Part II

Binary Data

Binary Data

Binary data may occur in two forms

- **ungrouped** in which the variable can take one of two values, say success/failure
- **grouped** in which the variable is the number of successes in a given number of trials

The natural distribution for such data is the Binomial(n, p) distribution, where in the first case $n = 1$

Exploring Binary Data

If our aim is to model a binary response, we would first like to explore the relationship between that response and potential explanatory variables.

When the explanatory variables are categorical, a simple approach is to calculate proportions within subgroups of the data.

When some of the explantory variables are continuous, plots can be more helpful.

Example: Bronchitis Data

Jones (1975) conducted a study of chronic bronchitis in Cardiff. The variables are

- `cigs` the number of cigarettes smoked per day
- `poll` the air pollution level in the locality of residence
- `bron` the presence/absence of bronchitis (indicated by 1/0)

Scatterplots for Binary Data

We can plot cigs against poll using the bronchitis values as labels:

```
plot(poll ~ cigs, xlab = "No. cigarettes/day",
     ylab = "Pollution level", type = "n")
text(cigs, poll, labels = bron)
legend(20, 65, legend = c("presence", "absence"),
       title = "Bronchitis",
       pch = c("1", "0"))
```

However the pattern of bronchitis cases is not that clear. We can get some idea of the relationships by considering cigs and poll separately.

Scatterplots of a binary response are not that helpful, but can be improved by adding jitter:

```
plot(bron ~ poll)
plot(jitter(bron, 0.1) ~ poll)
plot(bron ~ cigs)
plot(jitter(bron, 0.1) ~ cigs)
```

These plots suggest that risk of bronchitis increases with both `poll` and `cigs`, with `cigs` seeming to have the bigger effect.

Boxplots

An alternative to scatterplots is to use boxplots

```
boxplot(poll ~ bron,
    xlab = "Bronchitis presence/absence (1/0)",
    ylab = "Pollution level")
boxplot(cigs ~ bron,
    xlab = "Bronchitis presence/absence (1/0)",
    ylab = "No. cigarettes/day")
```

which can be easier to interpret.

Example: Budworm Data

Collett(1991) describes an experiment on the toxicity of the pyrethoid *trans - cypermethrin* to the tobacco budworm. Batches of 20 moths of each sex were exposed to varying doses of the pyrethoid for three days and the number knocked out in each batch was recorded:

	Dose (μ g)					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Since the doses are in powers of two, we will use $\log_2(\texttt{dose})$ as the response.

Scatterplots of Binomial Data

For grouped binary data, scatterplots are more helpful:

```
ldose <- rep(0:5, 2)
dead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))

plot(c(1, 32), c(0, 1), type = "n", xlab = "dose",
     ylab = "prob", log = "x")
text(2^ldose, dead/20, labels = sex)
```

Models for Binary Data

In Part I we saw that Binomial data may be modelled by a glm, with the canonical logit link. This model is known as the **logistic regression** model and is the most popular for binary data.

There are two other links commonly used in practice:

- **probit link** $g(\mu_i) = \Phi^{-1}(\mu_i)$ where Φ denotes the cumulative distribution function of N(0, 1)
- **complementary log-log link** $g(\mu_i) = \log(-\log(1 - \mu_i))$

Comparison of Links

The three links map the linear predictor η to the probability scale as follows:

```
mu.logit <- function(eta) 1/(1 + exp(-eta))
mu.probit <- function(eta) pnorm(eta, 0, pi/sqrt(3))
mu.cloglog <- function(eta) 1 - exp(-exp(eta))

plot(mu.logit, (-4): 4, xlim = c(-4, 4), ylim = c(0,1),
     xlab = expression(eta),
     ylab = expression(mu == g^-1 * (eta)))
curve(mu.probit, (-4):4, add = TRUE, lty = 2)
curve(mu.cloglog, (-4):4, add = TRUE, lty = 3)
legend(-4, 1, c("logit", "probit", "complementary log-log"),
       lty = 1:3)
```

Choice of Link

The logit and probit functions are symmetric and - once their variances are equated - are very similar. Therefore it is usually difficult to choose between them on the grounds of fit.

The logit is usually preferred over the probit because of its simple interpretation as the logarithm of the odds of success $(p_i/(1-p_i))$.

The complementary log-log is asymmetric and may therefore be useful when the logit and probit links are inappropriate.

We will concentrate on using the logit link.

