16

Binary Response Variable

Many statistical problems involve binary response variables. For example, we often classify individuals as

- dead or alive,
- occupied or empty,
- healthy or diseased,
- wilted or turgid,
- male or female,
- literate or illiterate,
- mature or immature,
- solvent or insolvent, or
- employed or unemployed.

It is interesting to understand the factors that are associated with an individual being in one class or the other. In a study of company insolvency, for instance, the data would consist of a list of measurements made on the insolvent companies (their age, size, turnover, location, management experience, workforce training and so on) and a similar list for the solvent companies. The question then becomes which, if any, of the explanatory variables increase the probability of an individual company being insolvent?

The response variable contains only 0s or 1s; for example, 0 to represent dead individuals and 1 to represent live ones. Thus, there is only a single column of numbers for the response, in contrast to proportion data where two vectors (successes and failures) were bound together to form the response (see Chapter 14). The way that R treats binary data is to assume that the 0s and 1s come from **a binomial trial with sample size 1**. If the probability that an individual is dead is p, then the probability of obtaining y (where y is

Statistics: An Introduction using R M. J. Crawley
 ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

either dead or alive, 0 or 1) is given by an abbreviated form of the binomial distribution with $n = 1$, known as the Bernoulli distribution:

$$P(y) = p^y(1-p)^{(1-y)}.$$

The random variable y has a mean of p and a variance of $p(1-p)$, and the objective is to determine how the explanatory variables influence the value of p. The trick for using binary response variables effectively is to know when it is worth using them, and when it is better to lump the successes and failures together and analyse the **total counts** of dead individuals, occupied patches, insolvent firms or whatever. The question you need to ask yourself is: **do I have unique values of one or more explanatory variables for each and every individual case?**

If the answer is 'yes', then analysis with a binary response variable is likely to be fruitful. If the answer is 'no', then there is nothing to be gained, and you should reduce your data by aggregating the counts to the resolution at which each count **does** have a unique set of explanatory variables. For example, suppose that all your explanatory variables were categorical such as gender (male or female), employment (employed or unemployed) and region (urban or rural). In this case there is nothing to be gained from analysis using a binary response variable because none of the individuals in the study have **unique** values of any of the explanatory variables. It might be worthwhile if you had each individual's body weight, for example, then you could ask the question 'when I control for gender and region, are heavy people more likely to be unemployed than light people?' In the absence of **unique** values for any explanatory variables, there are two useful options.

- Analyse the data as a contingency table using Poisson errors, with the count of the total number of individuals in each of the eight contingencies (2 x 2 x 2) as the response variable (see Chapter 13) in a dataframe with just eight rows.
- Decide which of your explanatory variables is the key (perhaps you are interested in gender differences), then express the data as proportions (the number of males and the number of females) and re-code the binary response as a count of a two-level factor. The analysis is now of proportion data (the proportion of all individuals that are female, for instance) using binomial errors (see Chapter 14).

If you **do** have unique measurements of one or more explanatory variables for each individual, these are likely to be continuous variables such as body weight, income, medical history, distance to the nuclear reprocessing plant, geographic isolation and so on. This being the case, successful analyses of binary response data tend to be multiple regression analyses or complex analyses of covariance, and you should consult Chapters 10 and 11 for details on model simplification and model criticism.

In order to carry out modelling on a binary response variable we take the following steps:

- create a single vector containing 0s and 1s as the response variable,
- use `glm` with `family = binomial`,
- you can change the link function from default logit to complementary log–log,

- fit the model in the usual way,
- test significance by deletion of terms from the maximal model, and compare the change in deviance with chi-square,
- note that there is no such thing as overdispersion with a binary response variable, and hence no need to change to using quasibinomial when the residual deviance is large.

Choice of link function is generally made by trying both links and selecting the link that gives the lowest deviance. The logit link that we used earlier is symmetric in p and q, but the complementary log–log link is asymmetric.

