polynomials of degree less than or equal to m, is sufficiently rich to cover many practical cases adequately.

As with conventional polynomials, the degree m is selected either informally on *a priori* grounds or by increasing m until no worthwhile improvement in the fit of the best fitting fractional polynomial is judged to have occured.

6.2.2 Deviance and Model Choice

We assume that all models are to be fitted by *maximum likelihood*. For given m, the best power vector $\tilde{\mathbf{p}} = \{\tilde{p_1},\ldots,\tilde{p_m}\}$ is that associated with the model with the highest likelihood or, equivalently, with the lowest deviance D. Thus $\tilde{\mathbf{p}}$ may be regarded as the *maximum likelihood estimate* (MLE) of $\mathbf{p}$ over the restricted parameter space based on $\mathcal{P}$. We use the deviance, as defined in Eq. 5.26, that is to say twice the difference between the log-likelihood of the full model and that of the model under investigation.

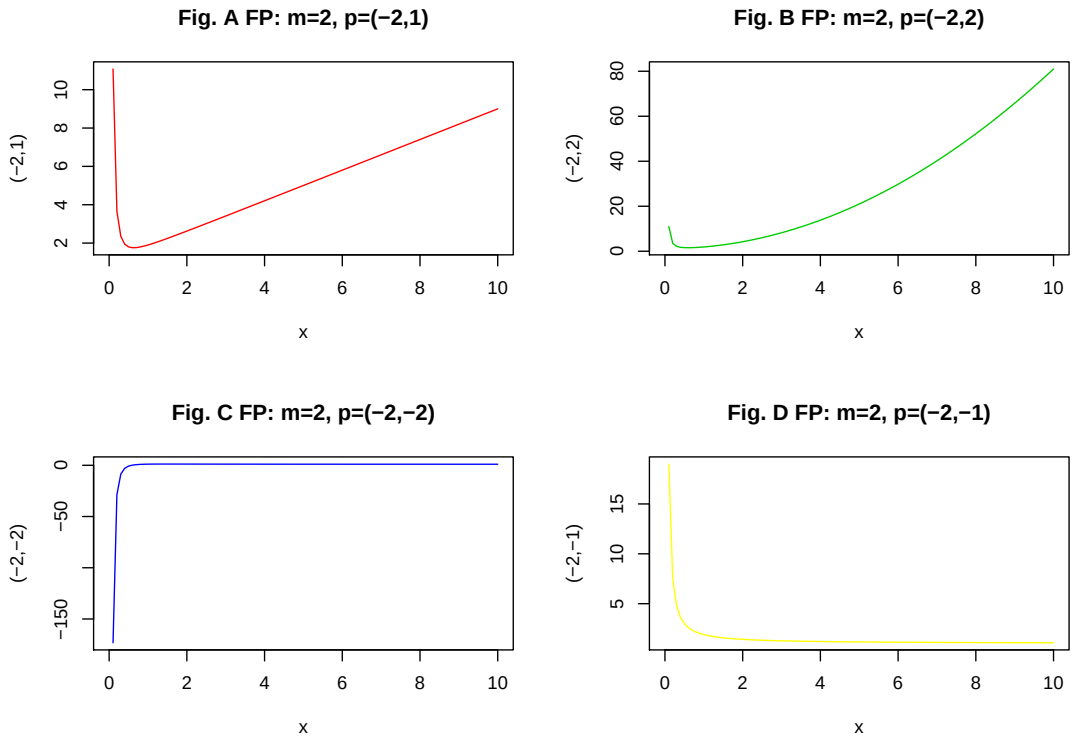


Figure 6.1: Examples of $\phi_2(X;p)$, for $p = (-2,1), (-2,2), (-2,-2)$ and $(-2,-1)$

Suppose that the elements of $\mathbf{p}$ are allowed to vary continuously, rather than being restricted to $\mathcal{P}$. Then $\phi_m(X;\mathbf{p})$ is a nonlinear model with parameters $\mathbf{p}$ and ξ. Let $\hat{\mathbf{p}}$ be the full MLE of $\mathbf{p}$. So, the quantity $D(m,\mathbf{p}) - D(m,\hat{\mathbf{p}})$ asymptotically has a χ^2 distribution on m degrees of freedom (DF).

Model Choice: the best power vector p

The statistic $D(m,\mathbf{p}) - D(m,\tilde{\mathbf{p}})$, where $\tilde{\mathbf{p}}$ is the best power vector, provides an (aymptotically conservative) test of a given value of $\mathbf{p}$ and it may be used as a guide to the adequacy of the conventional polynomial of degree m against fractional polynomial alternatives of the same degree.

For example, if we want to investigate the nonlinearity of some observed data, the Authors suggest to use the following criterion:

$$D(1,1) - D(1,\tilde{p}) > \chi^2_{1;0.90}, \tag{6.11}$$

where $D(1,1)$ is the deviance of a simple linear model, $D(1,\tilde{p})$ is the deviance of a fractional polynomial alternative of degree 1 and $\chi^2_{1;0.90}$, whose value is 2.7055, is the 90th percentile of χ^2 with 1 DF. This criterion furnishes a test with a significance level α of about 0.10 for $p = 1$ (linearity) against $p \neq 1$ (monotonic alternatives): if the criterion is true, the model $\phi_1(X;\tilde{p})$ is better than the linear model and so we refuse the hypothesis of linearity of the data; otherwise, the linear model $\phi_1(X;1)$ is the best choice.

If we have to choose between several fractional polynomial models with similar deviances, the Authors suggest, as a working rule, to use the following criterion:

$$D(m,\mathbf{p}) - D(m,\tilde{\mathbf{p}}) < \chi^2_{m;0.90}, \tag{6.12}$$

where $D(m,\mathbf{p})$ is the deviance of the FP model we want to test and $D(m,\tilde{\mathbf{p}})$ is the best FP model. We choose the FP model which minimizes the criterion.

Model Choice: the best degree m of the model

When we have to decide whether models with degree m are adequate or whether degree $m+1$ is required, first of all we have to note that two extra parameters (a power and a regression coefficient) are estimated when m is increased by 1. Therefore $D(m,\hat{\mathbf{p}}) - D(m+1,\hat{\mathbf{p}})$ is asymptotically distributed as χ^2 on 2 DF when the degree m is adequate. The Authors suggest the following criterion as a rule for preferring models with degree $m+1$ to those with degree m:

$$D(m, \tilde{\mathbf{p}}) - D(m+1, \tilde{\mathbf{p}}) > \chi^2_{2;0.90}, \tag{6.13}$$

where $\chi^2_{2;0.90} = 4.6052$. We expect the probability of a type I error (the probability of refusing the null hypothesis when it is true) associated with this rule to be near (but not exactly) 10%.

In general terms, when working with fractional polynomial models, it is convenient to use the deviance $D(1,1)$, associated with the simple linear model $\phi_1(X;1)$, as a baseline for reporting the deviances of other models. Thus we define the *gain* G for a model on a given data set as the deviance for $\phi_1(X;1)$ minus that for the model in question:

$$G = G(m, \mathbf{p}) = D(1,1) - D(m, \mathbf{p}). \tag{6.14}$$

Since G moves in the opposite direction to D, a *larger* gain indicates a *better* fit.

Once m and acceptable power vectors $\mathbf{p}$ have been selected as just described, the final choice must depend mainly on the appearence of the curves in relation to the data, especially at the extremes of X. Non-statistical considerations (mainly, the science of the problem) may also need to be taken into account.

6.3 Age-dependent prevalence and force of infection

In the following section we will use fractional polynomial models to estimate the prevalence data for mumps, rubella and parvovirus.

Although fractional polynomials provide a wide range of curve shapes, there is no guarantee that the prevalence $F(a)$ will be a monotone function of age and therefore fractional polynomials can result in a negative estimate for the force of infection. It is clear from Tab. 5.2 that the estimate for the force of infection is negative whenever $\eta'_m(a, \hat{\beta}, p) < 0$ (since $\delta(\eta_m(a, \hat{\beta}, p))$ is strictly positive). Therefore, one should fit model (6.7) subject to the constraints that $\eta'_m(a, \hat{\beta}, p) \geq 0$, for all ages a in the predefined range. In the framework of fractional polynomials this cannot be done analitically. But in practice, one can fit a large number of fractional polynomials, over a grid of powers, and check for each fitted model if $\eta'_m(a, \hat{\beta}, p) \geq 0$, for all ages a. In case that a given sequence of powers leads to a negative derivative of the linear predictor, the model is not considered an appropriate model. This means that we choose the model with the best goodness of fit among all fractional polynomials for which $\eta'_m(a, \hat{\beta}, p) \geq 0$.

In the following sections, for each data set, first- and second-order fractional polynomial models are fitted and the criterion proposed by Royston and Altman [35] is used to decide whether the second-order model is needed or not.

6.4 Analysis for mumps data

In this section, we show the results of the application of fractional polynomials to mumps data.

6.4.1 Fractional polynomial of degree 1 for mumps

We have started with a fractional polynomial of degree $m = 1$. Since the probability density function of the response variable Y is the Binomial distribution, we have implemented three models in accordance with three different link function: the logit link, the probit link and the complementary log-log link. Of course, the only covariate is the age of the individual, a. For every link function, the software R has determined the best power vector, the one that minimizes the deviance of the model.

For mumps, the best fractional polynomial of degree 1 has the logit link as link function and the best power for the covariate is $p = -0.2$. So the model $\phi_1(a; -0.2)$ is

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.2} \tag{6.15}$$

and the seroprevalence function is

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.2})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.2})}. \tag{6.16}$$

Tab. 6.1 reports the results of the parameter estimation performed by R [28], while Tab. 6.2 reports some measures of goodness of fit. Tables about the goodness of fit of the fitted models report the following measures:

1. *Degrees of freedom* (d.f.): these are the residual degrees of freedom of the model, which are $N - k$, where N is the number of covariate patterns and k is the number of parameters in the model.

2. *Deviance* (D): see Section 5.4.1. When comparing fitted models, the smaller the deviance, the better the fit.

3. *Pearson's chi-squared statistic* (X^2): see Section 5.4.2. When comparing fitted models, the smaller X^2, the better the fit.

4. *Likelihood ratio chi-squared statistic* (C): see Section 5.4.3. When comparing fitted models, the higher C, the better the fit.

5. *Pseudo* R^2: see Section 5.4.4. When comparing fitted models, the higher the pseudo R^2, the better the fit.

6. R^2 *based on the Kullback - Leibler divergence* (R^2_{KL}): see Section 5.4.5. When comparing fitted models, the higher this statistic, the better the fit.

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	11.4447	0.2569	44.55	0.0000
$\hat{\beta}_1$	−9.49 04	0.2336	−40.63	0.0000

Table 6.1: Mumps: summary of the estimated FP(m=1)-logit model for prevalence

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
24.00	65.40	74.89	2858.14	0.3613	0.9776

Table 6.2: Mumps: measures of goodness of fit for the estimated FP(m=1)-logit model for prevalence

Following the indications in Sec. 6.2.2, we evaluate the linear model $\phi_1(a;1)$, which is necessary for the study of the goodness of fit:

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i, \tag{6.17}$$

whose deviance $D(1,1)$ is 581.37.

The value of the gain G for the model $\phi_1(a;-0.2)$ is

$$\begin{aligned} G = G(1;-0.2) &= D(1;1) - D(1;-0.2) \\ &= 581.37 - 65.40 \\ &= 515.97. \end{aligned} \tag{6.18}$$

For this fitted model the deviance is about 65 on 24 degrees of freedom and the chi-squared goodness of fit statistics is about 75 (24 d.f.). G is 515.97, so the fitted model

is significantly different from the simple logistic model. The R^2 constructed from the Kullback-Leibler divergence is 0.98, that is about 98% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.3613, while the maximum achievable is 0.3696.