Scatterplot Scales

When fitting a logistic model, it can also be helpful to plot the data on the logit scale.

To avoid dividing by zero, we calculate the **empirical logits**

$$\log\left(\frac{(y_i + 0.5)/n_i}{1 - (y_i + 0.5)/n_i}\right) = \log\left(\frac{y_i + 0.5}{n_i + 0.5 - y_i}\right)$$

E.g. for the budworm data

```
emp.logits <- log((dead + 0.5)/(20.5 - dead))
plot(c(1, 32), range(emp.logits), type = "n", xlab = "dose",
     ylab = "emp.logit", log = "x")
text(2^ldose, emp.logits, labels = sex)
```

Modelling the Budworm Data

A linear logistic model appears to be appropriate. A reasonable approach might be to consider the following linear predictors:

- single line for both sexes ($\sim$ `ldose`)
- parallel lines for each sex ($\sim$ `ldose + sex`)
- separate lines for each sex ($\sim$ `ldose + sex + ldose:sex`)

How can we determine which model is best?

Nested models

The candidate models for the budworm data are an example of **nested models** where each model is a special case of the models that have a greater number of terms.

We can compare nested models by testing the hypothesis that some of the parameters of a larger model are equal to zero.

For example suppose we have the model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

we can test

$$\begin{aligned} & H_0 : \beta_{q+1} = \ldots = \beta_p = 0, \\ \text{versus} \quad & H_1 : \beta_j \neq 0, \text{for some } j \in \{q+1, p\} \end{aligned}$$

using the likelihood ratio statistic

$$LR = 2(l_{big} - l_{small})$$

where l_m is the maximised log-likelihood under model m, i.e. $l(\hat{\boldsymbol{\beta}}_m)$.

Under the null hypothesis, LR is approximately χ^2_d where $d = p - q$.

Binomial Responses and glm

Now we would like to fit our candidate models. Binomial responses can be specified to glm in three ways:

- a numeric vector giving the *proportion* of successes y_i/n_i, in which case a vector of the prior weights n_i must be passed to the weights argument
- a numeric 0/1 vector (0 = failure); a logical vector (FALSE = failure), or a factor (first level = failure)
- a two-column matrix with the number of successes and the number of failures

Better starting values are generated when the third format is used.

Single Line Model

We can fit the single line logistic model as follows

```
y <- cbind(dead, 20 - dead)
sing <- glm(y ~ ldose, family = binomial)
summary(sing)
```

The logit link is the default for the binomial family so doesn't need to be specified.

As expected, the coefficient of `ldose` is highly significant.

Parallel Lines Model

Now we fit the parallel lines model:

```
parr <- glm(y ~ ldose + sex, family = binomial)
```

The Wald tests suggest both `ldose` and `dose` are needed in the model.

In general, the likelihood ratio tests are a better way of comparing models. We can use `anova` to perform this test:

```
anova(sing, parr, test = "Chisq")
```

Separate Lines Model

Finally we consider the separate lines model:

```
sep <- glm(y ~ sex*ldose, family = binomial)
```

Using `anova` will test sequential addition of terms in this model:

```
anova(sep, test = "Chisq")
```

Allowing separate slopes does not significantly reduce the deviance.

Goodness-of-fit

Notice that the deviance

$$D = 2\phi(l_{sat} - l_{mod})$$

is ϕ times the likelihood ratio statistic comparing the fitted model to the saturated model.

Therefore the deviance can be used as goodness-of-fit statistic, tested against χ^2_{n-p}.

A good-fitting model will have

$$\frac{D}{\phi \approx d.f.}$$

For the budworm data, the parallel lines model has a deviance of 6.76 on 9 degrees of freedom, indicating that the model fits well.

Interpretation of Logistic Models

Consider the logistic model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i}$$

If we increase x_1 by one unit

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1(x_{1i}+1)\\ &= \beta_0 + \beta_1 x_{1i} + \beta_1\\ \Rightarrow \left(\frac{p_i}{1-p_i}\right) &= \exp(\beta_0 + \beta_1 x_{1i})\exp(\beta_1)\end{aligned}$$

the odds are multiplied by $\exp(\beta_1)$.

Interpretation of Budworm Model

For the budworm data, the parallel lines model is

$$\log\left(\frac{p_i}{1-p_i}\right) = -3.47 + 1.06\texttt{ldose}_\texttt{i} + \texttt{1.10}(\texttt{sex}_\texttt{i} == "\texttt{M}")$$

Therefore

- the odds of death for a male moth are $\exp(1.10) = 3.01$ times that for a female moth, given a fixed dose of the pyrethroid.
- the odds of death increase by a factor of $\exp(1.06) = 2.90$ for every log μg of pyrethroid, for male or female moths.