Incidence Functions

In this example, the response variable is called 'incidence'; a value of 1 means that an island was occupied by a particular species of bird, and 0 means that the bird did not breed there. The explanatory variables are the area of the island (km^2) and the isolation of the island (distance from the mainland, km).

```
island <-read.table("c:\\temp\\isolation.txt",header=T)
attach(island)
names(island)
```

```
[1] "incidence"  "area"  "isolation"
```

There are two continuous explanatory variables, so the appropriate analysis is multiple regression. The response is binary, so we shall do logistic regression with binomial errors. We begin by fitting a complex model involving an interaction between isolation and area:

```
model1 <-glm(incidence~area*isolation,binomial)
```

then fit a simpler model with only main effects for isolation and area:

```
model2 <-glm(incidence~area+isolation,binomial)
```

then compare the two models using Anova:

```
anova(model1,model2,test="Chi")
```

```
Analysis of Deviance Table

Model 1: incidence~area * isolation
Model 2: incidence~area + isolation
Resid.  Df  Resid. Dev  Df  Deviance  P(>|Chi|)
1       46    28.2517
2       47    28.4022  -1   -0.1504     0.6981
```

The simpler model is not significantly worse, so we accept this for the time being, and inspect the parameter estimates and standard errors:

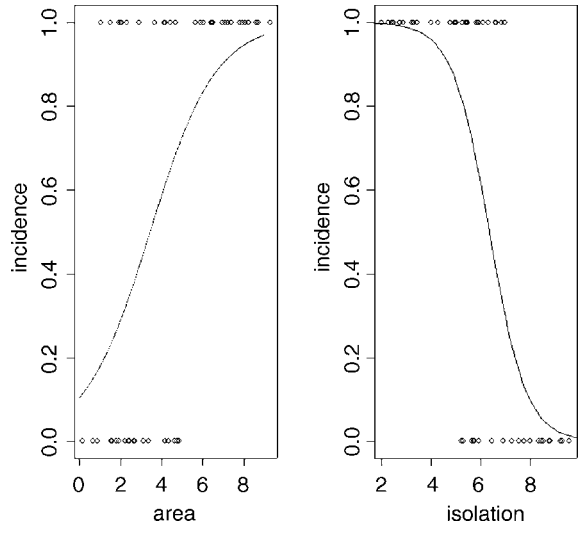
```
summary(model2)

Call:
glm(formula = incidence~area + isolation, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8189  -0.3089   0.0490   0.3635   2.1192

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    6.6417      2.9218    2.273   0.02302  *
area           0.5807      0.2478    2.344   0.01909  *
isolation     -1.3719      0.4769   -2.877   0.00401 **

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029 on 49 degrees of freedom
Residual deviance: 28.402 on 47 degrees of freedom
```

The estimates and their standard errors are in logits. Area has a significant positive effect (larger islands are more likely to be occupied), but isolation has a very strong negative effect (isolated islands are much less likely to be occupied). This is the minimal adequate model. We should plot the fitted model through the scatterplot of the data. It is much easier to do this for each variable separately, like this:

```
modela <-glm(incidence~area,binomial)
modeli <-glm(incidence~isolation,binomial)
par(mfrow=c(1,2))
xv <-seq(0,9,0.01)
yv <-predict(modela,list(area=xv),type="response")
plot(area,incidence)
lines(xv,yv)
xv2 <-seq(0,10,0.1)
yv2 <-predict(modeli,list(isolation=xv2),type="response")
plot(isolation,incidence)
lines(xv2,yv2)
```

This is all well and good, but it is very difficult to know how good the fit of the model is when the data are shown only as zeros or ones. It is sensible to compute one or more intermediate probabilities from the data, and to show these empirical estimates (ideally with their standard errors) on the plot in order to judge whether the fitted line is a reasonable description of the data.

For the purposes of demonstration, we take the central third of the data ranked by area and by isolation, calculate the mean proportion incidence, p, and add this to the plot



along with its standard error $\sqrt{p(1-p)/n}$. We use cut to obtain the central third of the data:

```
ac <-cut(area,3)
ic <-cut(isolation,3)
tapply(incidence,ac,sum)
```

```
(0.144,3.19] (3.19,6.23]    (6.23,9.28]
           7            8             14
```

```
tapply(incidence,ic,sum)
```

```
(2.02,4.54]  (4.54,7.06]    (7.06,9.58]
         12           17              0
```

Note the convention for labelling intervals: `(a, b]` means include their left-hand endpoint, a, but not their right-hand one, b. Now count the number of cases in each interval using table

```
table(ac)
ac
```

```
(0.144,3.19] (3.19,6.23]    (6.23,9.28]
          21           15             14
```

```
table(ic)
ic
```

```
(2.02,4.54]  (4.54,7.06]    (7.06,9.58]
         12           25             13
```

A sensible place to plot the mean probability associated with the central third of the explanatory variable is in the position defined by the median:

```
median(area)
```

```
[1] 4.1705
```

```
median(isolation)
```

```
[1] 5.8015
```