6.4.2 Fractional polynomial of degree 2 for mumps

Now, we will see a fractional polynomial of degree $m = 2$. The best fractional polynomial of degree 2 has the logit link as link function and the best power vector for the covariate a is $[-2, -0.8]$. So the model $\phi_2(a; -2, -0.8)$ is

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i^{-2} + \hat{\beta}_2 a_i^{-0.8} \tag{6.19}$$

and the seroprevalence function is

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-2} + \hat{\beta}_2 a_i^{-0.8})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-2} + \hat{\beta}_2 a_i^{-0.8})}. \tag{6.20}$$

Tab. 6.3 reports the results of the parameter estimation performed by R, while Tab. 6.4 reports some measures of goodness of fit.

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	4.7302	0.110931	42.64	0.0000
$\hat{\beta}_0$	0.1333	0.008971	14.86	0.0000
$\hat{\beta}_0$	−2.7421	0.094302	−29.08	0.0000

Table 6.3: Mumps: summary of the estimated FP(m=2)-logit model for prevalence

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
23.00	27.90	31.11	2895.64	0.3661	0.9904

Table 6.4: Mumps: measures of goodness of fit for the estimated FP(m=2)-logit model for prevalence

The value of the gain G for the model $\phi_2(a; -2, -0.8)$ is

$$
\begin{aligned}
G = G(2; -2, -0.8) &= D(1; 1) - D(2; -2, -0.8) \\
&= 581.37 - 27.90 \\
&= 553.47. \qquad (6.21)
\end{aligned}
$$

For this fitted model the deviance is about 28 on 23 degrees of freedom and the chi-squared goodness of fit statistics is about 31 (23 d.f.). G is 553.47, meaning that the fitted model is significantly different from the logistic model. The R^2 constructed from the Kullback-Leibler divergence is 0.99, that is about 99% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.3661, while the maximum achievable is 0.3696.

In Fig. 6.2 there are the plots of the two fractional polynomial models.

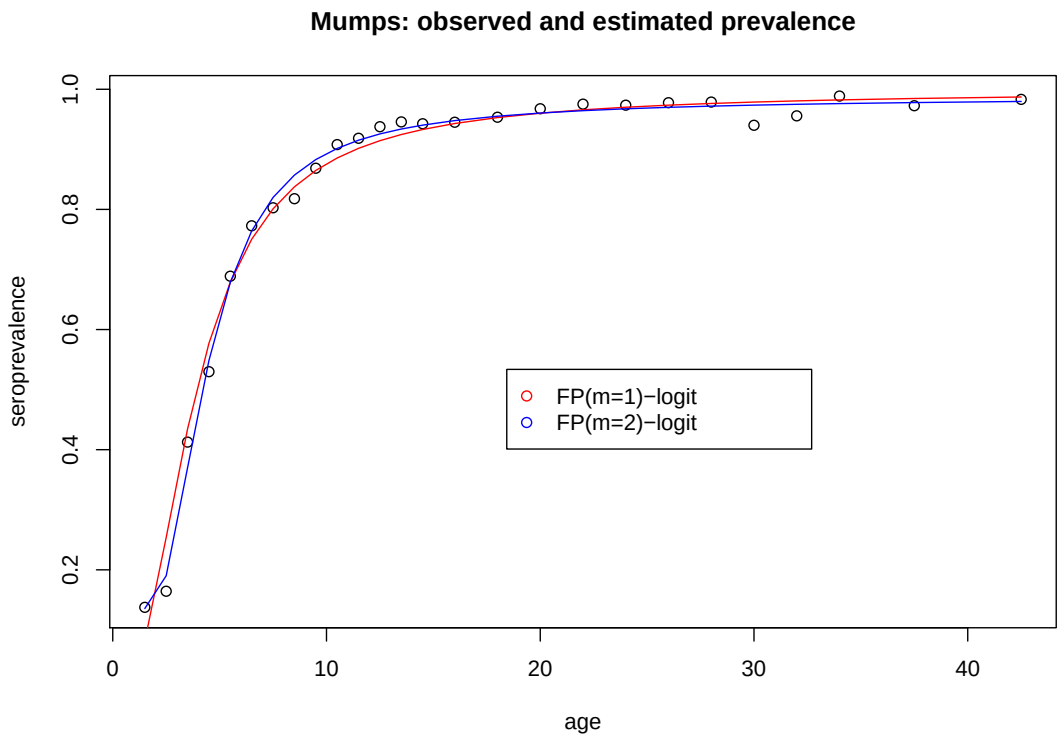


Figure 6.2: Mumps: observed and estimated prevalence with fractional polynomials of degree $m = 1$ and $m = 2$

We use now the criterion suggested by Royston and Altman [35] to choose between

the model of first degree and the model of second degree:

$$\begin{aligned} D(1;-0.1) - D(2;-1.2,-0.9) &= 65.40 - 27.90 \\ &= 37.50 > 4.6052 = \chi^2_{2,0.90}. \end{aligned} \tag{6.22}$$

Given that $D(1;-0.2) - D(2;-2,-0.8) > \chi^2_{2,0.90}$ and given that the plot of FP(m=2)-logit fits better observed data in the first age classes, we conclude that the best fractional polynomial model for mumps is that of degree 2.

6.4.3 Estimation of the force of infection

Now we plot the estimated force of infection for the best fractional polynomial model for prevalence, $\phi_2(a;-2,-0.8)$.

From the catalytic model, the function of the force of infection for the second-degree model is

$$\begin{aligned} \ell_2(a_i) &= \eta'(a_i)\frac{e^{\eta(a_i)}}{1+e^{\eta(a_i)}} \\ &= -(2\hat{\beta}_1 a_i^{-3} + 0.8\hat{\beta}_2 a_i^{-1.8})F(a_i). \end{aligned} \tag{6.23}$$

We also assume that $\ell_2(0) = 0$, accounting for the protective effect of maternal antibodies in the first year of life.

Fig. 6.3 shows the plot of the estimated force of infection for mumps. It reaches a peak at age 1.5 ($\ell_2(1.5) = 0.1325$) and drops down thereafter until $\ell_2 = 0.0025$ at age 42.5.

6.5 Analysis for rubella data

In this section, we show the results of the application of fractional polynomials to rubella data.

6.5.1 Fractional polynomial of degree 1 for rubella

In this case also, we start with a fractional polynomial of degree $m = 1$ and we implement three models for the preceding three link functions: logit, probit and complementary log-log.

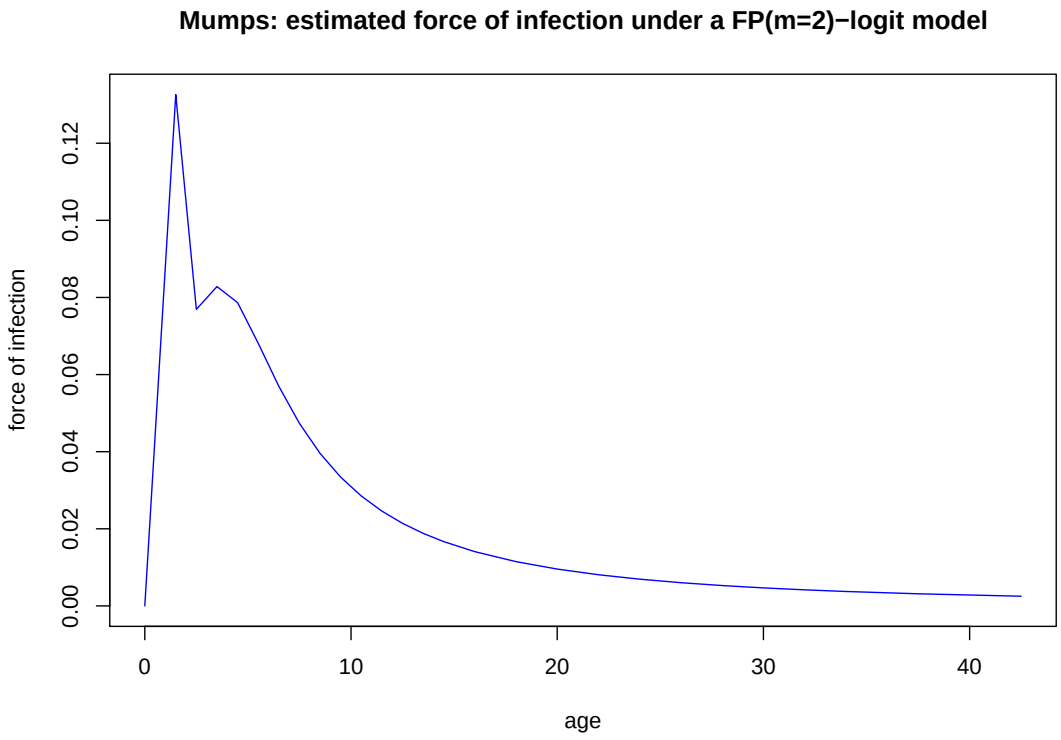


Figure 6.3: Mumps: estimated force of infection under a FP($m = 2$)-logit model for prevalence

For rubella, the best fractional polynomial of degree 1 has the logit link as link function and the best power for the covariate is $p = 0.1$. So the model $\phi_1(a; 0.1)$ is

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i^{0.1} \tag{6.24}$$

and the seroprevalence function is

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{0.1})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{0.1})} \tag{6.25}$$

Tab. 6.5 reports the results of the parameter estimation performed by R, while Tab. 6.6 reports some measures of goodness of fit.

The deviance $D(1, 1)$ for rubella is 208.84. So, the value of the gain G for the model $\phi_1(a; 0.1)$ is

$$\begin{aligned} G = G(1; 0.1) &= D(1; 1) - D(1; 0.1) \\ &= 208.84 - 44.22 \\ &= 164.62. \end{aligned} \tag{6.26}$$

For this fitted model the deviance is about 44 on 24 degrees of freedom and the chi-squared goodness of fit statistics is about 45 (24 d.f.). G is 164.62, meaning that the fitted model is significantly different from the logistic model. The R^2 constructed from the Kullback-Leibler divergence is 0.97, that is about 97% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.2522, while the maximum achievable is 0.2607.

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	−15.97	0.5616	−28.43	0.0000
$\hat{\beta}_1$	16.75	0.5647	29.66	0.0000

Table 6.5: Rubella: summary of the estimated FP(m=1)-logit model for prevalence

6.5.2 Fractional polynomial of degree 2 for rubella

Now, we see a fractional polynomial of degree $m = 2$. The best second-order fractional polynomial, the one for which the first derivative of the linear predictor $\eta'_m(a, \hat{\beta}, p)$ is

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
24.00	44.22	44.93	1312.67	0.2522	0.9674

Table 6.6: Rubella: measures of goodness of fit for the estimated FP(m=1)-logit model for prevalence

nonnegative, has the logit link as link function and the best power vector for the covariate a is $[-0.9, -0.9]$. So the model $\phi_2(a; -0.9, -0.9)$ is

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.9} + \hat{\beta}_2 a_i^{-0.9}\ln(a_i) \tag{6.27}$$

and the seroprevalence function is

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.9} + \hat{\beta}_2 a_i^{-0.9}\ln(a_i))}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.9} + \hat{\beta}_2 a_i^{-0.9}\ln(a_i))} \tag{6.28}$$

Tab. 6.7 reports the results of the parameter estimation performed by R [28], while Tab. 6.8 reports some measures of goodness of fit.

	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\beta}_0$	4.340	0.16462	26.3	0.0000
$\hat{\beta}_1$	−3.444	0.16502	−20.87	0.0000
$\hat{\beta}_2$	−1.239	0.08003	−15.48	0.0000

Table 6.7: Rubella: summary of the estimated FP(m=2)-logit model for prevalence

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
23.00	25.15	25.13	1331.74	0.2559	0.9815

Table 6.8: Rubella: measures of goodness of fit for the estimated FP(m=2)-logit model for prevalence

The deviance $D(1,1)$ of the simple logistic model for rubella is 208.84. Then, the value of the gain G for the model $\phi_2(a; -0.9, -0.9)$ is

$$
\begin{aligned}
G = G(2; -0.9, -0.9) &= D(1;1) - D(2; -0.9, -0.9) \\
&= 208.84 - 25.15 \\
&= 183.69. \qquad (6.29)
\end{aligned}
$$

For this fitted model the deviance is about 25 on 23 degrees of freedom and the chi-squared goodness of fit statistics is about 25 (23 d.f.). G is 183.69, meaning that the fitted model is significantly different from the logistic model. The R^2 constructed from the Kullback-Leibler divergence is 0.98, that is about 98% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.2559, while the maximum achievable is 0.2607.

