Imperial College
London

M3S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2019

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Statistical Modelling 2

Date: Thursday 16 May 2019

Time: 10.00 - 12.00

Time Allowed: 2 Hours

This paper has 4 Questions.

Candidates should use ONE main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Calculators may not be used.

Imperial College
London

M4/5S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2019

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Statistical Modelling 2

Date: Thursday 16 May 2019

Time: 10.00 - 12.30

Time Allowed: 2 Hours 30 Minutes

This paper has 5 Questions.

Candidates should use ONE main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Calculators may not be used.

1. This question concerns a Normal linear model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2 I_n)$. The design matrix may be assumed to have full rank.

(a) Give an expression for $\widehat{\boldsymbol{\beta}}$ in terms of X and $\boldsymbol{Y}$, and use standard properties of expectation and covariance to determine the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ in terms of σ^2.

Consider the specific case in which the design matrix is given by

$$X = \begin{bmatrix} 1 & -2 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

(b) Determine the variance-covariance matrix of the least squares estimator $\widehat{\boldsymbol{\beta}}$, in terms of σ^2.

(c) Find the variance of $\widehat{\beta_1} + \widehat{\beta_2} + \widehat{\beta_3}$.

(d) Suppose a large number of realizations $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ are drawn from the model whose design matrix is given above, and the estimates $\widehat{\boldsymbol{\beta}_1}, \ldots, \widehat{\boldsymbol{\beta}_m}$ computed for each realization. Two scatter plots are then made, one showing estimates of β_1 against those of β_2, and another showing estimates of β_2 against those of β_3. Describe the appearance of the two plots, noting in particular how they would differ.

(e) Consider the three QQ-plots in Fig 1, which show the residuals from linear models applied to three different datasets. Comment on the appearance of each plot, and suggest properties of the model or error structure that could give rise to each.

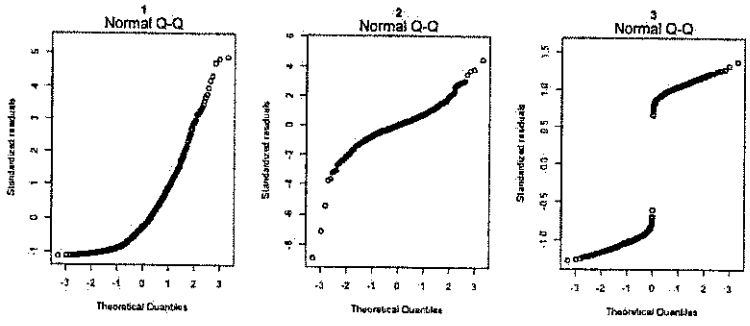

Figure 1: QQ plots of standardized residuals for linear models fitted to three different datasets.

2. (a) Define the three components of a generalized linear model (GLM).

Consider a binomial GLM, in which n independent observations Y_i are made of a response variable with probability distribution

$$\Pr(Y_i = k) = \binom{n_i}{k} p_i^k (1-p_i)^{n_i - k}, \qquad 0 \leq k \leq n_i,$$

where n_i is a fixed number for each observation i.

(b) State the relationship between $\mu_i = \mathrm{E}(Y_i)$ and p_i, and hence write down the log likelihood for this model in terms of μ_i.

(c) Show that the deviance for this model can be written as

$$2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\widehat{\mu_i}}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{\mu_i}}\right).$$

(d) Give the form of the deviance residual for this model.

Consider in what follows the case of binary logistic regression, i.e. suppose that $n_i = 1$ for each i, and let the linear predictor $\eta = X\beta$ be given, and related to the mean of the response variable by the logit link,

$$\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

(e) Describe the appearance of a plot of the deviance residuals against the fitted values μ_i, and give a brief explanation in terms of the deviance.

[Question continues overleaf]

A prospective study collected data to determine the effect of birth weight on the probability of developing a particular childhood illness. A data analysis in R is shown below. `ill` is a binary indicator for disease status, and `wt` is weight (in kg).

```
Call:
glm(formula = ill ~ wt, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09203    1.08690  -0.085   0.9325
wt           0.98006    0.38399   2.552   0.0107 *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.465  on 149  degrees of freedom
Residual deviance: 55.587  on 148  degrees of freedom
AIC: 59.587
```

(f) Give a plain language summary of the model output, with clear reference to the effect of birth weight.

(g) Comment on whether the residual deviance can be used as a measure of goodness of fit in this case.

(h) Explain how, if at all, your interpretation of these results would change if they had come from a retrospective, rather than a prospective, study.

3. This question concerns a response variable Y that is gamma distributed,

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}e^{-\frac{\nu y}{\mu}}, \quad \mu,\nu,y>0.$$

(a) Show that the response is a member of the exponential family.

(b) In the Michaelis-Menten model of enzyme kinetics, the reaction rate y can be modelled as

$$y = \frac{\alpha_0 x}{1+\alpha_1 x},$$

where α_1 and α_2 are constants and x is the concentration of a substrate. If this model is taken as determining the relationship between $\mathrm{E}(Y)$ and the non-random covariate x, show that we can form a gamma GLM with the canonical link, for

$$\eta = \beta_0 + \frac{\beta_1}{x},$$

where the relationships between the regression coefficients β and the parameters α_0 and α_1 are to be determined.

The code at the end of the question implements the iterated weighted least squares algorithm for a large random sample of size n from the Michaelis-Menten model, but some parts are missing, labelled #####. (In your responses below, you are not expected to produce any code.)

(c) Derive the form of the deviance function `D`.

(d) Derive the form of the adjusted response variable `z` and weights `w`.

(e) Suggest how a sensible starting value of the coefficients β could be chosen.

(f) Explain how the dispersion parameter could be estimated, assuming any necessary asymptotic results, which you should state.

(g) Explain how to use the output of this algorithm i.e. the parameter values and weights at convergence, to construct confidence intervals for

(i) the coefficient β_2,

(ii) the mean value of y for a particular value x of the covariate.

(h) An alternative model is proposed in which the linear predictor takes the form

$$\eta_i = \beta_0 + \frac{\beta_1}{x_i} + \beta_2 x_i.$$

Stating any asymptotic results you need, explain how to test the null hypothesis $\beta_2 = 0$.

```
beta <- #####   #initial guess
y<-dat$rate
#gamma deviance function
D <- function(mu,y){
   #####
}

#compute the initial deviance
oldD <- D(inv.link(X%*%beta),y)
jj <- 0

while(jj==0){
    eta <- X%*%beta
    mu <- inv.link(eta)
    z <- ####
    w <- ####
    lmod <- lm(z~X,weights=w,data=dat)
    beta <- as.numeric(lmod$coeff)
    newD <- D(inv.link(X%*%beta),y)
    control <- abs(newD-oldD)/(abs(newD)+0.1)
    if(control<1e-8)
        jj <- 1
    oldD <- newD
}
```

4. This question concerns the *penicillin* dataset, which was used in a tutorial class. The response variable is the yield of penicillin produced in four different processes A,B, C and D. The raw material used by these processes is produced in blends, which can be quite variable, and blends are only made in quantities large enough for four runs. The data therefore represent 20 observations, indexed y_{ij}, where $i \in \{1,2,3,4\}$ is the process type and $j \in \{1,2,3,4,5\}$ is the blend. Each process is tested with each blend.

(a) The code below is used to fit two models, and compare them.

```
> fit0<-lm(yield~treat,data=penicillin)
> fit1<-lm(yield~treat+blend,data=penicillin)
> anova(fit0,fit1)
Analysis of Variance Table

Model 1: yield ~ treat
Model 2: yield ~ treat + blend
  Res.Df RSS Df Sum of Sq      F  Pr(>F)
1     16 490
2     12 226  4       264 3.5044 0.04075
```

(i) Explain why four additional degrees of freedom are needed for `blend`.

(ii) State the null hypothesis that is tested using the anova command. Give the test statistic and state its distribution under the null hypothesis.

(iii) Summarize the conclusion of the hypothesis test.

(iv) Explain the problem that arises when attempting to assess goodness of fit for the model `yield ~ treat * blend`.

[Question continues overleaf]

(b) An alternative model is proposed, in which blend is considered as a random effect,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}.$$

The matrix $\boldsymbol{Z}$ specifies which measurements correspond to which blend, and $\boldsymbol{\nu} \sim (N, \sigma_\nu^2)$ encodes the random effects. The fixed effect structure encoded by $X\boldsymbol{\beta}$ is as before, using treatment contrasts, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$.