Next, calculate the two mean proportions:

```
8/15
```

```
[1] 0.5333333
```

```
17/25
```

```
[1] 0.68
```

and their two standard errors:

```
sqrt((8/15*7/15)/15)
```

```
[1] 0.1288122
```

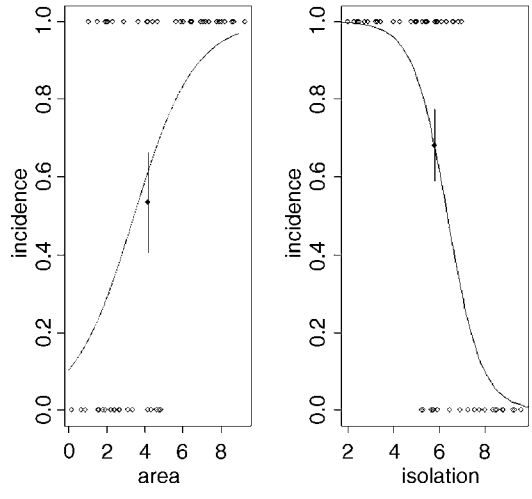
```
sqrt((17/25*8/25)/25)
```

```
[1] 0.09329523
```

Finally, re-plot the two graphs, adding the empirical estimates and their error bars:

```
plot(area,incidence)
lines(xv,yv)
points(4.1705,0.5333,pch=16)
lines(c(4.1705,4.1705),c(0.533333-0.1288,0.533333+0.1288))
plot(isolation,incidence)
lines(xv2,yv2)
points(5.8015,0.68,pch=16)
lines(c(5.8015,5.8015),c(0.68-0.093,0.68+0.093))
```

This shows that the fit to the central third of the data is excellent for the relationship between incidence and isolation, but less good (although not significantly far out) for the relationship with area. With a large data set, you can compute more of these empirical estimates (say three, five, or seven of them) and this would enable you to test quite sensitively for model failure. This approach would not work, of course, if there was an interaction between area and isolation; then you would need to produce conditioning plots of incidence against area for different degrees of isolation.



Ancova with a Binary Response Variable

In this example the binary response variable is parasite infection (infected or not) and the explanatory variables are weight and age (continuous) and gender (categorical). We begin with data inspection:

```
infection<-read.table("c:\\temp\\infection.txt",header=T)
attach(infection)
names(infection)
```

```
[1] "infected"  "age"  "weight"  "gender"
```

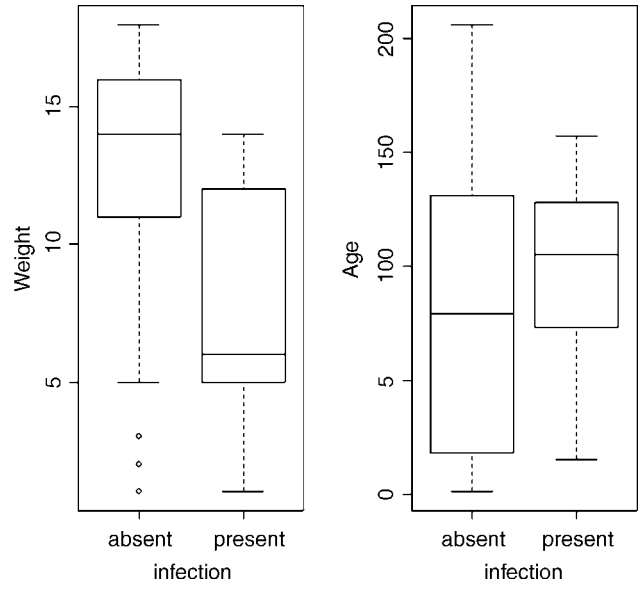
```
par(mfrow=c(1,2))
plot(infected,weight,xlab="Infection",ylab="Weight")
plot(infected,age,xlab="Infection",ylab="Age")
```

Infected individuals are substantially lighter than uninfected individuals and occur in a much narrower range of ages. To see the relationship between infection and gender (both categorical variables) we can use table:

```
table(infected,gender)
```

```
table(infected,gender)
                    gender
infected            female      male
   absent               17        47
  present               11         6
```

which indicates that the infection is much more prevalent in females (11/28) than in males (6/53).