In Fig. 6.4 we have the plot of the two fractional polynomial models for rubella.

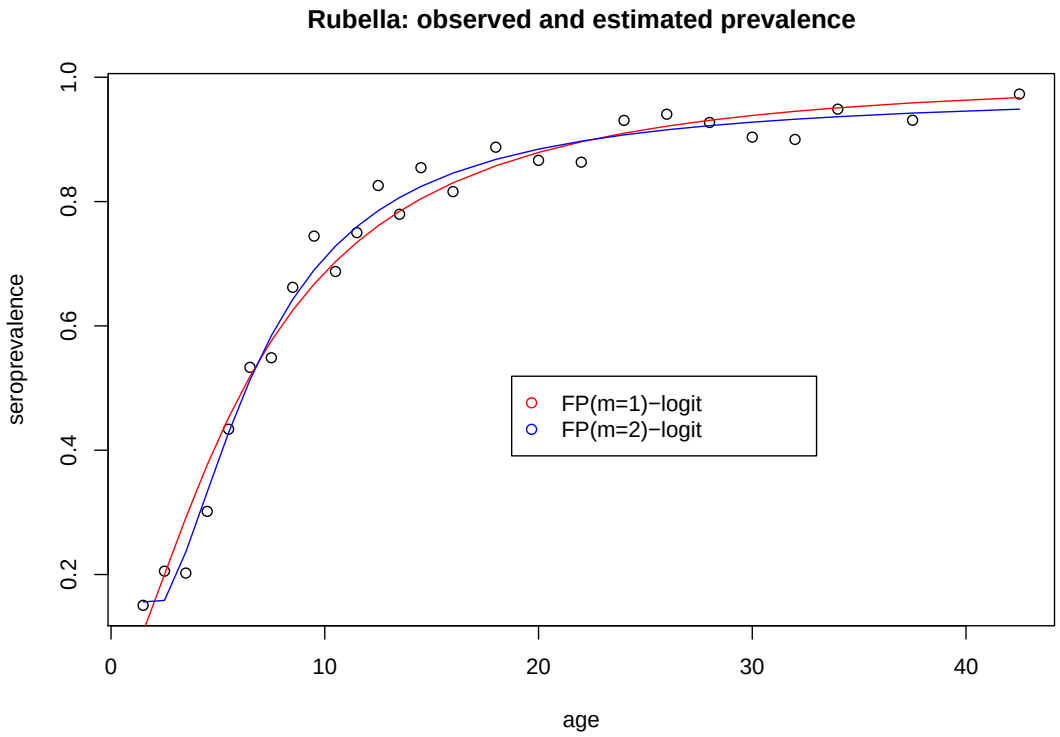


Figure 6.4: Rubella: observed and estimated prevalence with fractional polynomials of degree $m = 1$ and $m = 2$

Let us use now the criterion to choose between the model of first degree and the

model of second degree:

$$\begin{aligned} D(1;0.1) - D(2;-0.9,-0.9) &= 44.22 - 25.15 \\ &= 19.07 > 4.6052 = \chi^2_{2,0.90}. \end{aligned} \tag{6.30}$$

Given that $D(1;0.1) - D(2;-0.9,-0.9) > \chi^2_{2,0.90}$, we conclude that the best fractional polynomial model for rubella is that of degree 2.

6.5.3 Estimation of the force of infection

Now we plot the estimated force of infection for the best fractional polynomial model for prevalence, $\phi_2(a;-0.9,-0.9)$.

From the catalytic model, the function of the force of infection for the second-degree model is

$$\begin{aligned} \ell_2(a_i) &= \eta'(a_i)e^{\eta(a_i)} \\ &= -a_i^{-1.9}[0.9\hat{\beta}_1 + (1 - 0.9\ln a_i)\hat{\beta}_2]F(a_i). \end{aligned} \tag{6.31}$$

We assume that $\ell_2(0) = 0$, accounting for the protective effect of maternal antibodies in the first year of life.

Fig. 6.5 shows the plot of the estimated force of infection. It reaches a peak at age 1.5 ($\ell(1.5) = 0.1669$) and drops down thereafter until $\ell = 0.00462$ at age 42.5.

6.6 Analysis for parvovirus data

In this section, we show the results of the application of fractional polynomials to parvovirus data.

6.6.1 Fractional polynomial of degree 1 for parvovirus

In this case also, we start with a fractional polynomial of degree $m = 1$ and we implement three models in accordance to the three link functions: logit, probit and complementary log-log.

For parvovirus, the best fractional polynomial of degree 1 has the logit link as link function (differently from the GLM, where the best model is under the probit link) and the best power for the covariate is $p = -0.4$. So the model $\phi_1(a;-0.4)$ is

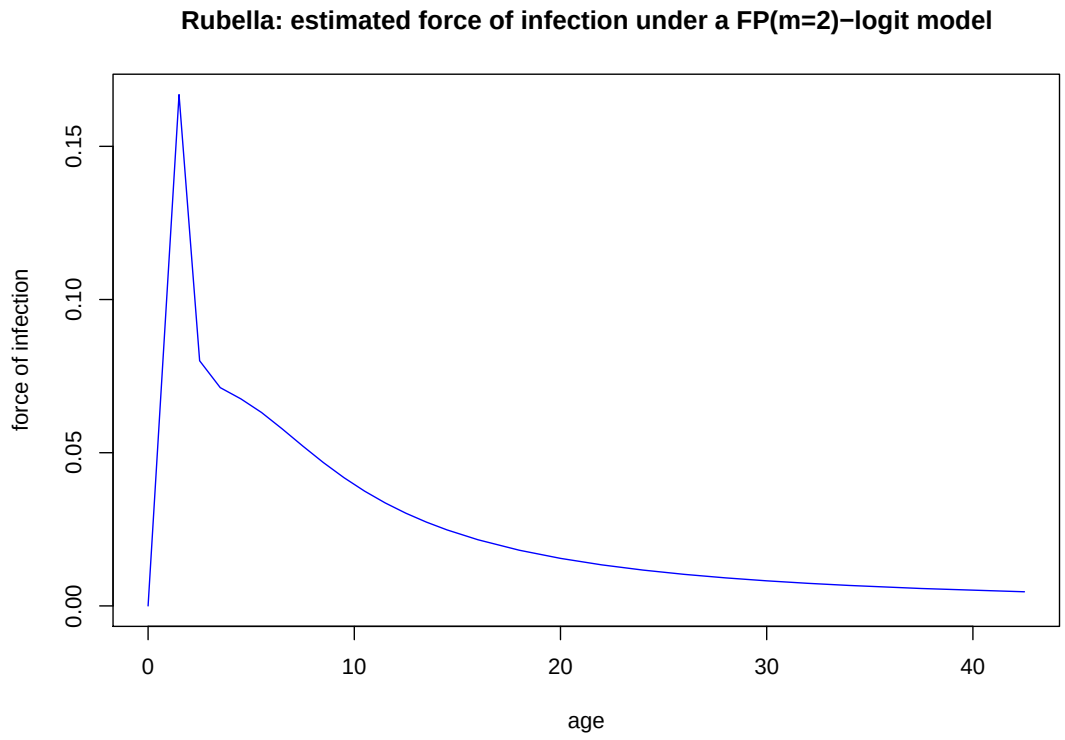


Figure 6.5: Rubella: estimated force of infection under a FP($m = 2$)-logit model for prevalence

$$\ln\left(\frac{F(a_i)}{1-F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.4} \tag{6.32}$$

and the seroprevalence function is

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.4})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-0.4})}. \tag{6.33}$$

Tab. 6.9 reports the results of the parameter estimation performed by R [28], while Tab. 6.10 reports some measures of goodness of fit.

The deviance $D(1,1)$for parvovirus under the logit link is 119.78. So, the value of the gain G for the model $\phi_1(a;-0.4)$ is

$$\begin{aligned} G = G(1;-0.4) &= D(1;1) - D(1;-0.4) \\ &= 119.78 - 50.91 \\ &= 68.87. \end{aligned} \tag{6.34}$$

For this fitted model the deviance is about 51 on 24 degrees of freedom and the chi-squared goodness of fit statistics is about 51 (24 d.f.). G is 68.87, meaning that the fitted model is significantly different from the logistic model. The R^2 constructed from the Kullback-Leibler divergence is 0.83, that is about 83% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.0572, while the maximum achievable is 0.0687.

	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\beta}_0$	1.629	0.1235	13.19	0.0000
$\hat{\beta}_1$	−1.991	0.1409	−14.13	0.0000

Table 6.9: Parvovirus: summary of the estimated FP(m=1)-logit model for prevalence

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
24.00	50.91	50.99	255.18	0.0572	0.8337

Table 6.10: Parvovirus: measures of goodness of fit for the estimated FP(m=1)-logit model for prevalence

6.6.2 Fractional polynomial of degree 2 for parvovirus

Now, we see a fractional polynomial of degree $m = 2$. The best fractional polynomial of degree 2 has the logit link as link function and the best power vector for the covariate a is $p = [-1.5, -1.4]$. So the model $\phi_2(a; -1.5, -1.4)$ is

$$\ln\left(\frac{F(a_i)}{1 - F(a_i)}\right) = \hat{\beta}_0 + \hat{\beta}_1 a_i^{-1.5} + \hat{\beta}_2 a_i^{-1.4} \tag{6.35}$$

and the seroprevalence function is

$$F(a_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-1.5} + \hat{\beta}_2 a_i^{-1.4})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i^{-1.5} + \hat{\beta}_2 a_i^{-1.4})}. \tag{6.36}$$

Tab. 6.11 reports the results of the parameter estimation performed by R [28], while Tab. 6.12 reports some measures of goodness of fit.

	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\beta}_0$	0.614	0.06423	9.560	0.0000
$\hat{\beta}_1$	3.666	0.41868	8.756	0.0000
$\hat{\beta}_2$	−4.605	0.49814	−9.244	0.0000

Table 6.11: Parvovirus: summary of the estimated FP(m=2)-logit model for prevalence

The deviance $D(1, 1)$ for parvovirus is 119.78. So, the value of the gain G for the model $\phi_2(a; -1.5, -1.4)$ is

$$\begin{aligned} G = G(2; -1.5, -1.4) &= D(1; 1) - D(2; -1.5, -1.4) \\ &= 119.78 - 40.97 \\ &= 78.81. \end{aligned} \tag{6.37}$$

For this fitted model the deviance is about 41 on 23 degrees of freedom and the chi-squared goodness of fit statistics is about 41 (23 d.f.). G is 78.81, meaning that the fitted model is significantly different from the logistic model. The R^2 constructed from the Kullback-Leibler divergence is 0.87, that is about 87% of the information provided by the full model with respect to the null model is explicated by the fitted model. The pseudo R^2 is 0.0595, while the maximum achievable is 0.0687.

Let us observe now the plot of the two fractional polynomial models in Fig. 6.6.

d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
23.00	40.97	40.98	265.12	0.0595	0.8661

Table 6.12: Parvovirus: measures of goodness of fit for the estimated FP(m=2)-cloglog model for prevalence

Let us use now the criterion to choose between the model of first degree and the model of second degree:

$$\begin{aligned} D(1;-0.4) - D(2;-1.5,-1.4) &= 50.91 - 40.97 \\ &= 9.94 > 4.6052 = \chi^2_{2,0.90}. \end{aligned} \tag{6.38}$$

Given that $D(1;-0.4) - D(2;-1.5,-1.4) > \chi^2_{2,0.90}$, we conclude that the best fractional polynomial model for parvovirus is that of degree 2.