The model is fitted using the code below. (Note that the output has been abridged.)

```
> fit2<-lmer(yield~treat +(1|blend),data=penicillin)
> summary(fit2)
Linear mixed model fit by REML ['lmerMod']
Formula: yield ~ treat + (1 | blend)
   Data: penicillin

REML criterion at convergence: 103.8

Random effects:
 Groups   Name        Variance
 blend    (Intercept) 11.79
 Residual             #???#
Number of obs: 20, groups:  blend, 5

Fixed effects:
            Estimate Std. Error
(Intercept)   84.000      2.475
treatB         1.000      2.745
treatC         5.000      2.745
treatD         2.000      #@@@#
```

(i) Suggest why blend might be included in the model as a random effect.

(ii) Give the distribution of $\boldsymbol{Y}$.

(iii) Determine the covariance between two observations with the same blend.

(iv) Explain briefly how REML is used to estimate parameters.

(v) Use the information given to determine the residual variance in `fit2`, labelled `#???#`. Leave your answer as a fraction.

(vi) Determine the missing standard error for `treatD`, labelled `#@@@#`, justifying your answer.

5. This question concerns the extract provided from *Generalized Additive Models by Simon Wood.*

These parts refer to Section 2.3.1 *Binomial Models and Heart Disease* and 2.3.2 *A Poisson Regression Model.*

(a) Give the reasoning that would lead the author to assert on page 84 that the deviance should follow a χ^2_{10} distribution if the model is fitting well.

(b) Suggest two advantages of using the canonical link for the binomial model.

(c) Explain why the pattern in the mean of the residuals in Fig 2.6 might be problematic.

(d) State how the solid line in three of the residual plots in Fig 2.6 should be interpreted.

(e) Justify the assertion in 2.3.2 that the log-link specifies a model of "unchecked spread of the disease".

The following parts refer to Section 2.2.1, *The Geometry of IRLS.*

(f) Explain in geometrical terms why fitting GLMs is typically more difficult than fitting linear models. Describe how weighted least squares addresses these difficulties.

(g) In Fig 2.2, explain why the collections of points (y_1, y_2) sharing the same MLE form straight lines.

(h) Give the equation of the bold model manifold curve in Fig 2.2 in the form $\mu_2 = f(\beta)\mu_1$.

(i) Suggest the approximate value of β corresponding to the observation $(y_1, y_2) = (5, 9)$.

and Nelder (1989) give examples, or in R you can type e.g.
`quasi(variance="mu^3")$dev.resids`
to access the form of q_i for any particular mean variance relationship there implemented. For mean variance relationships corresponding to an exponential family distribution from table 2.1, the form of the quasi-deviance corresponds exactly to the form of the deviance for that family.

One major practical use of quasi-likelihood is to provide a means of modelling count data that are more variable than the Poisson or binomial distributions (with their fixed scale parameters) predict: the quasi-likelihood approach assumes that ϕ is unknown. Such 'over-dispersed' data are common in practice. Another practical use is to provide a means of modelling data with a mean variance relationship for which there is no obvious exponential family distribution: for example continuous data for which the variance is expected to be proportional to the mean.

2.2 Geometry of GLMs

The geometry of GLMs and GLM fitting is less straightforward than the geometry of ordinary linear models, since the likelihood used to judge model fit does not generally mean that the fit can be judged by Euclidian distance between model and data. Figure 2.1 illustrates the geometric situation that prevails for GLMs, using the example of the fit to 3 data of a 2 parameter GLM with a Gamma distribution and a log link. The flat model subspace of section 1.4 is now replaced by a curved 'model manifold', consisting of all the possible fitted value vectors predictable by the model. Since Euclidean distance between model manifold and data is no longer the measure of fit being used then different means must be employed to illustrate the geometry of estimation. The black lines, in the right panel of figure 2.1, show all the combinations of the response variables, which give rise to the same estimated model. Notice how these lines are not generally parallel, and are not generally orthogonal to the model manifold.

To fully understand figure 2.1, it may help to consider what the figure would look like for some different 2 parameter models.

1. For an ordinary linear model, the model manifold would be a flat plane, to which all the lines of equal fit would be orthogonal (and hence parallel to each other).
2. For a GLM assuming a normal distribution (but non-identity link) the lines of equal fit would be orthogonal to the (tangent space of the) model manifold where they meet it.
3. For a 2 parameter fit to 4 data, the lines of equal fit would become planes of equal fit.

In general, the geometric picture presented in figure 2.1 applies to any GLM. With more data the lines of equal fit become $n-p$ dimensional planes of equal fit, where n and p are the number of data and parameters respectively: for any fixed $\boldsymbol{\beta}$, equation (2.3) gives the restrictions on $\mathbf{y}$ defining such a plane. Note that these planes

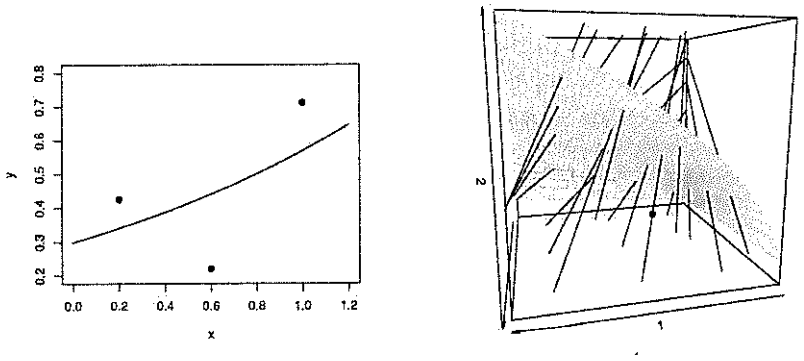

Figure 2.1 *The geometry of GLMs. The left panel illustrates the best fit of the generalized linear model* $\mathbb{E}(y) \equiv \mu = \exp(\beta_0 + \beta_1 x)$ *to the three* x, y *data shown, assuming that each* y_i *is an observation of a Gamma distributed random variable with mean given by the model. The right panel illustrates the geometry of GLM fitting using this model as an example. The unit cube shown, represents a space within which the vector* $(y_1, y_2, y_3)^\mathsf{T}$ *defines a single point,* •*. The grey surface shows all possible predicted values (within the unit cube) according to the model, i.e. it represents all possible* $(\mu_1, \mu_2, \mu_3)^\mathsf{T}$ *values. As the parameters* β_0 *and* β_1 *are allowed to vary, over all their possible values, this is the surface that the corresponding model 'fitted values' trace out: the 'model manifold'. The continuous lines, which each start at one face of the cube and leave at another, are lines of equivalent fit: the values of the response data* $(y_1, y_2, y_3)^\mathsf{T}$ *lying on such a line, each result in the same maximum likelihood estimates of* β_0, β_1 *and hence the same* $(\mu_1, \mu_2, \mu_3)^\mathsf{T}$*. Notice how the equivalent fit lines are neither parallel to each other nor orthogonal to the model manifold.*

can intersect — a point which will be returned to later. For discrete response data the pictures are no different, although the lines of equal fit strictly make sense only under continuous generalizations of the likelihoods (generally obtainable by replacing factorials by appropriate gamma functions in the probability functions). Only for the normal distribution are the lines/planes of equal fit orthogonal to the model manifold where-ever they meet it. For other distributions the lines/planes of equal fit may sometimes be parallel to each other, but are never all orthogonal to the model manifold.