Wald Confidence Intervals

Confidence intervals for the parameters can be based on the asymptotic normal distribution for $\hat{\beta}_j$.

For example a 95% confidence interval would be given by

$$\hat{\beta}_j \pm 1.96 \times s.e.(\hat{\beta}_j)$$

Such confidence intervals can be obtained as follows:

```
confint.lm(parr)
```

Profiling the Deviance

The Wald confidence intervals used standard errors based on the second-order Taylor expansion of the log-likelihood at $\hat{\boldsymbol{\beta}}$.

An alternative approach is to the profile the log-likelihood, or equivalently the deviance, around each $\hat{\beta}_j$ and base confidence intervals on this.

We set β_j to $\tilde{\beta}_j \neq \hat{\beta}_j$ and re-fit the model to maximise the likelihood/minimise the deviance under this constraint. Repeating this for a range of values around $\hat{\beta}_j$ gives a deviance profile for that parameter.

Likelihood Ratio Test

To test the hypothesis

$$\begin{aligned} & H_0 : \beta_j = \tilde{\beta}_j, \\ \text{versus} \quad & H_1 : \beta_j = \hat{\beta}_j \end{aligned}$$

We can use the likelihood ratio statistic

$$2(l(\hat{\beta}_j) - l(\tilde{\beta}_j))$$

which is asymptotically distributed χ^2_1. Thus

$$\begin{aligned} \tau &= \mathsf{sign}(\tilde{\beta}_j - \hat{\beta}_j)\sqrt{(}2(l(\hat{\beta}_j) - l(\tilde{\beta}_j))) \\ &= \mathsf{sign}(\tilde{\beta}_j - \hat{\beta}_j)\sqrt{(}(D(\tilde{\beta}_j) - D(\hat{\beta}_j))/\phi) \end{aligned}$$

is asymptotically $N(0, 1)$ and is analogous to the Wald statistic.

Profile Plots

If the log-likelihood were quadratic about $\hat{\beta}_j$, then a plot of τ vs. $\tilde{\beta}_j$ would be a straight line.

We can obtain such a plot as follows

```
plot(profile(parr, "ldose"))
```

The approximation is not unreasonable, but there is slight curvature.

Profile Confidence Intervals

Rather than use the quadratic approximation, we can directly estimate the values of β_j for which $\tau = \pm 1.96$ to obtain a 95% confindence interval for β_j.

This is the method used by `confint.glm`:

```
confint(parr)
```

Notice the confidence intervals are asymmetric.

Prediction

The `predict` method for GLMs has a `type` argument, which may be specified as

- `"link"` for predictions of η
- `"reponse"` for predictions of μ

If no new data is passed to `predict`, these options return `object$linear.predictor` and `object$fitted.values` respectively.

We can use `predict` to plot our fitted model on the logit scale:

```
plot(c(1, 32), range(emp.logits), type = "n", xlab = "dose",
     ylab = "emp.logit", log = "x")
text(2^ldose, emp.logits, labels = sex)
lines(2^ldose[sex == "M"],
      predict(parr, type = "link")[sex == "M"], col = 3)
lines(2^ldose[sex == "M"],
      predict(parr, type = "link")[sex == "F"], col = 2)
```

Or on the probability scale - generating new data over smaller intervals to obtain a smoother curve

```
plot(c(1, 32), c(0, 1), type = "n", xlab = "dose",
     ylab = "prob", log = "x")
text(2^ldose, dead/20, labels = sex)
ld <- seq(0, 5, 0.1)
newdat <- data.frame(ldose = c(ld, ld),
                     sex = gl(2, length(ld),
                     labels = c("F", "M")))
lines(2^ld, predict(parr, type = "response",
                    newdat = subset(newdat, sex == "M")),
      col = 3)
lines(2^ld, predict(parr, type = "response",
                    newdat = subset(newdat, sex == "F")),
      col = 2)
```

Residual Analysis

The deviance residuals can be used to check the model as with Normal models.

The standardized residuals for binomial data should have an approximate normal distribution, provided the numbers for each covariate pattern is not too small.

```
par(mfrow = c(2, 2))
plot(parr, 1:4)
```

Modelling Bronchitis Data

We saw that `bron` was related to both `cigs` and `poll`.