6.6.3 Estimation of the force of infection

Now we plot the estimated force of infection for the best fractional polynomial model for prevalence, $\phi_2(a;-1.5,-1.4)$.

From the catalytic model, the function of the force of infection for the second-degree model is

$$\begin{aligned} \ell_2(a_i) &= \eta'(a_i)\frac{e^{\eta(a_i)}}{1+e^{\eta(a_i)}} \\ &= -(1.5\hat{\beta}_1 a_i^{-2.5} + 1.4\hat{\beta}_2 a_i^{-2.4})F(a_i). \end{aligned} \tag{6.39}$$

We assume that $\ell_2(0) = 0$, accounting for the protective effect of maternal antibodies in the first year of life.

Fig. 6.5 shows the plot of the estimated force of infection. It reaches a peak at age 1.5 ($\ell(1.5) = 0.0598$) and drops down thereafter until $\ell = 0.0002$ at age 42.5.

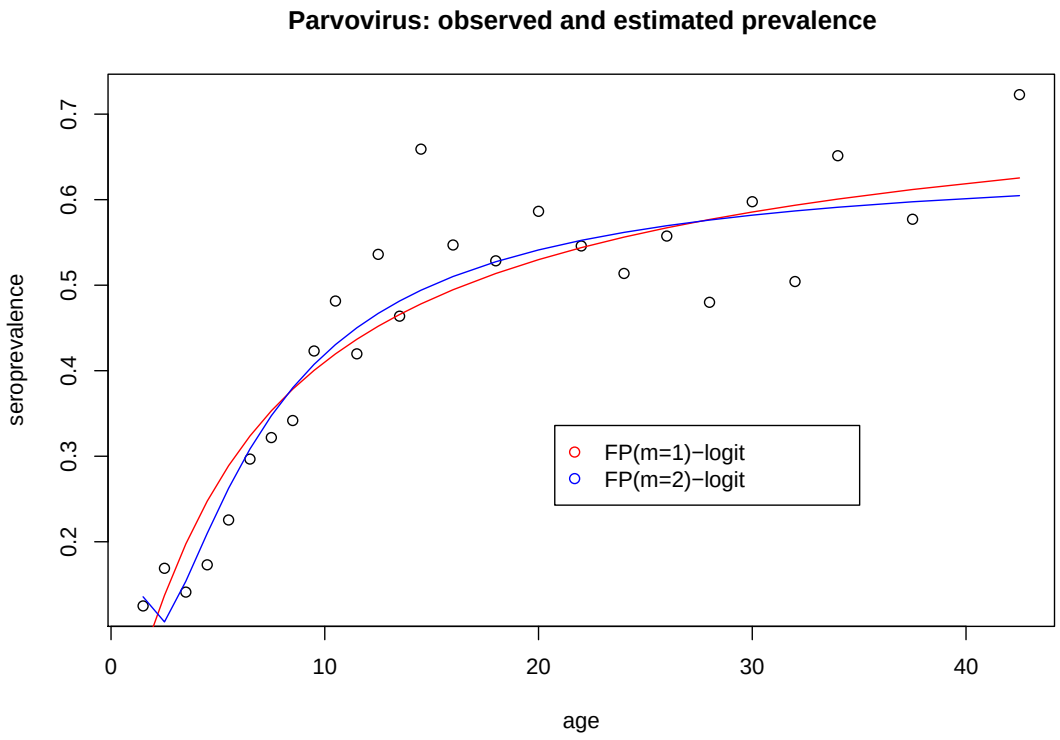


Figure 6.6: Parvovirus: observed and estimated prevalence with fractional polynomials of degree $m = 1$ and $m = 2$

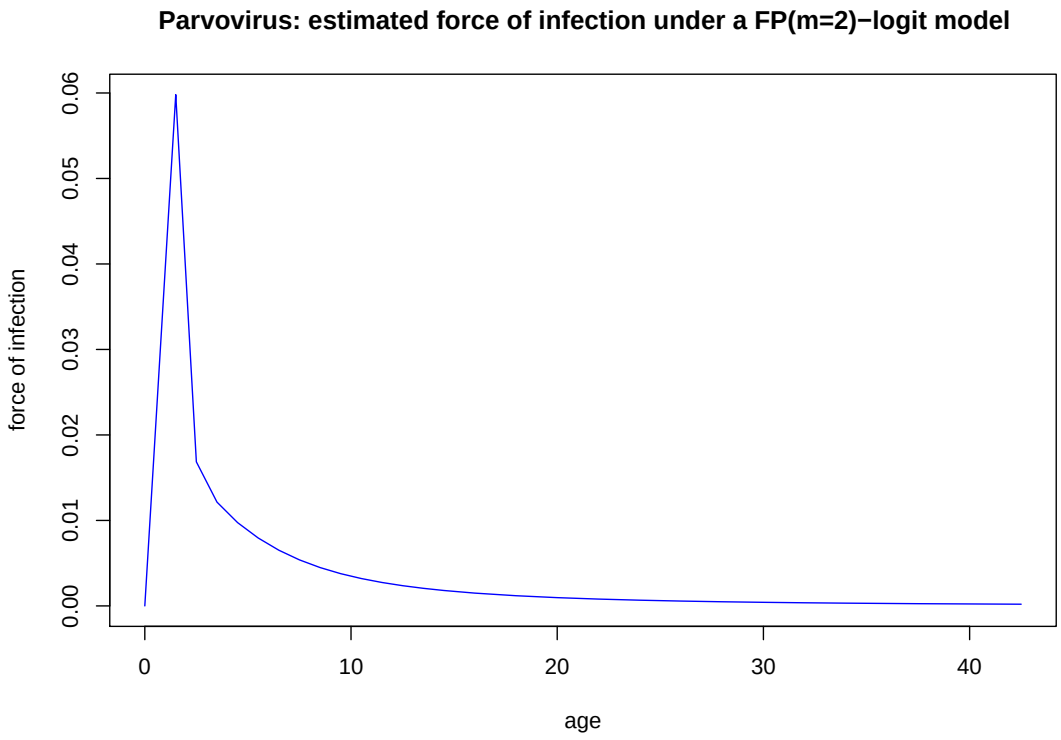


Figure 6.7: Parvovirus: estimated force of infection under a FP($m = 2$)-logit model for prevalence

Chapter 7

Conclusions

7.1 Confidence intervals for the seropositive proportions

In Chapter 2 we have started with the problem of the determination of the optimal sample size for a serological survey. Since in the serological surveys presented in literature it has been shown that the seropositive proportions vary with the age of the subject, we have posed the problem to determine the minimum sample size for every age class to estimate correctly the seroprevalence. The question is important: given that surveying units are specimens of human blood, it is not possible to gather a very large sample.

However, the classical formula to derive the sample size,

$$n = \frac{4\kappa^2 \hat{\pi}(1-\hat{\pi})}{A^2}, \tag{7.1}$$

where $\kappa = z_{\alpha/2}$ and A is the size of the standard interval confidence, has some problems:

1. the formula is valid if the event of interest is not rare, i.e. $0.1 \leq \hat{\pi} \leq 0.9$;

2. the formula is based on the inversion of the standard (or Wald) confidence interval (CI) for the binomial proportion, which, as we have shown, is not a good CI.

The relationship between these two questions is that the actual coverage probability of the standard CI is poor when $\hat{\pi} < 0.1$ or $\hat{\pi} > 0.9$.

In consequence of that, we have followed two directions:

1. Since in many cases presented in literature the seropositive proportions are lower than 0.1 or higher than 0.9, we have studied the problem of the convergence of the

binomial distribution to the Normal distribution in order to estimate correctly the seropositive proportions, whatever their value is.

2. We have shown the problems of the standard CI and then we have introduced some alternative CI presented in literature.

7.1.1 The convergence of the binomial distribution to the Normal

To study the convergence of the binomial distribution to the Normal distribution we have used the Cramer - Von Mises criterion, which is simply the euclidean distance between the cumulative distribution function of the binomial and the Normal cumulative distribution function with the same mean and variance of the binomial.

We have observed that the convergence is quickly if $\pi \geq 0.5$, otherwise it can be very slow.

7.1.2 Problems of the standard CI

We have shown that the Wald CI

$$CI_S = \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \tag{7.2}$$

has a poor coverage probability (CP) either when $\hat{\pi}$ is near 0.5 and so the event is not rare, either when n is large: indeed, the actual CP tends to be lower than the nominal CP and then it presents an oscillatory behaviour. Thus, the main problems of the Wald CI are:

1. A systematic negative bias, whatever is the value of $\hat{\pi}$ for fixed n, that is to say the actual average coverage probability is lower than the nominal CP. This bias is due to the fact the standard CI has the "wrong" center.

2. An oscillatory behaviour of the actual CP when n varies and $\hat{\pi}$ is fixed. This behaviour is due to the discreteness and the skewness of the binomial distribution.

To solve these problems of the standard CI, literature presents several alternative intervals: some are centered on a different point than that of the Wald CI; others are based on continuous distribution, as the Beta or the F distribution.

The Wilson (or score) interval and the Agresti-Coull interval are centered on a different point than the estimate X/n. This point is a weighted average between the estimate

X/n and 0.5: when n is not large, the center gets closer to $1/2$; when n is larger, then the center gets closer the estimte X/n. These intervals are given by the following formulae:

$$CI_W = \frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa\sqrt{n}}{n + \kappa^2}\sqrt{\hat{\pi}(1 - \hat{\pi}) + \frac{\kappa^2}{4n}} \tag{7.3}$$

for the Wilson interval and

$$CI_{AC} = \frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa}{\sqrt{n + \kappa^2}}\sqrt{\frac{X + \kappa^2/2}{n + \kappa^2}\left(1 - \frac{X + \kappa^2/2}{n + \kappa^2}\right)} \tag{7.4}$$

for the Agresti-Coull interval, where $\kappa = z_{\alpha/2}$.

The Jeffreys prior interval is based on the continuous Beta distribution, since the Beta is the standard conjugate prior distribution of the binomial: this means that, given a likelihood function based on the binomial distribution, the prior distribution Beta is conjugate to this likelihood function because the resulting posterior distribution is a Beta too. The interval is given by

$$CI_J = \left[B_{\alpha/2}\left(X + \frac{1}{2}, n - X + \frac{1}{2}\right), B_{1-\alpha/2}\left(X + \frac{1}{2}, n - X + \frac{1}{2}\right)\right]. \tag{7.5}$$

The Clopper-Pearson "exact" interval is based on the binomial distribution, which is discrete; however, its endpoints can be determined using the F distribution or the Beta distribution.

Tables in Appendix A report two kinds of tables:

1. tables with the estimated seropositive proportions for the three datasets and their confidence intervals in accordance to the five CI;
2. tables with the lengths of the five CI.

Tab. A.1 and Tab. A.2 report the confidence intervals for the estimated seropositive proportions for mumps data set. Tab. A.3 and Tab. A.4 report the confidence intervals for rubella data set. Tab. A.5 and Tab. A.6 report the confidence intervals for parvovirus data set.

Tab. A.7 reports the length of the five CI for mumps prevalence, while Tab. A.8 reports the length of the CI for parvovirus. We have chosen these two datasets because the first is characterized from high values of the seropositive proportions (higher than 0.9),

while the second dataset presents lower proportion (the prevalence reaches its maximum at 0.72).

In general we can observe that:

- The upper bound of the standard CI can be higher than 1, if the estimated proportion is next to 1 (see Tab. A.1 and from Tab. A.3);
- The shortest CI is the Jeffreys prior interval when $0.12 < \hat{\pi} < 0.23$ (see Tab. A.8) and when $0.77 < \hat{\pi} < 0.99$ (see Tab. A.7), that is when $\hat{\pi}$ is near the boundaries (but not when $\hat{\pi} = 0$ or 1). Instead, when $0.24 < \hat{\pi} < 0.76$ the shortest CI are the Wilson interval and the Agresti-Coull interval, although the Wilson interval is sometimes more parsimonious than the Agresti-Coull.
- The longest CI is the Clopper-Pearson "exact" interval, which is the more conservative interval: in effect this CI is characterized by its length, larger than that of the other interval, and by its actual coverage probability, which is very often over the nominal CP.