2.2.1 The geometry of IRLS

The geometry of the IRLS estimation algorithm is most easily appreciated by considering the fit of a one parameter model to 2 response data. Figure 2.2 illustrates the geometry of such a model: in this case a GLM with a log link and Gamma errors,

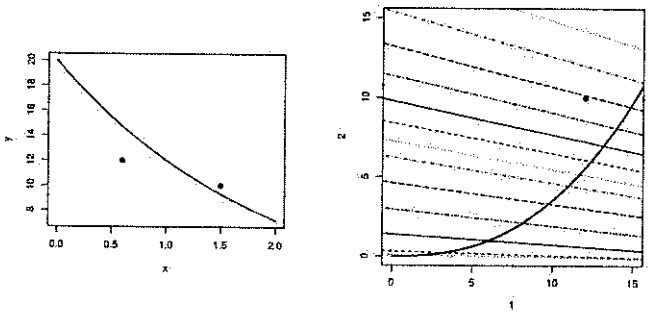

Figure 2.2 *Geometry of the GLM* $\mathbb{E}(y_i) \equiv \mu_i = 20\exp(-\beta x_i)$ *where* $y_i \sim$ Gamma *and* $i = 1, 2$. *The left panel illustrates the maximum likelihood estimate of the model (continuous line) fitted to the* 2 x, y *data shown as* •. *The right panel illustrates the fitting geometry. The* 15×15 *square is part of the space* $\Re^2$ *in which* (y_1, y_2) *defines a single point,* •. *The bold curve is the 'model manifold': it consists of all possible points* (μ_1, μ_2) *according to the model (i.e. as* β *varies* (μ_1, μ_2) *traces out this curve). The fine lines are examples of lines of equal fit. All points* (y_1, y_2) *lying on one of these lines share the same MLE of* β *and hence* (μ_1, μ_2)*: this MLE is where the equal fit line cuts the model manifold. The lines of equal fit are plotted for* $\beta = .1, .2, .3, .4, .5, .6, .7, .8, .9, 1, 1.2, 1.5, 2, 3, 4$. *(*$\beta = .1, .7$ *and* 2 *are represented by unbroken lines, with the* $\beta = 2$ *line being near the bottom of the plot. The* $\beta = .1$ *line is outside the plotting region in this plot, but appears in subsequent plots.)*

but similar pictures can be constructed for a GLM with any combination of link and distributional assumption.

Now the key problems in fitting a GLM are that the model manifold is not flat, and that the lines of equal fit are not orthogonal to the model manifold where they meet it. The IRLS method linearly translates and rescales the fitting problem, so that at the current estimate of $\boldsymbol{\mu}$, the model manifold and intersecting line of equal fit are orthogonal, and, in the rescaled space, the location of the current estimate of $\boldsymbol{\mu}$ is given by $\mathbf{X}$ multiplied by the current $\boldsymbol{\beta}$ estimate. This rescaling results in a fitting problem that can be treated as locally linear, so that the $\boldsymbol{\beta}$ estimate can be updated by least squares.

Figure 2.3 illustrates how the IRLS steps involved in forming pseudodata and weighting it, effectively transform the fitting problem into one that can be approximately solved by linear least squares. The figure illustrates the transformations involved in one IRLS step, which are redone repeatedly, as the IRLS method is iterated to convergence.

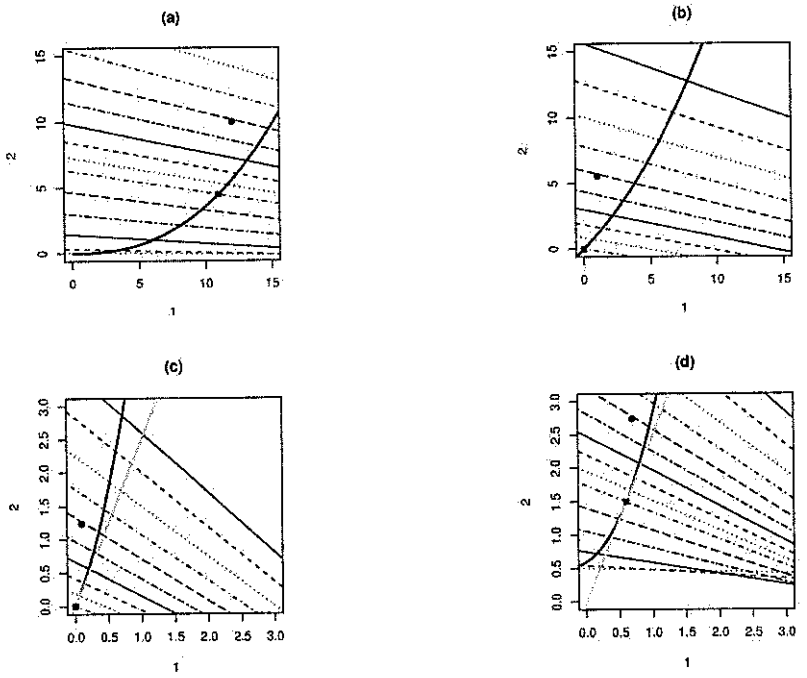

Figure 2.3 *Geometry of the IRLS estimation of a GLM, based on the example shown in figure 2.2. (a) shows the geometry of the fitting problem — the model manifold is the thick black curve, the equal fit lines are the thin lines (as figure 2.2), the data are at • and the current estimates of the fitted values, $\boldsymbol{\mu}^{[k]}$, are at ■. (b) The problem is re-centred around the current fitted values (y_i is replaced by $y_i - \mu_i^{[k]}$). (c) The problem is linearly re-scaled so that the columns of $\mathbf{X}$ now span the tangent space to the model manifold at ■. The tangent space is illustrated by the grey line (this step replaces $y_i - \mu_i^{[k]}$ by $g'(\mu_i^{[k]})(y_i - \mu_i^{[k]})$). (d) The problem is linearly translated so that the location of ■ is now given by $\mathbf{X}\boldsymbol{\beta}^{[k]}$. For most GLMs the problem would now have to be rescaled again by multiplying the components relative to each axis by $\sqrt{W_i}$, where the W_i are the iterative weights: this would ensure that the equal estimate line through ■, is orthogonal to the tangent space. In the current example these weights are all 1, so that the required orthogonality already holds. Now for the transformed problem, in the vicinity of ■, the model manifold can be approximated by the tangent space, to which the equal fit lines are approximately orthogonal: hence an updated estimate of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ can be obtained by finding the least squares projection of the transformed data, •, onto the tangent space (grey line).*

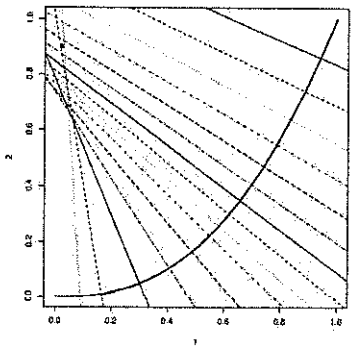

Figure 2.4 *Geometry of fitting and convergence problems. The geometry of a 1 parameter GLM with a log link and normal errors is illustrated. The thick curve is the model manifold — within the unit square, it contains all the possible fitted values of the data, according to the model. The thin lines are the equal fit lines (levels as in figure 2.2). Notice how the lines of equal fit meet and cross each other at the top left of the plot. Data in this overlap region will yield model likelihoods having local minima at more than one parameter value. Consideration of the operation of the IRLS fitting method reveals that, in this situation, it may converge to different estimates depending on the initial values used to start the fitting process.* • *illustrates the location of a problematic response vector, used to illustrate non-unique convergence in the text.*

2.2.2 *Geometry and IRLS convergence*

Figure 2.4 illustrates the geometry of fitting a model, $\mathbb{E}(y_i) \equiv \mu_i = \exp(-\beta x_i)$, where the y_i are normally distributed and there are two data, y_i, to fit, for which $x_1 = .6$ and $x_2 = 1.5$. As in the previous two sections, lines of equal fit are shown on a plot in which a response vector $(y_1, y_2)^\mathsf{T}$ would define a single point and the set of all possible fitted values $(\mu_1, \mu_2)^\mathsf{T}$, according to the model, is shown as a thick curve. In this example, the lines of equal fit intersect and cross in the top left hand corner of the plot (corresponding to very poor model fit). This crossing is problematic: in particular, the results of IRLS fitting to data lying in the upper left corner will depend on the initial parameter estimate from which the IRLS process is started, since each such data point lies on the intersection of two equal fit lines. If the IRLS iteration is started from fitted values in the top right of the plot then fitted values nearer the top right will be estimated, while starting the iteration with fitted values at the bottom left of the plot will result in estimated fitted values that are different, and closer to the bottom left of the plot.

That this indeed happens, in practice, is easily demonstrated in R, by fitting to the data $y_1 = .02$, $y_2 = .9$, illustrated as • in figure 2.4.

```
> ms<-exp(-x*4)    # set initial values at lower left
```

```
> glm(y~X-1,family=gaussian(link=log),mustart=ms)
Coefficients:
5.618
Residual Deviance: 0.8098       AIC: 7.868
> ms <- exp(-x*0.1)  # set initial values at upper right
> glm(y~X-1,family=gaussian(link=log),mustart=ms)
Coefficients:
0.544
Residual Deviance: 0.7017       AIC: 7.581
```

Notice that the second fit here actually has higher likelihood (lower deviance) — the fits are not equivalent in terms of likelihood. The type of fitting geometry that gives rise to these ambiguities does not always occur: for example some models have parallel lines/planes of equal fit, but for any model with intersecting lines/planes of equal fit there is some scope for ambiguity. Fortunately, if the model is a good model, it is often the case that data lying in the region of ambiguity is rather improbable. In the example in figure 2.4, the problematic region consists entirely of data that the model can only fit very poorly. It follows that very poor models of data may yield estimation problems of this sort: but it is not uncommon for very poor models to be a feature of early attempts to model any complex set of data. If such problems are encountered then it can be better to proceed by linear modelling of transformed response data, until good enough candidate models have been identified to switch back to GLMs.