We begin, as usual, by fitting a maximal model with different slopes for each level of the categorical variable:

```
model<-glm(infected~age*weight*gender,family=binomial)
summary(model)
```

```
Coefficients:
                        Estimate    Std. Error    z value   Pr(>|z|)
(Intercept)           -0.109124      1.375388    -0.079      0.937
age                    0.024128      0.020874     1.156      0.248
weight                -0.074156      0.147678    -0.502      0.616
gendermale            -5.969133      4.275952    -1.396      0.163
age:weight            -0.001977      0.002006    -0.985      0.325
age:gendermale         0.038086      0.041310     0.922      0.357
weight:gendermale      0.213835      0.342825     0.624      0.533
age:weight:gendermale -0.001651      0.003417    -0.483      0.629

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234 on 80 degrees of freedom
Residual deviance: 55.706 on 73 degrees of freedom
AIC: 71.706
```

It certainly does not look as if any of the high-order interactions are significant. Instead of using update and anova for model simplification, we can use step to compute the AIC for each term in turn (see p. 208).

```
model2<-step(model)
```

```
Start: AIC= 71.71
```

First, it tests whether the three-way interaction is required

```
                            Df   Deviance      AIC
- age:weight:gender          1     55.943   69.943
⟨none⟩                             55.706   71.706

Step: AIC= 69.94
```

This causes a reduction in AIC of just $71.7 - 69.9 = 1.8$ and hence is not significant. Next, it looks at the three two-way interactions and decides which to delete first:

```
                  Df  Deviance     AIC
- weight:gender    1    56.122  68.122
- age:gender       1    57.828  69.828
⟨none⟩                  55.943  69.943
- age:weight       1    58.674  70.674

Step: AIC= 68.12
```

Only the removal of weight: gender causes a substantial reduction in AIC, so this interaction in deleted and the other two interactions are retained. Let's see if we would have been this lenient:

summary(model2)

```
Call:
glm(formula = infected~age + weight + gender + age:weight + age:gender,
    family = binomial)

Coefficients:
                 Estimate  Std. Error  z value    Pr(>|z|)
(Intercept)     -0.391572   1.264850   -0.310      0.7569
age              0.025764   0.014918    1.727      0.0842   .
weight          -0.036493   0.128907   -0.283      0.7771
gendermale      -3.743698   1.786011   -2.096      0.0361   *
age:weight      -0.002221   0.001365   -1.627      0.1037
age:gendermale   0.020464   0.015199    1.346      0.1782

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234 on 80 degrees of freedom
Residual deviance: 56.122 on 75 degrees of freedom
AIC: 68.122
```

Neither of the two interactions retained by step would figure in our model ($p > 0.10$). We shall use update to simplify model2:

```
model3 <-update(model2, ~.-age:weight)
anova(model2,model3,test="Chi")
```

```
Analysis of Deviance Table

Model 1: infected~age + weight + gender + age:weight +age: gender
Model 2: infected~age + weight + gender + age:gender
Resid.  Df       Resid. Dev        Df     Deviance    P(>|Chi|)
1       75          56.122
2       76          58.899        -1       -2.777        0.096
```

so there is no really persuasive evidence of an age:weight term ($p = 0.096$)

```
model4 <-update(model2, ~.-age:gender)
anova(model2,model4,test="Chi")
```

Note that we are testing all the two-way interactions by deletion from the model that contains all two-way interactions (model 2): $p = 0.155$, so nothing there, then. What about the three main effects?

```
model5 <-glm(infected ~age + weight + gender,family = binomial)
summary(model5)
```

```
Coefficients:
              Estimate   Std. Error   z value    Pr(>|z|)
(Intercept)   0.609392     0.801303     0.761    0.446955
age           0.012649     0.006717     1.883    0.059654     .
weight       -0.227880     0.068138    -3.344    0.000825   ***
gendermale   -1.543151     0.681434    -2.265    0.023539     *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234 on 80 degrees of freedom
Residual deviance: 59.859 on 77 degrees of freedom
AIC: 67.859
```

Weight is highly significant, as we expected from the initial boxplot, gender is quite significant, and age is marginally significant. It is worth establishing whether there is any evidence of non-linearity in the response of infection to weight or age. We might begin by fitting quadratic terms for the two continuous explanatory variables:

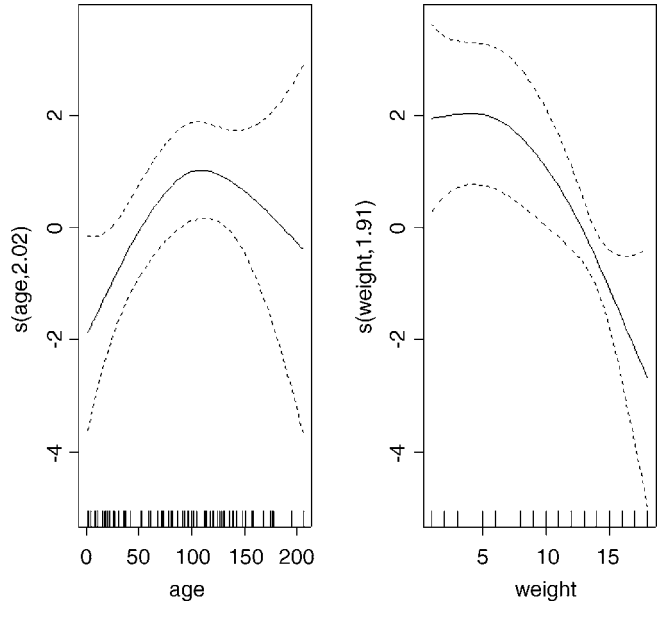
```
model6 <-
glm(infected ~age + weight + gender + I(weight^2) + I(age^2), family = binomial)
summary(model6)
```

```
Coefficients:
                Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)   -3.4469474    1.7825435   -1.934    0.0531   .
age            0.0829206    0.0355997    2.329    0.0198   *
weight         0.4465758    0.3355612    1.331    0.1832
```

```
gender          -1.2202485   0.7646071   -1.596     0.1105
I(weight^2)     -0.0415082   0.0208383   -1.992     0.0464   *
I(age^2)        -0.0004008   0.0001981   -2.023     0.0431   *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234 on 80 degrees of freedom
Residual deviance: 48.620 on 75 degrees of freedom
AIC: 60.62
```

Evidently, both relationships are significantly non-linear. It is worth looking at these non-linearities in more detail, to see if we can do better with other kinds of models (e.g. non-parametric smoothers, piece-wise linear models or step functions). A good start is often a gam (a generalized additive model) when we have continuous covariates:

```
library(mgcv)
model7 <-gam(infected ~ gender + s(age) + s(weight),family = binomial)
plot.gam(model7)
```



These non-parametric smoothers are excellent at showing the humped relationship between infection and age, and at highlighting the possibility of a threshold at weight ≈ 8 in the relationship between weight and infection. We can now return to a glm to incorporate these ideas. We shall fit **age** and **age**^2 as before, but try a piecewise linear fit for **weight**, estimating the threshold weight at a range of values (say 8–14) and selecting the threshold that gives the lowest residual deviance; this turns out to be a threshold = 12. The piecewise regression is specified by the term:

```
I((weight - 12) * (weight > 12))
```

The I ('as is') is necessary to stop the * as being evaluated as an interaction term in the model formula. What this expression says is 'regress infection on the value of weight − 12, but only do this when (weight > 12) is true'. Otherwise, assume that infection is independent of weight.

```
model8 < -glm(infected ~ sex + age + I(age^2) + I((weight–12)* (weight > 12)),
   family = binomial)
summary(model8)
```

```
Coefficients:
                                Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)                   -2.7511452   1.3672006   -2.012    0.0442   *
gender male                   -1.2864620   0.7343309   -1.752    0.0798   .
age                            0.0798630   0.0347926    2.295    0.0217   *
I(age^2)                      -0.0003892   0.0001953   -1.993    0.0463   *
I((weight - 12) * (weight > 12))  -1.3547080   0.5318043   -2.547    0.0109   *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234 on 80 degrees of freedom
Residual deviance: 48.687 on 76 degrees of freedom
AIC: 58.687
```

```
model9 < -update(model8, ~ .-gender)
anova(model8,model9,test = "Chi")
model10 < -update(model8, ~ .-I(age^2))
anova(model8,model10,test = "Chi")
```

The effect of gender on infection is not quite significant ($p = 0.071$ for a Chi-square test on deletion), so we leave it out. The quadratic term for age does not look highly significant here, but a deletion test gives $p = 0.011$, so we retain it. The minimal adequate model is therefore model9:

```
summary(model9)
```

```
Coefficients:
                                Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)                   -3.1207575   1.2663090   -2.464   0.01372   *
age                            0.0765785   0.0323274    2.369   0.01784   *
I(age^2)                      -0.0003843   0.0001845   -2.082   0.03732   *
I((weight - 12) * (weight > 12))  -1.3511514   0.5112930   -2.643   0.00823  **

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234 on 80 degrees of freedom
Residual deviance: 51.953 on 77 degrees of freedom
AIC: 59.953
```

We conclude there is a humped relationship between infection and age, and a threshold effect of weight on infection. The effect of gender is marginal, but might repay further investigation ($p = 0.071$).