7.1.3 The optimal sample size

Finally, in order to determine the optimal sample size, we have constructed the function $A(n, \pi)$, which furnishes the value of the difference between the upper bound and the lower bound of a CI, which is the *length* of the interval. After that the researcher has chosen *a priori* a certain level of this total error A and the probability of success π, he can use the plotted curves of the function $A(n, \pi)$ to determine the optimal sample size n. Of course these curves can be constructed for everyone of the alternative confidence intervals introduced in Chapter 2.

In Fig. 7.1 we plot the function $A(n, \pi)$ for the five CI, comparing what happens when the probability of success is $\pi = 0.1$ or $\pi = 0.9$. In this case, we can see that the total error size A is larger if we use the Clopper-Pearson interval, while A is smaller when we use the Jeffreys prior interval or the standard interval.

In Fig. 7.2 we plot the function $A(n, \pi)$ for the five CI, comparing what happens when the probability of success is $\pi = 0.2$ or $\pi = 0.8$. In this case, the total error size A is larger than the case $\pi = 0.1$ or $\pi = 0.9$. Now the larger values of A are furnished by the Clopper-Pearson interval again, while the lowest values are furnished by the Wilson interval and the Jeffreys prior interval.

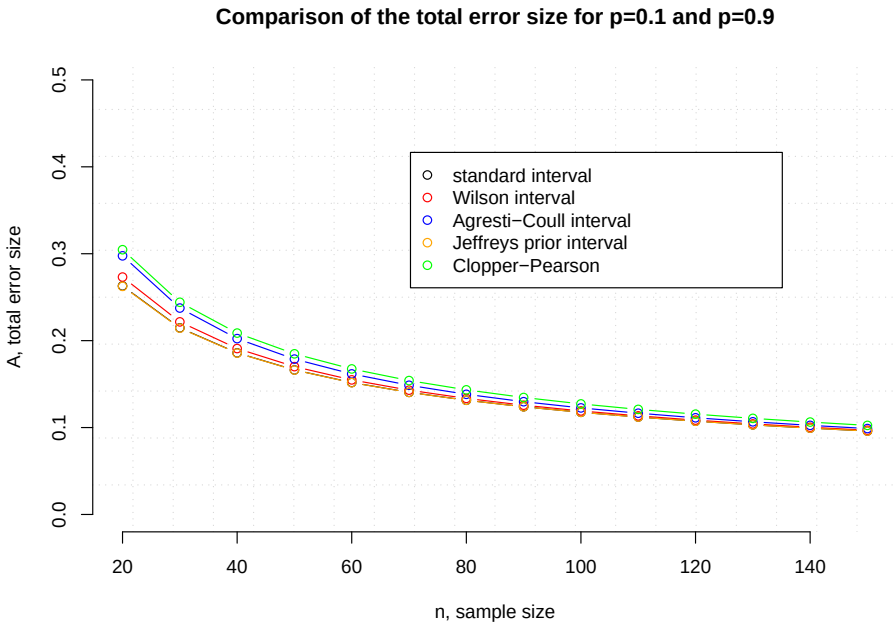


Figure 7.1: Comparison of the function $A(n, \pi)$ between the five CI for $\pi = 0.1$ or $\pi = 0.9$

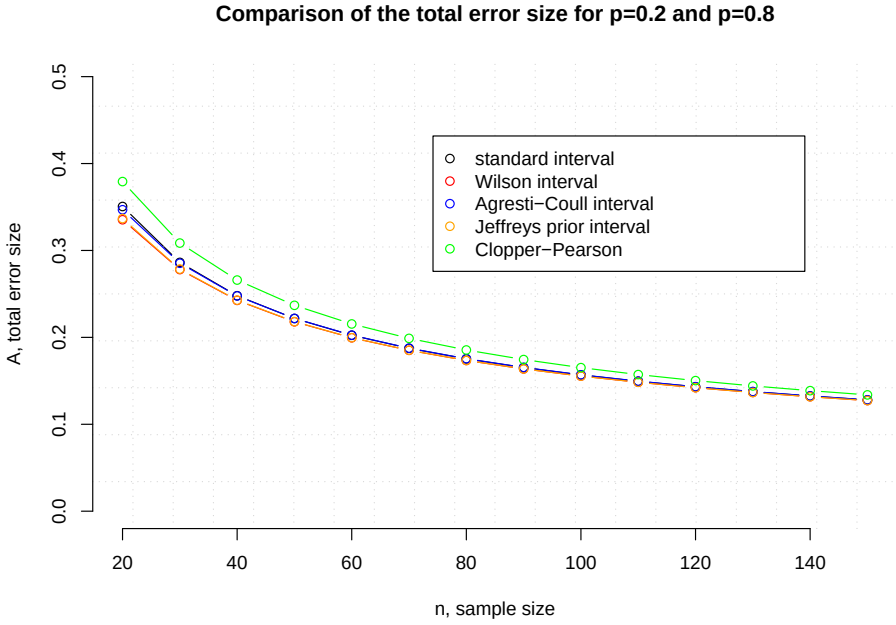


Figure 7.2: Comparison of the function $A(n, \pi)$ between the five CI for $\pi = 0.2$ or $\pi = 0.8$

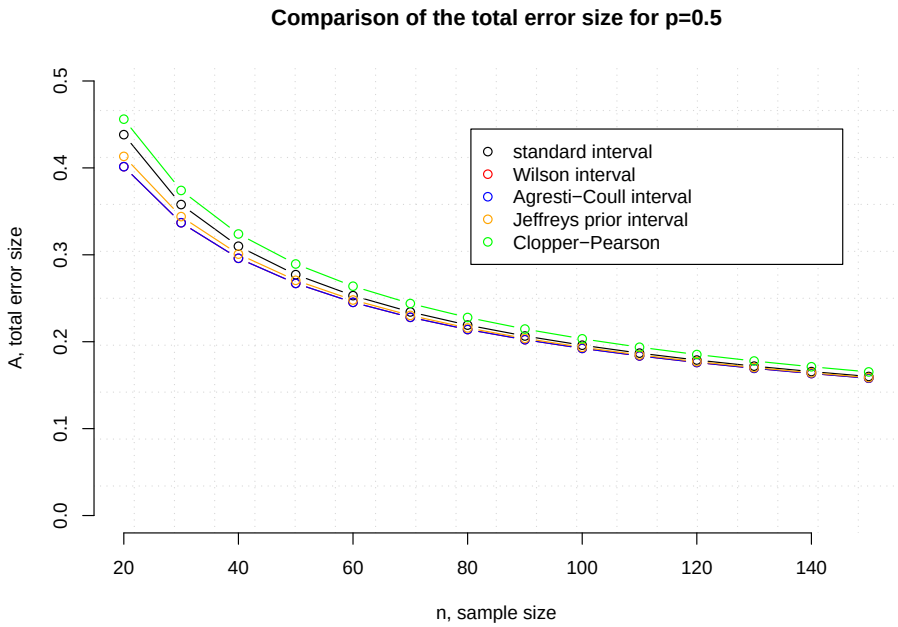


Figure 7.3: Comparison of the function $A(n, \pi)$ between the five CI for $\pi = 0.5$

Eventually, Fig. 7.3 shows the plots of $A(n, \pi)$ when $\pi = 0.5$. This is the case of maximum variance. So, the values of A are generally higher than the previous cases. We have the larger values of A for the Clopper-Pearson interval and the lowest values for the Wilson interval and the Agresti-Coull interval.

In conclusion, what we can say about these five intervals is:

1. The standard interval has a very poor coverage probability, lower that the nominal CP, either when n is large either when $\pi \approx 0.5$.

2. The Wilson interval has an actual coverage probability (except when $\pi = 0$ or $\pi = 1$) at the level of the nominal CP, even when n is not too large; its length is the shortest when when $0.2 < \hat{\pi} < 0.8$.

3. The Agresti-Coull interval is simple to comprehend and to evaluate, has a good coverage probability, similar to that of the Wilson interval (except for $\pi = 0$ or $\pi = 1$, when its CP is higher than the nominal CP) and its length is larger or equal to that of the Wilson interval.

4. The Jeffreys prior interval has a good coverage probability, similar to that of the Wilson interval (with the same problems at the extremes of the range of π); its length is the shortest when $\pi < 0.2$ and $\pi > 0.8$.

5. The Clopper-Pearson interval is the more conservative interval, because its coverage probability is always higher than the nominal CP, particularly when π is near the extrems of its range; thus, its length is always the largest, whatever the probability of success π is.

7.2 Comparison of the goodness-of-fit measures for the fitted prevalence models

In the previous chpaters, we have tried to describe the behaviour of the seropositive proportions for three infections (mumps, rubella and parvovirus) using some parametric models: in Chapter 4 we have fitted the non-linear least squares model of Farrington [1], in Chapter 5 we have used the generalized linear models and in Chapter 6 we have used the fractional polynomial models.

In this chapter we will compare the four fitted models for every data set, that are the Farrington's nonlinear least squares model, the GLM model, the first-order FP and the second-order FP models. Of course, the aim of this comparison is to see which is the best parametric model in terms of adequacy to our data.

7.2.1 Mumps: Parametric models for prevalence

	d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
FP(m=2)-logit	23	27.90	31.11	2895.64	0.3661	0.9904
Nonlinear L. S.	23	46.48	49.89	2877.06	0.3637	0.9841
FP(m=1)-logit	24	65.40	74.89	2858.14	0.3613	0.9776
GLM-logit	24	581.37	1755.63	2342.17	0.2961	0.8011

Table 7.1: Mumps: comparing goodness-of-fit measures for the fitted models

For mumps we have fitted the following models:

1. Farrington's nonlinear least squares model;

2. a generalized linear model under the logit link;

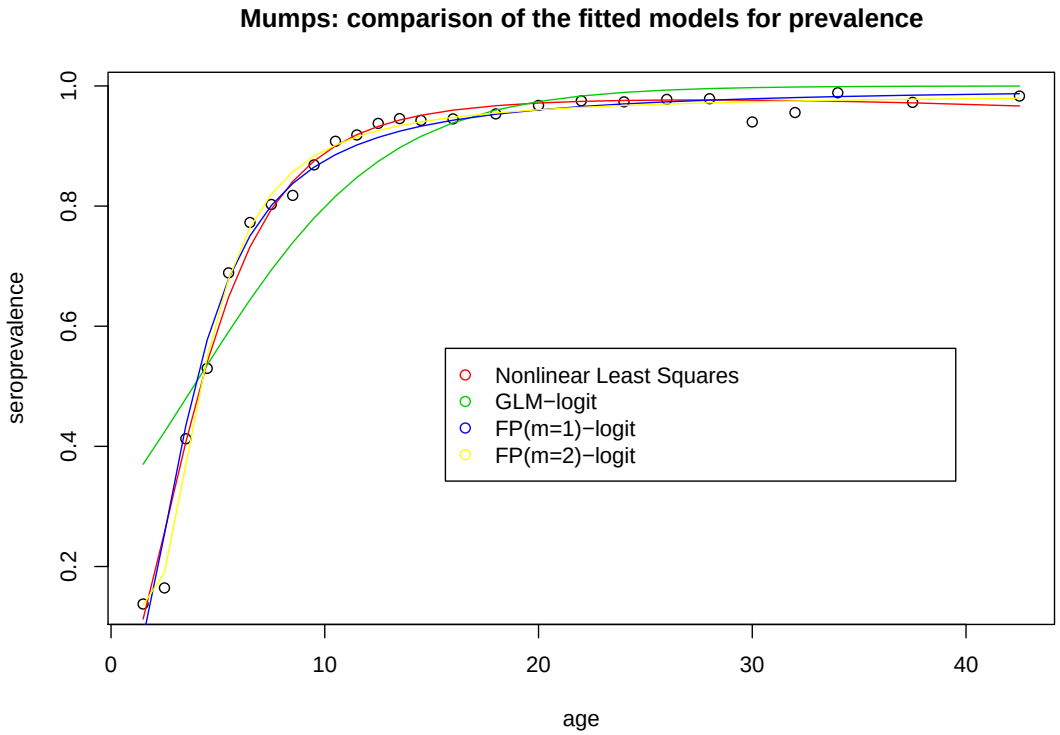


Figure 7.4: Comparison of the plots for the four fitted models for mumps prevalence

3. a first-order fractional polynomial model under the logit link;

4. a second-order fractional polynomial model under the logit link.