Of course, if reasonable starting values are chosen, then the ambiguity in the fitting process is unlikely to cause major problems when fitting GLMs: the algorithm will converge to one of the local minima of the likelihood, after all. However the ambiguity can cause more serious convergence problems for GAM estimation by "performance iteration", when it becomes possible to cycle between alternative minima without ever converging.

2.3 GLMs with R

The `glm` function provides the means for using GLMs in R. Its use is similar to that of the `lm` function but with two differences. The right hand side of the model formula, specifying the form for the linear predictor, now gives the link function of the mean of the response, rather than the mean of the response directly. Also `glm` takes a `family` argument, which is used to specify the distribution from the exponential family to use, and the link function that is to go with it. In this section the use of the `glm` function with a variety of simple GLMs will be presented, to illustrate the wide variety of model structures that the GLM encompasses.

2.3.1 Binomial models and heart disease

Early diagnosis of heart attack is important if the best care is to be given to patients. One suggested diagnostic aid is the level of the enzyme creatinine kinase (CK) in

CK value	Patients with Heart attack	Patients without heart attack
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

Table 2.2 *Data (from Hand et al., 1994) on heart attack probability as a function of CK level.*

the blood stream. A study was conducted (Smith, 1967) in which the level of CK was measured for 360 patients suspected of suffering from a heart attack. Whether or not each patient had really suffered a heart attack was established later, after more prolonged medical investigation. The data are given in table 2.2. The original paper classified patients according to ranges of CK level, but in the table only midpoints of the range have been given.

It would be good to be able to base diagnostic criteria on data like these, so that CK level can be used to estimate the probability that a patient has had a heart attack. We can go some way towards such a goal, by constructing a model which tries to explain the proportion of patients suffering a heart attack, from the CK levels. In the following the data were read into a data.frame called `heart`. It contains variables `ha`, `ok` and `ck`, giving numbers of patients who subsequently turned out to have had, or not to have had, heart attacks, at each CK level. It makes sense to plot the observed proportions against CK level first.

```
p<-heart$ha/(heart$ha+heart$ok)
plot(heart$ck,p,xlab="Creatinine kinase level",
     lab="Proportion Heart Attack")
```

The resulting plot is figure 2.5.

A particularly convenient model for describing these proportions is

$$\mathbb{E}(p_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}},$$

where p_i is the proportion with heart attacks at CK level x_i. This curve is sigmoid in

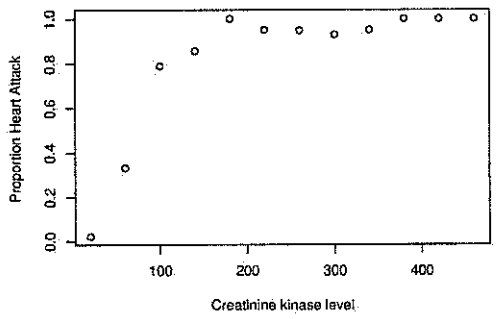

Figure 2.5 *Observed proportion of patients subsequently diagnosed as having had a heart attack, against CK level at admittance .*

shape, and bounded between 0 and 1. (Obviously the `heart` data do not show the lower tail of this proposed sigmoid curve.) This means that the expected number of heart attack sufferers is given by

$$\mu_i \equiv \mathbb{E}(p_i N_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} N_i,$$

where N_i is the known total number of patients at each CK level. This model is somewhat non-linear in its parameters, but if the 'logit' link,

$$g(\mu_i) = \log\left(\frac{\mu_i}{N_i - \mu_i}\right),$$

is applied to it we obtain

$$g(\mu_i) = \beta_0 + \beta_1 x_i,$$

the r.h.s. of which is linear in the model parameters. The logit link is the canonical link for binomial models, and hence the default in R.

In R there are two ways of specifying binomial models with `glm`.

1. The response variable can be the observed proportion of successful binomial trials, in which case an array giving the number of trials must be supplied as the `weights` argument to `glm`. For binary data, no weights vector need be supplied, as the default weights of 1 suffice.
2. The response variable can be supplied as a two column array, in which the first column gives the number of binomial 'successes', and the second column is the number of binomial 'failures'.

For the current example the second method will be used. Supplying 2 arrays of the

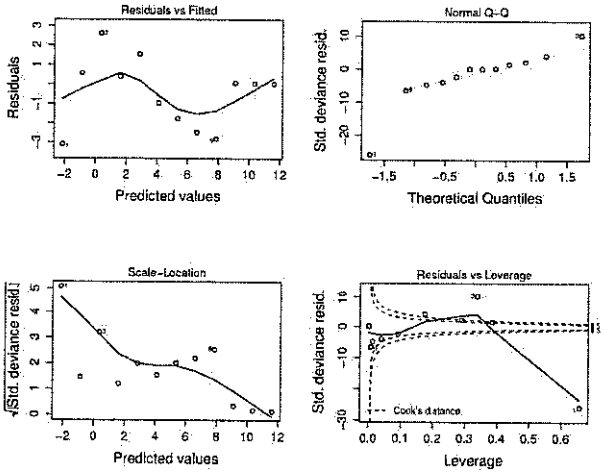

Figure 2.6 *Model checking plots for the first attempt to fit the CK data.*

r.h.s. of the model formula involves using `cbind`. Here is a `glm` call which will fit the heart attack model:

```
> mod.0<-glm(cbind(ha,ok)~ck,family=binomial(link=logit),
+ data=heart)
```

or we could have used

```
mod.0<-glm(cbind(ha,ok)~ck,family=binomial,data=heart)
```

since the logit link is canonical for the binomial and hence the R default. Here is the default information printed about the model:

```
> mod.0

Call: glm(formula=cbind(ha,ok)~ck,family=binomial,data=heart)

Coefficients:
(Intercept)           ck
   -2.75834      0.03124

Degrees of Freedom: 11 Total (i.e. Null);  10 Residual
Null Deviance:      271.7
Residual Deviance: 36.93        AIC: 62.33
```

The `Null deviance` is the deviance for a model with just a constant term, while

the `Residual deviance` is the deviance of the fitted model (and also the scaled deviance in the case of a binomial model). These can be combined to give the *proportion deviance explained*, a generalization of r^2, as follows:

```
> (271.7-36.93)/271.7
[1] 0.864078
```

AIC is the Akaike Information Criteria for the model, discussed in sections 2.1.4 and 2.4.7 (it could also have been extracted using `AIC(mod.))`).

Notice that the deviance is quite high for the χ^2_{10} random variable that it should approximate if the model is fitting well. In fact

```
> 1-pchisq(36.93,10)
[1] 5.819325e-05
```

shows that there is a very small probability of a χ^2_{10} random variable being as large as 36.93. The residual plots (shown in figure 2.6) also suggest a poor fit.

```
> op<-par(mfrow=c(2,2))
> plot(mod.0)
```

The plots have the same interpretation as the model checking plots for an ordinary linear model, discussed in detail in section 1.5.1, except that it is now the deviance residuals that are plotted, the `Predicted values` are on the scale of the linear predictor rather than the response, and some departure from a straight line relationship in the Normal QQ plot is often to be expected. The plots are not easy to interpret when there are so few data, but there appears to be a trend in the mean of the residuals plotted against fitted value, which would cause concern. Furthermore, the first point has very high influence. Note that the interpretation of the residuals would be much more difficult for binary data: exercise 2 explores simple approaches that can be taken in the binary case.

Notice how the problems do not stand out so clearly from a plot of the fitted values overlayed on the raw estimated probabilities (see figure 2.7):

```
> plot(heart$ck,p,xlab="Creatinine kinase level",
+ ylab="Proportion Heart Attack")
> lines(heart$ck,fitted(mod.0))
```

Note also that the fitted values provided by `glm` for binomial models are the estimated p_i's, rather than the estimated μ_i's.