A logistic regression with linear effects of both variables is a good place to start

```
model1 <- glm(bron ~ cigs + poll, family = binomial)
```

How can we evaluate this model?

Deviance for Binary Data

We have seen that the deviance may be viewed as a likelihood ratio statistic with approximate distribution χ^2_{n-p}.

However the χ^2 distribution of the likelihood ratio statistic is based on the limit as $n \rightarrow \infty$ with the number of parameters in the nested models both fixed. This does not apply to the deviance.

The χ^2_{n-p} distribution is still reasonable where the information content of each observation is large e.g. binomial models with large n_i, Poisson models with larger μ_i, gamma models with small ϕ.

For binary data, the χ^2 approximation does not apply.

In fact for the logistic regression model it can be shown that

$$D = -2 \sum_{i=1}^{n} \{\hat{p}_i \log[\hat{p}_i/(1-\hat{p}_i)] + \log(1-\hat{p}_i)\}$$

which depends only on y_i through $\hat{p}_i$ therefore can tell us nothing about agreement between y_i and $\hat{p}_i$.

Instead we shall analyse the residuals and consider alternative models.

Residual Plots for Binary Data

For binary data, or binomial data where n_i is small for most covariate patterns, there are few distinct values of the residuals and the plots may be uninformative:

```
plot(model1)
```

Therefore we consider “large” residuals

Checking Outliers

We check residuals >2:

```
r <- residuals(model1)
r[abs(r) > 2]
Data[abs(r) > 2, ]
sum(bron[cigs == 0])
```

The model appears to fit poorly for non-smokers

More Complex Models

Further investigation shows that

- modelling non-smokers separately, i.e.

  ```
  model2 <- glm(bron ∼ smoker + smoker:(cigs + poll),
                     family = binomial)
  ```

  does not improve the model
- adding second and third order terms does improve the model, but results in very complex model

We suspect that we are missing an important aspect of the data.

Grouping the Data

A useful technique in evaluating models fit to binary data is to group the data and treat as binomial instead.

We select category boundaries to give roughly equal numbers in each category:

```
cutCigs <- cut(cigs, c(-1, 0, 1, 3, 5, 8, 50))
cutPoll <- cut(poll, c(-1, 55, 57.5, 60, 62.5, 65, 100))
xtabs(∼ cutCigs)
xtabs(∼ cutPoll)
```

Then we compute the proportions from the original data:

```
total <- xtabs( ~ cutCigs + cutPoll)
presence <- xtabs(bron ~ cutCigs + cutPoll)
absence <- c(total) - c(presence)
binData <- as.data.frame.table(presence,
                               responseName = "presence")
binData <- cbind(binData, absence)
```

Modelling Binomial Data

We go back to a model with first order terms:

```
model4 <- glm(cbind(presence, absence) ~ cutCigs + cutPoll,
              family = binomial, data = binData)
summary(model4)
```

The deviance can now be used as a measure of goodness-of-fit, showing that the model fits well.

Checking for Linearity

Using anova, we see that `cutPoll` is close to significance

```
anova(model4, test = "Chisq")
```

We should look more closely before dropping this variable.

The linear trend in the fitted effects for `cutPoll` suggests this factor could be treated as a continuous variable.

```
model5 <- glm(cbind(presence, absence) ~ cutCigs + c(cutPoll),
              family = binomial, data = binData)
anova(model5, test = "Chisq")
```

The difference in deviance is not significant and now both terms are significant in the model.

Looking at the fitted effects for `cutCigs` suggests we can simplify further.

Checking for Nonlinearity

```
levels(binData$cutCigs) <- c("0", "(0,3]", "(0,3]", "(3,8]",
                                    "(3,8]", "8+")
model6 <- update(model5)
```

The difference in deviance is minimal.

The large negative effect for `cutCigs = (0,3]` `corresponds to` $\hat{p} \approx 0$. This occurs because there were no cases of bronchitis observed for smokers of < 3 cigarettes/day.

There were cases of bronchitis amongst non-smokers however, explaining why models based on the continuous variable `cigs` required higher order terms.

Back to Binary Data

Now we can apply what we have found to the original data.