Tab. 7.1, reporting the goodness-of-fit measures for the fitted models, is sorted in ascending order by the deviance D. We can observe that the best fitting model is the second-order FP, whose deviance is 27.90 on 23 d.f. and whose R^2 from the Kullback-Leibler divergence is about 0.99. We remember that for mumps the maximum value of the pseudo R^2 is 0.3696.

After the fractional polynomial of degree 2, we have the nonlinear least squares model proposed by Farrington, whose deviance is twice that of FP(m=2)-logit model. Then we have the first-order FP and the GLM-logit model, whose deviance is 20 times larger than the deviance of FP(m=2)-logit model. From Fig. 7.4 also we can see that the GLM fits badly the seropositive proportions, particularly in age range [5.5,14.5], while the other three models are very similar to each other.

7.2.2 Rubella: parametric models for prevalence

	d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
FP(m=2)-logit	23	25.15	25.13	1331.74	0.2559	0.9815
FP(m=1)-logit	24	44.22	44.93	1312.67	0.2522	0.9674
Nonlinear L. S.	23	47.40	54.18	1309.49	0.2516	0.9651
GLM-logit	24	208.84	249.59	1148.05	0.2206	0.8461

Table 7.2: Rubella: comparing goodness-of-fit measures for the fitted models

For rubella we have fitted the following models:

1. Farrington's nonlinear least squares model;

2. a generalized linear model under the logit link;

3. a first-order fractional polynomial model under the logit link;

4. a second-order fractional polynomial model under the logit link.

Tab. 7.2, reporting the goodness-of-fit measures for the fitted models, is sorted in ascending order by the deviance D. We can observe that the best fitting model for rubella is the second-order FP again, whose deviance is 25.15 on 23 d.f. and whose R^2 from the

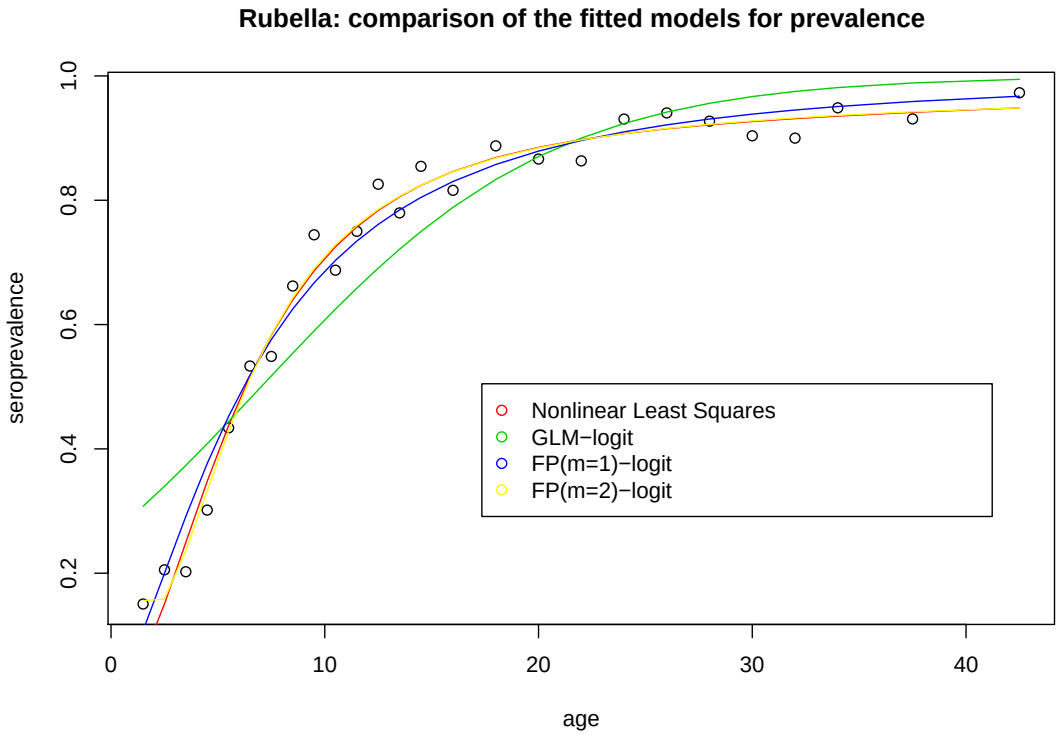


Figure 7.5: Comparison of the plots for the four fitted models for rubella prevalence

Kullback-Leibler divergence is about 0.98. We remember that for rubella the maximum value of the pseudo R^2 is 0.2607.

After the fractional polynomial of degree 2, this time we have the fractional polynomial of degree 1. It is followed by the nonlinear least squares model proposed by Farrington, whose deviance is very next to that of the preceding model. The last model is always the GLM-logit model, whose deviance is about 9 times larger than the deviance of FP(m=2)-logit model. From Fig. 7.5 also we can see that the GLM fits badly the seropositive proportions, particularly in age range [5.5,17], while the other three models are very similar to each other.

7.2.3 Parvovirus: parametric models for prevalence

	d.f.	D	X^2	C	Pseudo R^2	R^2_{KL}
FP(m=2)-logit	23	40.97	40.98	265.12	0.0595	0.8661
Nonlinear L. S.	23	49.34	54.27	256.75	0.0576	0.8388
FP(m=1)-logit	24	50.91	50.99	255.18	0.0572	0.8337
GLM-probit	24	118.97	115.17	187.12	0.0420	0.6113

Table 7.3: Parvovirus: comparing goodness-of-fit measures for the fitted models

For parvovirus we have fitted the following models:

1. Farrington's nonlinear least squares model;
2. a generalized linear model under the probit link;
3. a first-order fractional polynomial model under the logit link;
4. a second-order fractional polynomial model under the logit link.

Tab. 7.3, reporting the goodness-of-fit measures for the fitted models, is sorted in ascending order by the deviance D. We can observe that the best fitting model for parvovirus is the second-order FP again, whose deviance is 40.97 on 23 d.f. and whose R^2 from the Kullback-Leibler divergence is about 0.87. We remember that for parvovirus the maximum value of the pseudo R^2 is 0.0687.

After the fractional polynomial of degree 2, we have Farrington's model, as it happes for mumps. The third model is the first-order FP, whose deviance is very next to that of Farrington's model. The last model is always the GLM model, but this time its deviance is only about 3 times larger than the deviance of FP(m=2) model. From Fig. 7.6 we

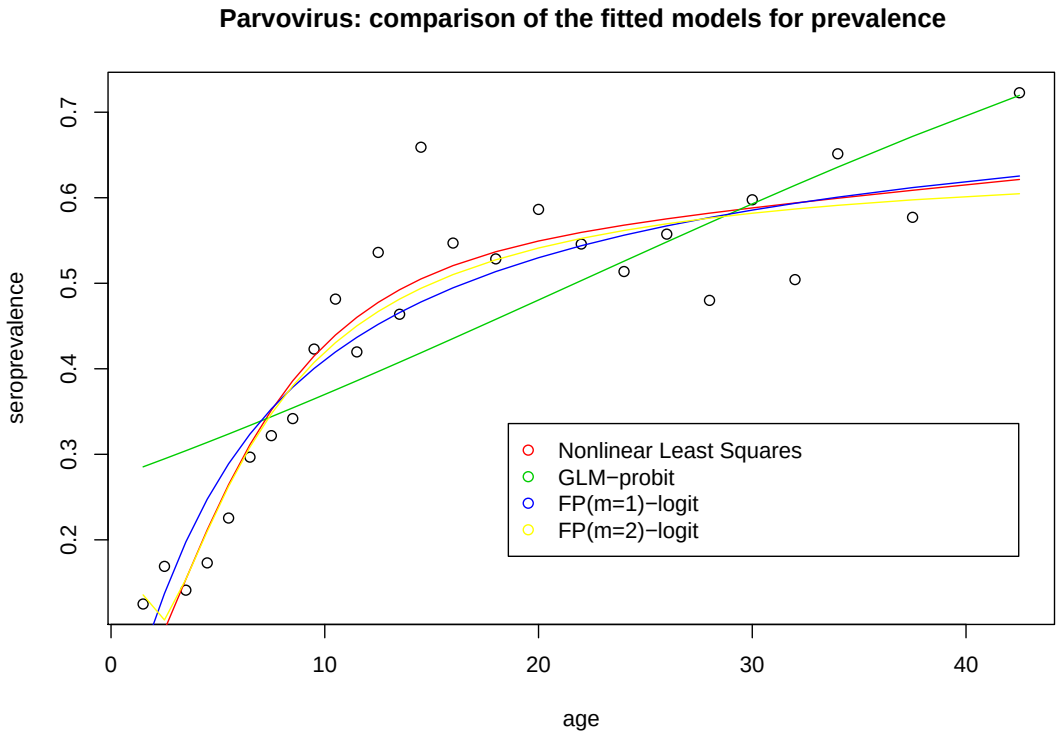


Figure 7.6: Comparison of the plots for the four fitted models for parvovirus prevalence

can see that these four models do not fit very well the observed seropositive proportions. While the GLM-probit model reduces the plot to a straight line, the other three models, although they are certainly nonlinear models, however they are not able to describe the behaviour of data after $a = 20$, i.e. a slight decrease followed by a more noticeable increase.

Appendix A

Age	X	n	mean	L_S	U_S	L_W	U_W	L_{AC}	U_{AC}
1.5	56	407	0.1376	0.1041	0.1711	0.1075	0.1745	0.1073	0.1746
2.5	48	292	0.1644	0.1219	0.2069	0.1263	0.2112	0.1261	0.2114
3.5	137	332	0.4127	0.3597	0.4656	0.361	0.4663	0.361	0.4663
4.5	195	368	0.5299	0.4789	0.5809	0.4789	0.5803	0.4789	0.5803
5.5	290	421	0.6888	0.6446	0.7331	0.6431	0.7312	0.643	0.7312
6.5	255	330	0.7727	0.7275	0.8179	0.7245	0.8147	0.7244	0.8148
7.5	236	294	0.8027	0.7572	0.8482	0.7535	0.8442	0.7533	0.8443
8.5	211	258	0.8178	0.7707	0.8649	0.7662	0.8602	0.766	0.8604
9.5	271	312	0.8686	0.8311	0.9061	0.8266	0.9016	0.8263	0.9019
10.5	276	304	0.9079	0.8754	0.9404	0.8701	0.9355	0.8697	0.9359
11.5	259	282	0.9184	0.8865	0.9504	0.8806	0.945	0.8801	0.9455
12.5	301	321	0.9377	0.9113	0.9641	0.9057	0.9593	0.9052	0.9598
13.5	296	313	0.9457	0.9206	0.9708	0.9148	0.9658	0.9142	0.9664
14.5	345	366	0.9426	0.9188	0.9664	0.9139	0.9622	0.9135	0.9626
16	224	237	0.9451	0.9162	0.9741	0.9084	0.9677	0.9076	0.9685
18	328	344	0.9535	0.9312	0.9757	0.9258	0.9712	0.9252	0.9717
20	358	370	0.9676	0.9495	0.9856	0.9442	0.9814	0.9436	0.982
22	355	364	0.9753	0.9593	0.9912	0.9537	0.9869	0.953	0.9877
24	331	340	0.9735	0.9565	0.9906	0.9505	0.986	0.9497	0.9868
26	350	358	0.9777	0.9623	0.993	0.9565	0.9886	0.9558	0.9894
28	322	329	0.9787	0.9631	0.9943	0.9567	0.9897	0.9558	0.9905
30	251	267	0.9401	0.9116	0.9685	0.9049	0.9628	0.9042	0.9634
32	216	226	0.9558	0.9289	0.9826	0.9205	0.9758	0.9195	0.9768
34	175	177	0.9887	0.9731	1.004	0.9597	0.9969	0.9571	0.9995
37.5	320	329	0.9726	0.955	0.9903	0.9488	0.9855	0.9481	0.9863
42.5	234	238	0.9832	0.9669	0.9995	0.9576	0.9934	0.956	0.995