The residual plots suggest trying a cubic linear predictor, rather than the initial straight line.

```
> mod.2<-glm(cbind(ha,ok)~ck+I(ck^2)+I(ck^3),family=binomial,
+ data=heart)
> mod.2

Call: glm(formula=cbind(ha,ok)~ck+I(ck^2)+I(ck^3),
```

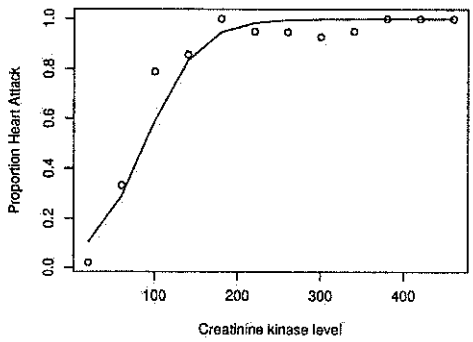

Figure 2.7 *Predicted and observed probability of heart attack against CK level.*

```
          family=binomial,data=heart)

Coefficients:
(Intercept)            ck       I(ck^2)       I(ck^3)
 -5.786e+00     1.102e-01    -4.648e-04     6.448e-07

Degrees of Freedom: 11 Total (i.e. Null);  8 Residual
Null Deviance:      271.7
Residual Deviance: 4.252        AIC: 33.66
> par(mfrow=c(2,2))
> plot(mod.2)
```

Clearly 4.252 is not too large for consistency with a χ^2_8 distribution (it is less than the expected value, in fact) and the AIC has improved substantially. The residual plots (figure 2.8) now show less clear patterns than for the previous model, although if we had more data then such a departure from constant variance would be a cause for concern. Furthermore the fit is clearly closer to the data now (see figure 2.9):

```
par(mfrow=c(1,1))
plot(heart$ck,p,xlab="Creatinine kinase level",
     ylab="Proportion Heart Attack")
lines(heart$ck,fitted(mod.2))
```

We can also get R to test the null hypothesis that `mod.0` is correct against the alternative that `mod.2` is required. Somewhat confusingly the `anova` function is used to do this, although it is an analysis of **deviance** (i.e. a generalized likelihood ratio test) that is being performed, and not an analysis of variance.

```
> anova(mod.0,mod.2,test="Chisq")
Analysis of Deviance Table
```

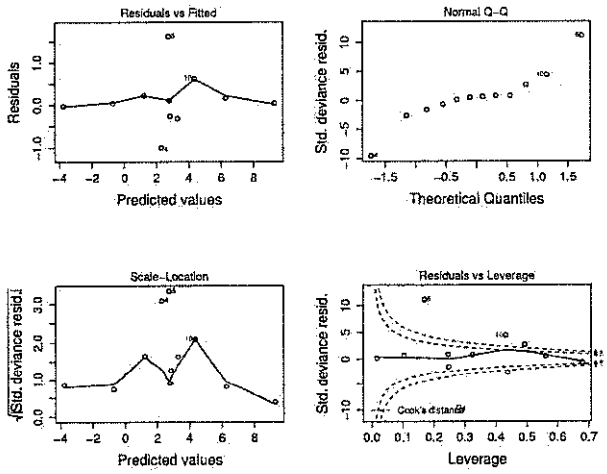

Figure 2.8 *Model checking plots for the second attempt to fit the CK data.*

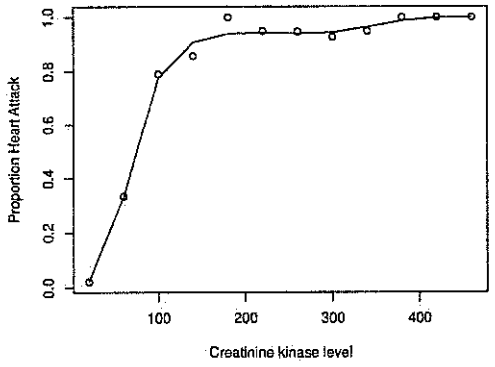

Figure 2.9 *Predicted and observed probability of heart attack against CK level.*

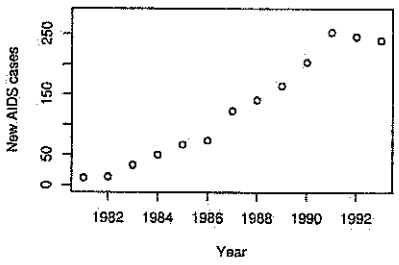

Figure 2.10 *AIDS cases per year in Belgium*

```
Model 1: cbind(ha, ok) ~ ck
Model 2: cbind(ha, ok) ~ ck + I(ck^2) + I(ck^3)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        10     36.929
2         8      4.252  2   32.676 8.025e-08
```

A p-value this low indicates very strong evidence against the null hypothesis - we really do need model 2. Recall that this comparison of models has a much firmer theoretical basis than the examination of the individual deviances had.

2.3.2 *A Poisson regression epidemic model*

The introduction to this chapter included a simple model for the early stages of an epidemic. Venables and Ripley (2003) provide some data on the number of new AIDS cases each year, in Belgium, from 1981 onwards. The data can be entered into R and plotted as follows.

```
y<- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
t<-1:13
plot(t+1980,y,xlab="Year",ylab="New AIDS cases",ylim=c(0,280))
```

Figure 2.10 shows the resulting plot. The scientifically interesting question, relating to such data, is whether they provide any evidence that the increase in the underlying rate of new case generation is slowing. The simple model from the introduction might provide a plausible model from which to start investigating this question. The model assumes that the underlying expected number of cases per year, μ_i, increases according to:

$$\mu_i = c\exp(bt_i)$$

where c and b are unknown parameters, and t_i is time in years since the start of the

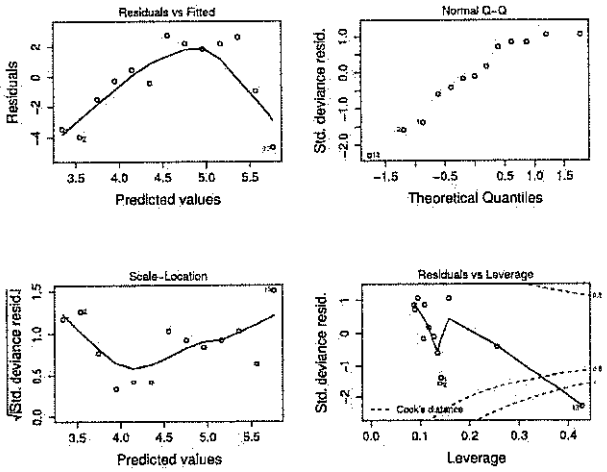

Figure 2.11 *Residual plots for* m0 *fitted to the AIDS data.*

data. A log link turns this into a GLM,

$$\log(\mu_i) = \log(c) + bt_i = \beta_0 + t_i\beta_1,$$

and we assume that $y_i \sim \text{Poi}(\mu_i)$ where y_i is the observed number of new cases in year t_i. The y_i are assumed independent. This is essentially a model of unchecked spread of the disease.

The following fits the model (the log link is canonical for the Poisson distribution, and hence the R default) and checks it.

```
> m0 <- glm(y~t,poisson)
> m0

Call:  glm(formula = y ~ t, family = poisson)

Coefficients:
(Intercept)            t
     3.1406       0.2021

Degrees of Freedom: 12 Total (i.e. Null);  11 Residual
Null Deviance:      872.2
Residual Deviance: 80.69        AIC: 166.4
> par(mfrow=c(2,2))
> plot(m0)
```

The deviance is very high for the observation of a χ^2_{11} random variable that it ought

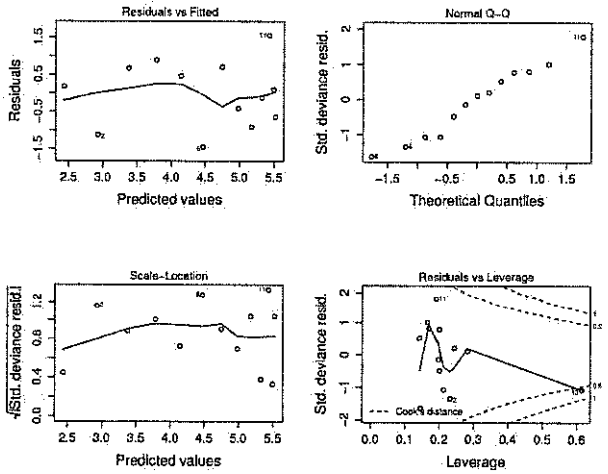

Figure 2.12 *Residual plots for* `m1` *fitted to the AIDS data.*

to approximate, if the model is a good fit. The residual plots shown in figure 2.11 are also worrying. In particular the clear pattern in the mean of the residuals, plotted against the fitted values, shows violation of the independence assumption, and probably results from omission of something important from the model. Since, for this model, the fitted values increase monotonically with time, we would get the same sort of pattern if residuals were plotted against time — i.e. it appears that a quadratic term in time could usefully be added to the model. The very high influence of the final year's data, evident in the residuals versus leverage plot, is also worrying. Note that the interpretation of residual plots can become difficult if the Poisson mean is low, so that the data are mostly zeroes and ones. In such cases the simulation approaches covered in exercise 2 can prove useful, if adapted to the Poisson case.

