A8 A study has n_i independent binary observations $\{y_{i1}, \ldots, y_{in_i}\}$ at x_i, $i = 1, \ldots, N$, with $n = \sum_{i=1}^{N} n_i$. Consider the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, where $\pi_i = \Pr(Y_{ij} = 1)$, $j = 1, \ldots, n_i$.

(a) Show that the kernel of the likelihood (i.e. the part of the likelihood that depends on the parameters) is the same if treating the data as n Bernoulli observations or N binomial observations. Explain what this result implies in terms of parameter estimates for grouped and ungrouped data.

(b) For the saturated model, explain why the likelihood function is different for these two data forms. Hence, the deviance reported by the software depends on the data entry.

(c) Use the following data to illustrate your findings in (a) and (b).

x	Number of trials	Number of successes
0	3	1
1	4	2
3	5	4

Create a data files in two ways, entering the data as (i) ungrouped data, (ii) grouped data. Fit the models M_0 with $\text{logit}(\pi_i) = \beta_0$ and M_1 with $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$. For the two forms of data entry, compare parameter estimates in M_1, the deviance of M_1, and the difference between deviances of M_1 and M_0.

A9 Suppose the logistic model (i.e. binomial GLM with binary response and the logit link) holds in which x is uniformly distributed between 0 and 1, and $\text{logit}(\pi_i) = -2.0 + 0.4 x_i$.

(a) Generate a sample of size $n = 100$ from this model as follows. Set the seed to your favourite number, for example

```
set.seed(1)
```

Next, generate a sample of size 100 from the uniform distribution between 0 and 1 and calculate the vector π of length 100 of the probabilities π following the above logit model. Finally, for each $i = 1, \ldots, n$, draw independently one observation from the Bernoulli distribution with probability π_i. To these randomly generated responses, fit the above binomial GLM and plot the Pearson and the deviance residuals. Explain why the residual plots have this appearance? What are the "lines" that are visible on the plot? Calculate the curves they correspond to and plot them on the same picture as the residuals.

(b) As in part (a) set the seed to your favourite number, for example

```
set.seed(1)
```

and generate a sample of size 100 from the uniform distribution between 0 and 1, and calculate the vector π of length 100 of the probabilities π following the above logit model. Next, $N = 1000$ times, generate, independently, a sample of size 100 from the logit model as in part (a). For each generated sample (you should have $N = 1000$ of them), fit the above binomial GLM and store the value of the deviance. Plot the histogram of these 1000 values of the deviance and overlay the density of the χ^2 distribution with $n - p = 100 - 2 = 98$ degrees of freedom. What do you observe?

(c) Repeat (b) with $n = 1000$ instead of $n = 100$. What is your finding now? Think of a possible explanation.

A10 You plan to study the relation between $x =$ `age` and Y, whether the person belongs to a social network such as Facebook ($1 =$ `yes`). A priori, you predict that $\Pr(Y = 1|x = 18)$ is currently between 0.8 and 0.9, and that $\Pr(Y = 1|x = 65)$ is between about 0.2 and 0.3. If the logistic regression model describes this relation well, what is the plausible range of values for the effect β_1 of `age` in the model?

A11 Consider the following data on home-well contamination in 3020 households in Arahazar upazila, Bangladesh. The response variable is `switch` (binary variable whether or not the household switched to another well from an unsafe well). Other variables collected for each household were `arsenic` (the level of arsenic contamination in the household's original well, in hundreds of micrograms per liter), `dist100` (distance in 100-meter units to the closest known safe well), `educ` (years of education of the head of the household) and `assoc` (whether or not any members of the household participated in any community organizations: no or yes). The data is available in `MyCourses` under `Assignments`. Load the data and compute `dist100` as follows.

```
wells <- read.table("wells.dat")
attach(wells)
dist100 <- dist/100
```

(a) Fit a logistic regression model with the intercept and `dist100`. Test the adequacy of this model using the Pearson X^2 and the likelihood ratio G^2 statistics. Conclude at the 5% level.

(b) Using deviance-based comparisons, find the most appropriate logistic regression model for the data and interpret it.

(c) Plot the ROC curve of the model you found in part (b) and compare it to the ROC curve of the model in part (a). Interpret the plot.