Table A.1: Confidence intervals for the estimated seropositive proportions for mumps: the standard interval, the Wilson interval and the Agresti-Coull interval

Age	X	n	mean	L_J	U_J	L_{CP}	U_{CP}
1.5	56	407	0.1385	0.1067	0.1736	0.1056	0.1749
2.5	48	292	0.1655	0.1253	0.2101	0.1238	0.212
3.5	137	332	0.4129	0.3606	0.4662	0.3592	0.4677
4.5	195	368	0.5298	0.4788	0.5805	0.4775	0.5818
5.5	290	421	0.6884	0.6434	0.7317	0.6422	0.7328
6.5	255	330	0.7719	0.7253	0.8154	0.7237	0.8168
7.5	236	294	0.8017	0.7544	0.8451	0.7526	0.8467
8.5	211	258	0.8166	0.7673	0.8612	0.7652	0.863
9.5	271	312	0.8674	0.8278	0.9026	0.826	0.904
10.5	276	304	0.9066	0.8715	0.9366	0.8696	0.9379
11.5	259	282	0.917	0.8822	0.9462	0.8801	0.9476
12.5	301	321	0.9363	0.9073	0.9603	0.9054	0.9615
13.5	296	313	0.9443	0.9164	0.9668	0.9145	0.968
14.5	345	366	0.9414	0.9152	0.963	0.9136	0.9641
16	224	237	0.9433	0.9106	0.9689	0.908	0.9705
18	328	344	0.9522	0.9273	0.9721	0.9256	0.9732
20	358	370	0.9663	0.9457	0.9822	0.944	0.9831
22	355	364	0.974	0.9554	0.9877	0.9536	0.9886
24	331	340	0.9721	0.9522	0.9868	0.9503	0.9878
26	350	358	0.9763	0.9583	0.9894	0.9564	0.9903
28	322	329	0.9773	0.9587	0.9904	0.9567	0.9914
30	251	267	0.9384	0.9068	0.9639	0.9045	0.9654
32	216	226	0.9537	0.9229	0.977	0.9201	0.9786
34	175	177	0.986	0.9642	0.9976	0.9598	0.9986
37.5	320	329	0.9712	0.9507	0.9864	0.9487	0.9874
42.5	234	238	0.9812	0.9605	0.9943	0.9575	0.9954

Table A.2: Confidence intervals for the estimated seropositive proportions for mumps: the Jeffreys prior interval and the Clopper-Pearson "exact" interval

Age	X	n	mean	L_S	U_S	L_W	U_W	L_{AC}	U_{AC}
1.5	31	206	0.1505	0.1017	0.1993	0.1081	0.2057	0.1077	0.2061
2.5	30	146	0.2055	0.1399	0.271	0.1479	0.2782	0.1475	0.2786
3.5	34	168	0.2024	0.1416	0.2631	0.1486	0.2695	0.1482	0.2698
4.5	57	189	0.3016	0.2362	0.367	0.2406	0.3704	0.2405	0.3706
5.5	95	219	0.4338	0.3682	0.4994	0.3699	0.5	0.3698	0.5
6.5	104	195	0.5333	0.4633	0.6034	0.4633	0.602	0.4633	0.602
7.5	90	164	0.5488	0.4726	0.6249	0.4724	0.623	0.4724	0.623
8.5	96	145	0.6621	0.5851	0.7391	0.5818	0.734	0.5817	0.7341
9.5	134	180	0.7444	0.6807	0.8082	0.6761	0.8026	0.6759	0.8028
10.5	110	160	0.6875	0.6157	0.7593	0.612	0.7542	0.6119	0.7543
11.5	111	148	0.75	0.6802	0.8198	0.6745	0.8128	0.6742	0.8131
12.5	147	178	0.8258	0.7701	0.8816	0.7634	0.8745	0.763	0.8749
13.5	138	177	0.7797	0.7186	0.8407	0.713	0.8344	0.7127	0.8347
14.5	141	165	0.8545	0.8008	0.9083	0.7927	0.9003	0.7921	0.9009
16	102	125	0.816	0.7481	0.8839	0.739	0.8741	0.7384	0.8748
18	142	160	0.8875	0.8385	0.9365	0.8292	0.9276	0.8284	0.9285
20	162	187	0.8663	0.8175	0.9151	0.8101	0.9078	0.8096	0.9083
22	158	183	0.8634	0.8136	0.9131	0.8061	0.9057	0.8056	0.9063
24	161	173	0.9306	0.8928	0.9685	0.8827	0.9599	0.8816	0.961
26	174	185	0.9405	0.9065	0.9746	0.8967	0.9665	0.8956	0.9676
28	153	165	0.9273	0.8876	0.9669	0.8772	0.9579	0.8761	0.959
30	122	135	0.9037	0.8539	0.9535	0.8422	0.9429	0.841	0.944
32	90	100	0.9	0.8412	0.9588	0.8256	0.9448	0.8239	0.9465
34	74	78	0.9487	0.8998	0.9977	0.8754	0.9799	0.8715	0.9838
37.5	175	188	0.9309	0.8946	0.9671	0.8853	0.9591	0.8843	0.9601
42.5	108	111	0.973	0.9428	1.003	0.9235	0.9908	0.9201	0.9942

Table A.3: Confidence intervals for the estimated seropositive proportions for rubella: standard, Wilson and Agresti-Coull

Age	X	n	mean	L_J	U_J	L_{CP}	U_{CP}
1.5	31	206	0.1522	0.1067	0.204	0.1046	0.2068
2.5	30	146	0.2075	0.1461	0.2764	0.1431	0.2802
3.5	34	168	0.2041	0.147	0.2679	0.1444	0.2712
4.5	57	189	0.3026	0.2395	0.3697	0.2371	0.3724
6.5	104	195	0.5332	0.4633	0.6024	0.4607	0.6049
7.5	90	164	0.5485	0.4723	0.6235	0.4693	0.6265
8.5	96	145	0.661	0.5825	0.7353	0.5789	0.7385
9.5	134	180	0.7431	0.6772	0.8039	0.6742	0.8064
10.5	110	160	0.6863	0.6128	0.7555	0.6096	0.7583
11.5	111	148	0.7483	0.6759	0.8145	0.6722	0.8175
12.5	147	178	0.824	0.7651	0.8761	0.762	0.8785
13.5	138	177	0.7781	0.7144	0.8359	0.7113	0.8384
14.5	141	165	0.8524	0.7948	0.902	0.7913	0.9045
16	102	125	0.8135	0.7413	0.8763	0.7368	0.8796
18	142	160	0.8851	0.8317	0.9295	0.828	0.9319
20	162	187	0.8644	0.812	0.9094	0.809	0.9116
22	158	183	0.8614	0.8081	0.9073	0.805	0.9096
24	161	173	0.9282	0.8855	0.9615	0.882	0.9636
26	174	185	0.9382	0.8995	0.968	0.8961	0.9699
28	153	165	0.9247	0.8801	0.9596	0.8764	0.9619
30	122	135	0.9007	0.8454	0.9449	0.841	0.9477
32	90	100	0.896	0.8299	0.9474	0.8238	0.951
34	74	78	0.943	0.8827	0.9824	0.8739	0.9859
37.5	175	188	0.9286	0.8879	0.9607	0.8847	0.9627
42.5	108	111	0.9688	0.9296	0.9923	0.923	0.9944

Table A.4: Confidence intervals for the estimated seropositive proportions for rubella: the Jeffreys prior interval and the Clopper-Pearson "exact" interval

Age	X	n	mean	L_S	U_S	L_W	U_W	L_{AC}	U_{AC}
1.5	9	72	0.125	0.04861	0.2014	0.06718	0.2208	0.06498	0.223
2.5	12	71	0.169	0.08184	0.2562	0.09941	0.2726	0.09785	0.2742
3.5	12	85	0.1412	0.06715	0.2152	0.08264	0.2307	0.0811	0.2323
4.5	18	104	0.1731	0.1004	0.2458	0.1124	0.2571	0.1115	0.258
5.5	23	102	0.2255	0.1444	0.3066	0.1552	0.3157	0.1546	0.3163
6.5	27	91	0.2967	0.2028	0.3906	0.2126	0.3972	0.2123	0.3976
7.5	28	87	0.3218	0.2237	0.42	0.233	0.4257	0.2327	0.426
8.5	27	79	0.3418	0.2372	0.4464	0.2467	0.4515	0.2465	0.4518
9.5	44	104	0.4231	0.3281	0.518	0.3325	0.5191	0.3325	0.5191
10.5	39	81	0.4815	0.3727	0.5903	0.376	0.5886	0.376	0.5886
11.5	34	81	0.4198	0.3123	0.5272	0.3183	0.5285	0.3182	0.5285
12.5	52	97	0.5361	0.4368	0.6353	0.4374	0.6321	0.4374	0.6321
13.5	32	69	0.4638	0.3461	0.5814	0.3511	0.5802	0.3511	0.5802
14.5	58	88	0.6591	0.5601	0.7581	0.5553	0.7496	0.555	0.7498
16	64	117	0.547	0.4568	0.6372	0.4567	0.6343	0.4567	0.6343
18	102	193	0.5285	0.4581	0.5989	0.4582	0.5977	0.4582	0.5977
20	112	191	0.5864	0.5165	0.6562	0.5155	0.6539	0.5155	0.6539
22	113	207	0.5459	0.4781	0.6137	0.4778	0.6123	0.4778	0.6123
24	131	255	0.5137	0.4524	0.5751	0.4526	0.5744	0.4526	0.5744
26	97	174	0.5575	0.4837	0.6313	0.4832	0.6292	0.4832	0.6292
28	84	175	0.48	0.406	0.554	0.4072	0.5537	0.4072	0.5537
30	101	169	0.5976	0.5237	0.6716	0.5223	0.6686	0.5223	0.6686
32	59	117	0.5043	0.4137	0.5949	0.415	0.5933	0.415	0.5933
34	71	109	0.6514	0.5619	0.7408	0.5581	0.7343	0.558	0.7344
37.5	116	201	0.5771	0.5088	0.6454	0.508	0.6433	0.508	0.6434
42.5	73	101	0.7228	0.6355	0.8101	0.6285	0.8007	0.6282	0.8011

Table A.5: Confidence intervals for the estimated seropositive proportions for parvovirus: standard, Wilson and Agresti-Coull

Age	X	n	mean	L_J	U_J	L_{CP}	U_{CP}
1.5	9	72	0.1301	0.06369	0.2157	0.05878	0.2241
2.5	12	71	0.1736	0.09585	0.2685	0.0905	0.2766
3.5	12	85	0.1453	0.07956	0.2267	0.07513	0.2336
4.5	18	104	0.1762	0.1098	0.2542	0.1059	0.2597
5.5	23	102	0.2282	0.1528	0.3135	0.1486	0.3189
6.5	27	91	0.2989	0.2103	0.3958	0.2055	0.4016
7.5	28	87	0.3239	0.2307	0.4246	0.2256	0.4306
8.5	27	79	0.3438	0.2443	0.4506	0.2387	0.4571
9.5	44	104	0.4238	0.3313	0.5191	0.3268	0.5239
10.5	39	81	0.4817	0.3749	0.5893	0.369	0.5953
11.5	34	81	0.4207	0.3166	0.5285	0.3109	0.5346
12.5	52	97	0.5357	0.437	0.633	0.4319	0.638
13.5	32	69	0.4643	0.3496	0.5809	0.3428	0.588
14.5	58	88	0.6573	0.5562	0.7517	0.5503	0.7568
16	64	117	0.5466	0.4566	0.6351	0.4523	0.6392
18	102	193	0.5284	0.4581	0.5981	0.4555	0.6006
20	112	191	0.5859	0.5157	0.6545	0.513	0.657
22	113	207	0.5457	0.4778	0.6127	0.4754	0.615
24	131	255	0.5137	0.4525	0.5746	0.4506	0.5766
26	97	174	0.5571	0.4832	0.6298	0.4804	0.6326
28	84	175	0.4801	0.4068	0.5538	0.404	0.5567
30	101	169	0.5971	0.5226	0.6694	0.5196	0.6722
32	59	117	0.5042	0.4145	0.5938	0.4103	0.598
34	71	109	0.65	0.5589	0.7359	0.5542	0.7401
37.5	116	201	0.5767	0.5081	0.6439	0.5056	0.6463
42.5	73	101	0.7206	0.6301	0.8029	0.6248	0.8072