It seems sensible to amend the model by adding a quadratic term to obtain:

$$\mu_i = \exp(\beta_0 + \beta_1 t_i + \beta_2 t_i^2).$$

This model allows situations other than unrestricted spread of the disease to be represented. The following fits and checks it:

```
> m1 <- glm(y~t+I(t^2),poisson)
> plot(m1)
> summary(m1)

Call:
glm(formula = y ~ t + I(t^2), family = poisson)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.45903  -0.64491   0.08927   0.67117   1.54596

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.901459   0.186877  10.175  < 2e-16 ***
t            0.556003   0.045780  12.145  < 2e-16 ***
I(t^2)      -0.021346   0.002659  -8.029 9.82e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 872.2058  on 12  degrees of freedom
Residual deviance:   9.2402  on 10  degrees of freedom
AIC: 96.924

Number of Fisher Scoring iterations: 4
```

Notice how the residual plots shown in figure 2.12 are now much improved: the clear trend in the mean has gone, the (vertical) spread of the residuals is reasonably even, the influence of point 13 is much reduced and the QQ plot is straighter. The fuller model summary shown for this model also indicates improvement: the deviance is now quite reasonable, i.e. close to what is expected for a χ^2_{10} r.v., and the AIC has dropped massively. All in all this model appears to be quite reasonable.

Notice also, how the structure of the `glm summary` is similar to an `lm summary`. The standard errors and p-values in the table of coefficient estimates is now based on the large sample distribution of the parameter estimators given in section 2.1.5. The `z value` column simply reports the parameter estimates divided by their estimated standard deviations. Since no dispersion parameter estimate is required for the Poisson, these z-values should be observations of N(0,1) r.v.s, if the true value of the corresponding parameter is zero (at least in the large sample limit), and the reported p-value is based on this distributional approximation. For `mod.1` the reported p-values are very low: i.e for each parameter there is clear evidence that it is not zero. Note that `Fisher Scoring iterations` are another name for IRLS iterations in the GLM context.

Examination of the coefficient summary table indicates that the hypothesis that $\beta_2 = 0$ can be firmly rejected, providing clear evidence that `mod.1` is preferable to `mod.0`. The same question can also be addressed using a generalized likelihood ratio test:

```
> anova(m0,m1,test="Chisq")
Analysis of Deviance Table

Model 1: y ~ t
Model 2: y ~ t + I(t^2)
```

```
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        11     80.686
2        10      9.240  1   71.446 2.849e-17
```

The conclusion is the same as before: the tiny p-value indicates that `mod.0` should be firmly rejected in favour of `mod.1`. Notice that the p-value from the summary and the analysis of deviance table are different, since they are based on fundamentally different approximate distributional results. The `test="Chisq"` argument to `anova` is justified because the scale parameter is known for this model, had it been estimated it would be preferable to set `test` to `"F"`.

The hypothesis testing approach to model selection is appropriate here, as the main question of interest is whether there is evidence, from these data, that the epidemic is spreading un-checked, or not. It would be prudent not to declare that things are improving if the evidence is not quite firm that this is true. If we had been more interested in simply finding the best model for predicting the data then comparison of AIC would be more appropriate, but leads to the same conclusion for these data.

The parameter β_1 can be interpreted as the rate of spread of the disease at the epidemic start: that is, as a sort of intrinsic rate of increase of the disease in a new population where no control measures are in place. Notice how the estimate of this parameter has actually increased substantially between the first and second models: it would have been possible to be quite badly misled if we had stuck with the first poorly fitting model. An approximate confidence interval for β_1 can be obtained in the usual manner, based on the large sample results from section 2.1.5. The required estimate and standard error are easily extracted using the `summary` function, as the following illustrates:

```
> beta.1 <- summary(m1)$coefficients[2,]
> ci <- c(beta.1[1]-1.96*beta.1[2],beta.1[1]+1.96*beta.1[2])
> ci # print 95% CI for beta_1
0.4662750 0.6457316
```

The use of the critical points of the standard normal distribution is appropriate, because the scale parameter is known for this model. Had it been estimated, then we would have had to use critical points from the t distribution, with degrees of freedom set to the residual degrees of freedom of the model (i.e number of data less number of estimated β parameters).

Another obvious thing to want to do, is to use the model to find a confidence interval for the underlying rate of case generation at any time. The following R code illustrates how to use the `predict.glm` function to find CI's for the underlying rate, over the whole period of the data, and plot these.

```
new.t<-seq(1,13,length=100)
fv <- predict(m1,data.frame(t=new.t),se=TRUE)
plot(t+1980,y,xlab="Year",ylab="New AIDS cases",ylim=c(0,280))
lines(new.t+1980,exp(fv$fit))
lines(new.t+1980,exp(fv$fit+2*fv$se.fit),lty=2)
lines(new.t+1980,exp(fv$fit-2*fv$se.fit),lty=2)
```

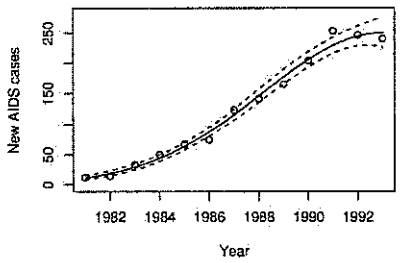

Figure 2.13 *Underlying AIDS case rate according to model* `m1` *shown as a continuous curve with 95% confidence limits shown as dashed curves.*

The plot is shown in figure 2.13. Notice that by default the `predict.glm` function predicts on the scale of the linear predictor: we have to apply the inverse of the link function to get back onto the original response scale.

So the data provide quite firm evidence to suggest that the unfettered exponential increase model is overly pessimistic: by the end of the data there is good evidence that the rate of increase is slowing. Of course this model contains no mechanistic content — it says nothing about how or why the slowing might be occurring: as such it is entirely in-appropriate for prediction beyond the range of the data. The model allows us to be reasonably confident that the apparent slowing in the rate of increase in new cases is real, and not just the result of chance variation, but it says little or nothing about what may happen later.

2.3.3 Log-linear models for categorical data

The following table classifies a sample of women and men according to their belief in the afterlife:

	Believer	Non-Believer
Female	435	147
Male	375	134

The data (reported in Agresti, 1996) come from the US General Social Survey (1991), and the 'non-believer' category includes 'undecideds'. Are there differences between males and females in the holding of this belief? We can address this question by using analysis of deviance to compare the fit of 2 competing models of these data: one in which belief is modelled as independent of gender, and a second in which there is some interaction between belief and gender. First consider the model of indepen-

1. (a) [3 marks] From the normal equations, $\widehat{\boldsymbol{\beta}}$ satisfies

$$X^T X\widehat{\beta} = X^T \boldsymbol{y},$$

So now, since X has full rank,

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

By standard properties of covariance,

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= \mathrm{Var}((X^T X)^{-1} X^T \boldsymbol{y}) = ((X^T X)^{-1} X^T)\mathrm{Var}(\boldsymbol{y})((X^T X)^{-1} X^T)^T \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} (X^T X)(X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}\end{aligned}$$

[Seen]

(b) [4 marks]

$$X^T X = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 6 & -2 \\ 0 & -2 & 4 \end{bmatrix}.$$

Since this matrix is block diagonal, the inverse is straightforwardly computed to be

$$(X^T X)^{-1} = \begin{bmatrix} .2 & 0 & 0 \\ 0 & .2 & .1 \\ 0 & .1 & .3 \end{bmatrix}.$$

Hence

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \begin{bmatrix} .2 & 0 & 0 \\ 0 & .2 & .1 \\ 0 & .1 & .3 \end{bmatrix}.$$

[Seen Method]