We create a `cigs` factor and use this instead of a continuous variable:

```
cigs <- cut(cigs, c(-1, 0, 3, 8, 50))
model7 <- glm(bron ~ cigs + poll, family = binomial)
```

There are now only three residuals with absolute value greater than two, all corresponding to non-smokers with bronchitis.

```
r <- residuals(model1)
r[abs(r) > 2]
anova(model1, model7, test = "Chisq")
```

The model is a significant improvement on model 1, whilst being simple to interpret.

Bronchitis Model

```
Call:  glm(formula = bron ~ cigs + poll, family = binomial)

Coefficients:
(Intercept)    cigs(0,3]    cigs(3,8]   cigs(8,50]
    -9.3667     -17.3944       1.5416       2.4657
       poll
     0.1241

Degrees of Freedom: 211 Total (i.e. Null);  207 Residual
Null Deviance:     221.8
Residual Deviance: 143.7  AIC: 153.7
```

Interpretation of the Model

- An increase in `Poll` of one unit multiplies the odds on bronchitis by $\exp(0.1241) =$ 1.132.
- Smokers of 3-8 and more than 8 cigarettes per day have their odds on bronchitis multiplied by $\exp(1.54) =$ 4.67 and $\exp(2.47) = 11.77$ respectively, compared with non-smokers.
- There were no observed cases amongst smokers of < 3 cigarettes a day.

Exercises

In our first exercise we consider data related to the NASA Space Shuttle Challenger Disaster.

1. The Challenger had two booster rockets, each with three joints sealed by O-rings. A damaged O-ring can allow a gas leak, which may lead to disaster. The forecast for the time that the Challenger was due to launch was $31\,^{\circ}\mathrm{F}$, whilst the coldest previous launch temperature was $53\,^{\circ}\mathrm{F}$. The day before launch, the engineers met to decide whether the flight should go ahead. They considered a plot of damaged O-rings against temperature using data from previous flights.

Read in the data `shuttle.txt` using `read.table` and reproduce this plot, i.e. plot `damage` vs. `tempF` for observations where `damage == 1` (consider the limits of the y-axis carefully!).

It was decided to go ahead with the flight. Does this decision seem reasonable given the plotted data?

Now plot `damage` vs. `tempF` for all O-rings. Does the decision seem reasonable in the light of this plot?

2. To help interpret the plot, add a piecewise linear model to the plot:

- use `cut` to convert `tempF` to a factor with one level for each 5-degree interval of temperature, i.e. 50-55, 55-60, . . . 80-85
- use `tapply` to compute the mean of `damage` for each level of the new factor
- create a vector of the left endpoints $(50, 55, \ldots)$ and a vector of the right endpoints $(55, 60, \ldots)$ of the temperature intervals
- use `segments` to add the mean line in each interval to the plot

3. The piecewise linear model gives an idea of the trend but a logistic model will quantify the effect of temperature on the expected odds of damage. Use `glm` to fit this logistic model and summarise the results. Is there a significant relationship between `tempF` and `damage`?

Since the data are binary, we cannot use the deviance to check goodness-of-fit. Look instead at the data corresponding to the large residuals. What do you notice?

4. Use `predict` to add a fitted line to data plot. Compare this to the pattern in the residuals.

5. Since there are six O-rings altogether, we have repeated observations at each temperature. Therefore we can also analyse the data as binomial observations:

- create a factor grouping each set of six observations
- use `tapply` to sum `damage` over the levels of the new factor
- compute the number of “failures” (undamaged O-rings) in each launch
- use `tapply` with `min` to obtain the temperature at each launch

Fit a binomial logistic model equivalent to that in question 3 and compare the results. What is the same? What is different?

6. Plot the observed proportions damaged and use `predict` to add the fitted line from the binomial model. Compare this to your observations in question 5.

7. Interpret the coefficient of `tempF` in the binomial model. Use `predict` to predict the probability of damage at $31\,^{\circ}\mathrm{F}$. Assuming the six O-rings fail independently, how many failures does the model predict will fail at this temperature? Would you have recommended that the flight should go ahead?

8. Use `read.table` to read the file `"car_income.txt"`, which records for 33 families

- `purchase` – whether or not the family purchased a car in the past 12 months
- `income` – annual family income ($1000)
- `age` – age of family car

Use `plot` to explore the bivariate relationships of `purchase` with the other variables.

9. Fit a logistic model regressing `purchase` on `income` and `age`. Are both variables significant?

10. Repeat the analysis in question 9 treating age as a factor. The effect for cars that are 6 years old has a large standard error, why is this?

Look at the fitted effects for age. Can age be replaced by a factor with fewer levels?