Table A.6: Confidence intervals for the estimated seropositive proportions for parvovirus: the Jeffreys prior interval and the Clopper-Pearson "exact" interval

Age	mean	Standard	Wilson	Agresti-Coull	Jeffreys	Clopper-Pearson
1.5	0.1376	0.06693	0.06696	0.0673	0.06669*	0.06929**
2.5	0.1644	0.08502	0.08491	0.08535	0.08461*	0.08821**
3.5	0.4127	0.1059	0.1053*	0.1053*	0.1055	0.1085**
4.5	0.5299	0.102	0.1015*	0.1015*	0.1017	0.1043**
5.5	0.6888	0.08845	0.08811*	0.08818	0.08819	0.09058**
6.5	0.7727	0.09043	0.09013	0.09034	0.09007*	0.09318**
7.5	0.8027	0.09098	0.09072	0.09106	0.09056*	0.09409**
8.5	0.8178	0.0942	0.09397	0.09442	0.0937*	0.09775**
9.5	0.8686	0.07498	0.07506	0.07558	0.07462*	0.07805**
10.5	0.9079	0.06501	0.0654	0.06618	0.06466*	0.06829**
11.5	0.9184	0.06389	0.06445	0.06541	0.06349*	0.06746**
12.5	0.9377	0.05288	0.05358	0.05456	0.05256*	0.05612**
13.5	0.9457	0.05021	0.05107	0.05218	0.04988*	0.05358**
14.5	0.9426	0.04765	0.04829	0.04915	0.04739*	0.05051**
16	0.9451	0.05798	0.05924	0.06089	0.05747*	0.06245**
18	0.9535	0.04451	0.04538	0.04646	0.04421*	0.04762**
20	0.9676	0.0361	0.03718	0.03839	0.03583*	0.0391**
22	0.9753	0.03191	0.03325	0.03469	0.03162*	0.03505**
24	0.9735	0.03413	0.03555	0.03707	0.03381*	0.03748**
26	0.9777	0.03062	0.0321	0.03365	0.03032*	0.03386**
28	0.9787	0.03119	0.03292	0.0347	0.03085*	0.03475**
30	0.9401	0.05694	0.0579	0.05921	0.05651*	0.06086**
32	0.9558	0.05362	0.05531	0.05734	0.05308*	0.05845**
34	0.9887	0.03114	0.03715	0.04245**	0.03041*	0.03885
37.5	0.9726	0.03525	0.03671	0.03828	0.03492*	0.03871**
42.5	0.9832	0.03266	0.03585	0.03895**	0.03212*	0.03787

Table A.7: Length of the confidence intervals for mumps prevalence

Age	mean	Standard	Wilson	Agresti-Coull	Jeffreys	Clopper-Pearson
1.5	0.125	0.1528	0.1536	0.158	0.1497*	0.1653**
2.5	0.169	0.1743	0.1732	0.1763	0.171*	0.1861**
3.5	0.1412	0.148	0.1481	0.1512	0.1456*	0.1585**
4.5	0.1731	0.1454	0.1447	0.1465	0.1435*	0.1537**
5.5	0.2255	0.1622	0.1605	0.1617	0.1601*	0.1703**
6.5	0.2967	0.1877	0.1846*	0.1853	0.1852	0.1961**
7.5	0.3218	0.1963	0.1927*	0.1933	0.1936	0.205**
8.5	0.3418	0.2092	0.2048*	0.2053	0.206	0.2184**
9.5	0.4231	0.1899	0.1866*	0.1866*	0.1877	0.1971**
10.5	0.4815	0.2176	0.2126*	0.2127	0.2144	0.2263**
11.5	0.4198	0.215	0.2102*	0.2103	0.2118	0.2237**
12.5	0.5361	0.1985	0.1947*	0.1947*	0.196	0.206**
13.5	0.4638	0.2353	0.2291*	0.2291*	0.2313	0.2452**
14.5	0.6591	0.1981	0.1943*	0.1948	0.1953	0.2065**
16	0.547	0.1804	0.1775*	0.1776	0.1785	0.1869**
18	0.5285	0.1409	0.1395*	0.1395*	0.14	0.1451**
20	0.5864	0.1397	0.1383*	0.1384	0.1388	0.144**
22	0.5459	0.1357	0.1344*	0.1344*	0.1349	0.1396**
24	0.5137	0.1227	0.1218*	0.1218*	0.1221	0.126**
26	0.5575	0.1476	0.146*	0.146*	0.1466	0.1522**
28	0.48	0.148	0.1464*	0.1464*	0.147	0.1526*
30	0.5976	0.1479	0.1463*	0.1463*	0.1468	0.1526**
32	0.5043	0.1812	0.1783*	0.1783*	0.1793	0.1877**
34	0.6514	0.1789	0.1762*	0.1764	0.1769	0.186**
37.5	0.5771	0.1366	0.1353*	0.1354	0.1358	0.1407**
42.5	0.7228	0.1746	0.1721*	0.1729	0.1724	0.1824**

Table A.8: Length of the confidence intervals for parvovirus prevalence

Bibliography

[1] Farrington C. P. Modelling forces of infection for measles, mumps and rubella. *Statistics in Medicine*, 9:953–967, 1990.

[2] Thiry N. Beutels P. Shkedy Z. Vranckx R. Vandermeulen C. Van Der Wielen M. and Van Damme P. The seroepidemiology of primary varicella-zoster virus infection in flanders (belgium). *European Journal of Pediatrics*, 161:588–593, 2002.

[3] Farrington C. P. Kanaan M. N. and Gay N. J. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Applied Statistics*, 50:251–292, 2001.

[4] Keiding N. Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series A*, 154:371–412, 1991.

[5] Muench H. *Catalytic Models in Epidemiology.* Harvard University Press, Cambridge, Massachusetts, 1959.

[6] Yamaguchi K. *Event History Analysis. Applied Social Research Methods Series*, volume 28. SAGE Publications, Newbury Park-London New Delhi, 1991.

[7] Mossong J. Putz L. and Schneider F. Seroprevalence and force of infection of varicella-zoster virus in luxembourg. *Epidemiology and Infection*, 132:1121–1127, 2004.

[8] Cohen D. I. Davidovici B. B. Smetana Z. Balicer R. D. Klement E. Mendelson E. and Green M. S. Seroepidemiology of varicella zoster in israel prior to large-scale use of varicella vaccines. *Infection*, 34(4):208–213, 2006.

[9] Brown L. D. Cay T. T. and DasGupta A. Interval estimation for a binomial proportion (with discussion). *Statistical Science*, 16(2):101–133, 2001.

[10] Wilson E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[11] Agresti A. and Coull B. A. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[12] Clopper C. J. and Pearson E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.

[13] Griffiths D. A. A catalytic model of infection for measles. *Applied Statistics*, 23(3):330–339, 1974.

[14] Grenfell B. T. and Anderson R. M. The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene*, 95:419–436, 1985.

[15] Edmunds W. J. Gay N. J. Kretzschmar M. Pebody R. G. and Wachmann H. The pre vaccination epidemiology of measles, mumps and rubella in europe: implications for modeling studies. *Epidemiology and Infection*, 125:635–650, 2000.

[16] Becker N. G. *Analysis of Infectious Disease Data*. Chapman and Hall, New York, 1989.

[17] Diamond I. D. and McDonald J. M. Analysis of current-status data. In Trussel J. Hankinson R. and Tiltan J., editors, *Demographic Application of Event History Analysis*, chapter 12. Oxford University Press, Oxford, 1992.

[18] Keiding N. Begtrup K. Scheike T. H. and Hasibeder G. Estimation from current status data in continuous time. *Lifetime Data Analysis*, 2:119–129, 1996.

[19] Jewell N. P. and Van Der Laan M. Generalizations of current status data with applications. *Lifetime Data Analysis*, 1:101–109, 1995.

[20] Grummer-Strawn L. M. Regression analysis of current status data: an application to breast feeding. *Journal of the American Statistical Association*, 88:758–765, 1993.

[21] Shkedy Z. Aerts M. Molenberghs G. Beutels P. H. and Van Damme P. Modeling forces of infection by using monotone local polynomials. *Applied Statistics*, 52(4):469–486, 2003.

[22] Shiboski S. C. Generalized additive models for current status data. *Lifetime Data Analysis*, 4:29–50, 1998.

[23] Hastie T. J. and Tibshirani R. J. *Generalized Additive Models*. Chapman and Hall, New York, 1990.

[24] Rossini A. J. and Tsiatis A. A. A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91(423):713–721, 1996.

[25] Martinussen T. and Scheike T. H. A flexible additive multiplicative hazard model. *Biometrika*, 89(2):283–298, 2002.

[26] Lin D. Y. Oakes D. and Ying Z. Additive hazard regression with current status data. *Biometrika*, 85(2):289–298, 1998.

[27] Bard Y. *A Function Maximization Method with Application to Parameter Estimation*. New York Scientific Center Report 322.0902, IBM, New York, 1967.

[28] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[29] Nelder J. A. and Wedderburn R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[30] McCullagh P. and Nelder J. A. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[31] Dobson A. J. *An Introduction to Generalized Linear Models*. Chapman and Hall, London, 1990.

[32] Kullback S. *Information Theory and Statistics*. Wiley, New York, 1959.

[33] Hastie T. A closer look at the deviance. *The American Statistician*, 41:16–20, 1987.

[34] Wedderburn R. W. M. Quasi-likelihood functions, generalized linear models and the gauss-newton method. *Biometrika*, 61:439–447, 1974.

[35] Royston P. and Altman D. G. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics*, 43(3):429–467, 1994.

[36] Whitaker H. J. and Farrington C. P. Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. *Statistics in Medicine*, 23:2429–2443, 2004.

[37] Gay N J. Analysis of serological surveys using mixture models: application to a survey of parvovirus b19. *Statistics in Medicine*, 15:1567–1573, 1996.

[38] Van Herck K. Beutels P. Van Damme P. Beutels M. Van den Dries J. Briantais Ph. and Vidor E. Mathematical models for assessment of long-term persistence of antibodies after vaccination with two inactivated hepatitis a vaccines. *Journal of Medical Virology*, 60:1–7, 2000.

[39] Kanaan M. N. and Farrington C. P. Estimation of waning vaccine efficacy. *Journal of the American Statistical Association*, 97(458):389–397, 2002.

[40] Farrington C. P. Estimating prevalence by group testing using generalized linear models. *Statistics in Medicine*, 11:1591–1597, 1992.

[41] Farrington C. P. Interval censored survival data: a generalized linear modelling approach. *Statistics in Medicine*, 15:283–292, 1996.

[42] Brown L. D. Cay T. T. and DasGupta A. Interval estimation in exponential families. *Statistica Sinica*, 13:14–49, 2003.

[43] Blyth C. R. and Still H. A. Binomial confidence intervals. *Journal of the American Statistical Association*, 78(381):108–116, 1983.

[44] Farrington C. P. Interval-censored survival data with informative examination times: parametric models and approximate inference. *Statistics in Medicine*, 18:1235–1248, 1999.