(c) [3 marks] Need $\mathrm{Var}(\boldsymbol{c}^T\widehat{\boldsymbol{\beta}})$, for $c = (1, 1, 1)$. This can be determined as follows:

$$c^T (X^T X)^{-1} = (0.2, 0.3, 0.4),$$

hence $\mathrm{Var}(\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}) = \sigma^2(0.2 + 0.3 + 0.4) = 0.9\sigma^2$. (Or just sum entries of covariance matrix). *[Seen Method]*

(d) [4 marks] From the covariance matrix, we see that estimates of β_1 and β_2 are uncorrelated and have the same variance. Hence a scatter plot will show circular symmetry, reflecting a bivariate normal distribution with identity covariance. When plotting estimates of β_2 against those of β_3, the shape will reveal bivariate normality but the two parameter estimates are now positively correlated, and $\widehat{\beta}_2$ has a smaller variance. Hence an ellipse would be apparent, whose principal axes would not be aligned to the coordinate axes. *[Seen Similar]*

(e) [6 marks]

* The first plot shows bowing away from the line of identity, suggesting a skewed error distribution. *[Seen Similar]*
* The second plot shows a symmetric sigmoid pattern, consistent with an error distribution with lighter tails than the Normal, but no skew. *[Seen Similar]*
* The third plot shows two split lines, suggestive of e.g. a mixture of normal distributions, perhaps evidence of an omitted covariate. *[Unseen]*

2. (a) [3 marks]

* The **random component** specifies the probability distribution of the response variables. Specifically, the components of $\boldsymbol{y}$ have pdf or pmf from an exponential family of distributions, with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$.
* The **systematic component** specifies a linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ as a function of the covariates and the unknown parameters.
* The **link function** g may be any monotonic differentiable function. The link function provides a functional relationship between the systematic component and the expectation of the response in the random component; namely $\boldsymbol{\eta} = g(\boldsymbol{\mu})$. *[Seen]*

(b) [2 marks] For this model, $\mathrm{E}(Y_i) = n_i p_i$. The log likelihood is given by

$$\sum_{i=1}^{n} \log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i)\log(1 - p_i).$$

Using $\mu_i = \mathrm{E}(Y_i) = n_i p_i$, we get

$$\sum_{i=1}^{n} \log \binom{n_i}{y_i} + y_i \log\left(\frac{\mu_i}{n_i}\right) + (n_i - y_i)\log\left(1 - \frac{\mu_i}{n_i}\right).$$

[Seen]

(c) [2 marks] The deviance is given by

$$\begin{aligned} D &= 2(l(\boldsymbol{y}, \boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}}, \boldsymbol{y})) \\ &= 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(1 - \frac{y_i}{n_i}\right) - y_i \log\left(\frac{\mu_i}{n_i}\right) - (n_i - y_i)\log\left(1 - \frac{\mu_i}{n_i}\right) \\ &= 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right). \end{aligned}$$

[Seen]

(d) [2 marks] The deviance residual is given by

$$r_i = \mathrm{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i},$$

where d_i is the contribution to the deviance for observation i, i.e. $D = \sum_{i=1}^{n} d_i^2$. Hence in this case

$$d_i = \mathrm{sign}\sqrt{2\left(y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right)\right)}.$$

[Seen Similar]

(e) [4 marks] For the binary case, $y_i \in \{0, 1\}$, so terms in $y \log y$ and $(1-y)\log(1-y)$ are zero, so that the absolute deviance residual takes a simpler form

$$\sqrt{2\left(-y_i \log \mu_i - (1 - y_i)\log(1 - \mu_i)\right)},$$

i.e. if $y_i = 1$

$$d_i = \sqrt{-2 \log \mu_i},$$

and if $y_i = 0$,

$$d_i = -\sqrt{-2 \log(1 - \mu_i)}.$$

Hence, in this case, the deviance residuals will lie on one of two smooth curves when plotted against the fitted value μ_i. *[Unseen]*

(f) [3 marks] Model shows that the log odds of developing the condition increase by around 0.98 for each kilo of birth weight. Hence odds of developing the disease increase by a multiplicative factor of $\exp(0.98) \approx 2.7$ for each kilo. Assuming the sample is large enough that asymptotic results apply, the coefficient is significantly different from zero, and an approximate 95% confidence interval for the log odds (working without a calculator) is $1 \pm 2 \times 0.4$, i.e (0.2, 1.8). *NB checked by simulation that confidence intervals from normal theory are reasonable in this case. [Seen Similar]*

(g) [2 marks] For binary logistic regression, deviance depends on the data only through the $\widehat{\beta}$. Its distribution conditional on the $\widehat{\beta}$ is degenerate, and certainly not independent of $\widehat{\beta}$. Since a goodness of fit statistic is expected to be independent of the fixed effects, its use is not appropriate here. Asymptotic theory does not hold. *[Seen Similar]*

(h) [2 marks] With the logistic link function, bias due to a different sampling mechanism cancels in the computation of the log odds, so the estimate of birth weight would be unchanged. The estimate of the intercept, however, would change, since this reflects the sampled proportions. *[Seen Similar]*

3. (a) [4 marks] Want to show Gamma pdf belongs to exponential family

$$\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

For the gamma distribution:

$$\begin{aligned} f(y;\mu,\nu) &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}e^{-\nu y/\mu} \\ &= \exp\left\{\frac{y(-1/\mu) - \log(\mu)}{1/\nu} + (\nu-1)\log(y) + \nu\log(\nu) - \log\Gamma(\nu)\right\} \end{aligned}$$

We then identify $\theta = -1/\mu$ and $a(\phi) = 1/\nu$. Take $\phi = 1/\nu$ then

$$\begin{aligned} b(\theta) &= \log(\mu) = -\log(-\theta) \\ c(y,\phi) &= (\phi^{-1}-1)\log(y) - \phi^{-1}\log(\phi) - \log\Gamma(\phi^{-1}). \end{aligned}$$

[Seen]

(b) [2 marks] By inspecting the exponential family representation, the canonical link function is: $g(\mu) = -1/\mu$.
Given in the question that $y = \frac{\alpha_0 x}{1+\alpha_1 x}$, hence

$$\frac{-1}{y} = \frac{1+\alpha_1 x}{\alpha_0 x} = \frac{\alpha_1}{\alpha_0} + \frac{1}{\alpha_0 x},$$

hence $\beta_0 = \frac{\alpha_1}{\alpha_0}$ and $\beta_1 = \frac{-1}{\alpha_0}$ (minus sign not important).
[Unseen]

(c) [3 marks] We first obtain the log likelihood of the saturated model. Hence we set the expected values μ_i to be the observations y_i.

$$l(\boldsymbol{y},\phi,\boldsymbol{y}) = \sum_{i=1}^{n}\left[\frac{y_i(-1/y_i) - \log(y_i)}{\phi} + c(y_i,\phi)\right] = \sum_{i=1}^{n}\left[\frac{-1-\log(y_i)}{\phi} + c(y_i,\phi)\right]$$

For a model with expected values $\widehat{\mu}_i$, the log-likelihood is

$$l(\widehat{\boldsymbol{\mu}},\phi,\boldsymbol{y}) = \sum_{i=1}^{n}\left[\frac{y(-1/\widehat{\mu}_i) - \log(\widehat{\mu}_i)}{\phi} + c(y_i,\phi)\right]$$

Now substitute in to the definition of deviance to get

$$\begin{aligned} D &= 2\phi\{\ell(\boldsymbol{y},\phi,\boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}},\phi,\boldsymbol{y})\} \\ &= 2\phi\left\{\sum_{i=1}^{n}\left[\frac{-1-\log(y_i)}{\phi} + c(y_i,\phi) - \frac{y_i(-1/\widehat{\mu}_i) - \log(\widehat{\mu}_i)}{\phi} - c(y_i,\phi)\right]\right\} \\ &= 2\phi\left\{\sum_{i=1}^{n}\left[\frac{-\log(y_i/\widehat{\mu}_i) + (y-\widehat{\mu}_i)/\widehat{\mu}_i}{\phi}\right]\right\} \\ &= 2\sum_{i=1}^{n}\left[-\log\left(\frac{y_i}{\widehat{\mu}_i}\right) + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right] \end{aligned}$$

[Seen]

(d) [2 marks]

$$z_i = \widehat{\eta}_i + (y_i - \widehat{\mu}_i)\frac{\partial \eta}{\partial \mu} = \widehat{\eta}_i + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i^2}$$

$$\widetilde{w}_{ii} = \left(\frac{\partial \eta}{\partial \mu}\right)^{-2} \frac{1}{V(\mu)} = \widehat{\mu}_i^2$$

(e) [2 marks] e.g. use a simple moment estimator: perform linear regression of $1/y$ against $1/x$. *[Seen Similar]*

(f) [2 marks] Use Pearson's chi-square statistic,

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

Asymptotically, $\frac{X^2}{\phi}$ has an approximate chi square distribution with $n - p$ degrees of freedom. A moment estimator of the dispersion is therefore $\frac{X^2}{n-p}$. *[Seen Similar]*

(g) (i) [2 marks] Asymptotically, we would expect $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \phi(X^T\widetilde{W}X)^{-1}$. (Strictly, it is approximately t-distributed since we have estimated ϕ, but given the large sample assumptions, the difference should be small.) Hence

$$\widehat{\beta}_2 \pm 2\widehat{\phi}((X^T\widetilde{W}X)^{-1})_{22}$$

is an approximate 95% confidence interval.

(ii) [1 mark] Apply the inverse of the link function to obtain a confidence interval on the response scale.

(h) [2 marks] Under the null hypothesis we have (approximately, for large samples, if the dispersion parameter is small) that the scaled deviances $D^* = D/\phi$ for the two models are χ^2:

$$D_1^* \sim \chi^2_{n-3} \quad \text{and} \quad D_0^* - D_1^* \sim \chi^2_1.$$

If we consider D_1^* and $D_0^* - D_1^*$ as asymptotically independent, then

$$\frac{(D_0^* - D_1^*)}{D_1^*/(n-3)} \sim F(1, n-3).$$

We can multiply top and bottom by ϕ to get a test statistic based on the deviance:

$$\frac{(D_0 - D_1)}{D_1/(n-3)} \sim F(1, n-3).$$

[NB checked by simulation that F-distribution result roughly holds here, for small dispersion parameter and large sample size.] [Seen Similar]

4. (a) (i) [2 marks] There are five different blends, hence can estimate four linearly independent contrasts; the first model already includes an intercept. *[Seen Similar]*

(ii) [3 marks] This is a test of the null hypothesis that there are no differences in mean yield for different blends. The test statistic is

$$\frac{RSS_0 - RSS_1}{RSS}\frac{20-12}{4} = \frac{2(RSS_0 - RSS_1)}{RSS}.$$

where RSS_0 is the RSS for the null model and RSS_1 for the larger alternative. Under the null hypothesis, this statistic has an $F(4, 12)$ distribution.
[Seen Method]

(iii) [2 marks] The test statistic is inside the critical region for a test at the 5% level of significance. A chance improvement would be unlikely to have decreased the RSS by so much. Hence we reject the null hypothesis that differences in the yield for different blends are due to chance.

(iv) [2 marks] This is a saturated model. There are as many observations as parameters. Hence, the model fits the data perfectly, and all residuals are zero. There are therefore no remaining degrees of freedom with which to assess goodness of fit. *[Seen Method]*

(b) *dataset seen, but not with random effects models*

(i) [2 marks] We seek to model the variability due to differences in blend - not really interested in the effect of any given blend. If fixed effects were used, this would be treating blends as individual entities with nothing in common. Instead, a random effects model posits a Normal distribution of blend effects, from which we have sampled five. This allows our analysis to extend to other samples from the population. *[Seen Similar]*

(ii) [2 marks] As a linear combination of multivariate normal variables, $\boldsymbol{Y}$ must be multivariate normal. It suffices to compute its mean vector and variance-covariance matrix.

$$\begin{aligned}
\mathrm{E}(\boldsymbol{Y}) &= \mathrm{E}(X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}) \\
&= X\boldsymbol{\beta} + Z\underbrace{\mathrm{E}(\boldsymbol{\nu})}_{=0} + \underbrace{\mathrm{E}(\boldsymbol{\epsilon})}_{=0} \\
&= X\boldsymbol{\beta}
\end{aligned}$$

And for the variance-covariance matrix,

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Y}) &= \mathrm{Cov}(X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon}) \\
&= Z\mathrm{Cov}(\boldsymbol{\nu})Z^T + \mathrm{Cov}(\boldsymbol{\epsilon}) \quad \text{(since } \boldsymbol{\epsilon}, \boldsymbol{\nu} \text{ indep.)} \\
&= ZI_m\sigma_\nu^2 Z^T + I_n\sigma_\epsilon^2 \\
&= \sigma_\epsilon^2\left(ZI_m\frac{\sigma_\nu^2}{\sigma_\epsilon^2}Z^T + I_n\right) \\
&= \sigma_\epsilon^2\left(I_n + Z\Psi Z^T + I_m\right),
\end{aligned}$$

where $\Psi = \frac{\sigma_\nu^2}{\sigma_\epsilon^2}I_n$. *[Seen]*

(iii) [2 marks] The covariance is

$$\begin{aligned}
\mathrm{E}\left[\left(Y_{1,j} - \mathrm{E}(Y_{1,j})\right)\left(Y_{2,j} - \mathrm{E}(Y_{2,j})\right)\right] &= \mathrm{E}\left[\left(\nu_j + \epsilon_{1,j}\right)\right)\left(\nu_j + \epsilon_{2,j}\right))\right] \\
&= \mathrm{E}(\nu_j^2) + \mathrm{E}(\nu_j(\epsilon_{1,j} + \epsilon_{2,j})) + \mathrm{E}(\epsilon_{1,j}\epsilon_{2,j}) \\
&= \mathrm{E}(\nu_j^2) = \sigma_\nu^2,
\end{aligned}$$

where we used the independence between the ν_j and ϵ_{ij}. *[Seen]*

(iv) [2 marks] To perform restricted maximum likelihood, we apply a linear transformation L to the response $\boldsymbol{Y}$ that projects the data onto the space orthogonal to that spanned by the columns of the design matrix. $L\boldsymbol{Y}$ is then independent of the fixed effects β, and we can maximize the resulting restricted likelihood for the parameters $(\sigma^2_\nu, \sigma^2_\epsilon)$, obtainining unbiased estimates of these variances. These estimates can then be plugged in to the likelihood, which can then be maximized over the fixed effects β.

(v) [2 marks] Data are balanced, so REML estimates are equal to the ANOVA estimates. Hence using `fit1`, residual variance is $\frac{226}{12}$. *[Unseen]*

(vi) [1 mark] Balanced design with treatment contrasts, so same variance for each treatment. Hence the standard error is 2.745, as for B and C. *[Unseen]*

5. (a) [3 marks] There are 12 observations, and 2 parameters have been estimated. Assuming asymptotic theory is applicable, the residual deviance should have an approximate chi square distribution with 10 degrees of freedom.

   (b) [2 marks] Observed information is equal to expected information (so always positive definite); simple interpretation in terms of log odds; model interpretable with distorted sampling proportions (retrospective vs prospective). [Any 2]

   (c) [1 mark] Suggests that some unmodelled structure remains in the residuals, so the systematic part of the model is not correct.

   (d) [2 marks] It is a smoother, producing a local average of nearby residuals at each point. It helps to guide the eye in discerning trends in the residuals.

   (e) [1 mark] The log link is equivalent to assuming

$$\mu_i = c\exp(bt_i),$$

   i.e. exponential increase in the mean number of infected individuals, for $b, c > 0$.

   (f) [3 marks] The model manifold (i.e the set of all possible fitted value vectors the model could predict) is curved, and not a flat plane. Moreover, lines of equal fit do not intersect the model manifold orthogonally. Iterated reweighted least squares amounts to a succession of translations and rescalings, one at each iteration, such that at the current best estimate, the model manifold intersects the lines of equal fit roughly orthogonally. Locally, then, the problem is reduced to one of orthogonal projection, and so can be solved efficiently by least squares.

   (g) [3 marks] The MLE is defined by setting the score function (derivative of the log likelihood) to zero. For a general member of the exponential family, this has the form

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0.$$

   For constant $\boldsymbol{\beta}$, this defines a linear system for $\boldsymbol{y}$.

   (g) [3 marks] The x-values of the points given are $(0.6, 1.5)$. It follows that $(\mu_1, \mu_2) = (20\exp(-0.6\beta), 20\exp(-1.5\beta))$. Hence $\mu_2 = 20\exp(-0.9\beta)\mu_1$.

   (h) [2 marks] Looking carefully at the plot, the given point $(y_1, y_2) = (5, 9)$ corresponds to the higher unbroken line, which is given in the text as $\beta = 0.7$.