Binary responses

A dataset from *Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression* concerns factors affecting the development of myopia (short-sightedness) in children. Our response is binary, 1 indicating that the child developed myopia within the period of study, and 0 indicating they did not. The explanatory variables available to us are the gender of the child, the myopia status of each of their parents, and the number of hours per week spent by the child outside school on various activities: sport, reading for pleasure, playing computer games, studying for school assignments and watching television. Download the data from the course webpage:

```
> file_path <- "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> Myopia <- read.csv(paste0(file_path, "Myopia.csv"))
> Myopia[1:3, ]
> attach(Myopia)
```

To perform logistic regression, or to fit any generalised linear model, we can use the `glm` function (as usual with a new function, do `?glm` to find out more). This works rather like the `lm` in terms of how arguments must be specified and also in terms of procedures for accessing details of the fitted model. We may fit a binomial logistic regression model to our data with `myopic` as the response with the following code:

```
> MyopiaLogReg1 <- glm(myopic ~ ., data = Myopia, family = binomial)
> summary(MyopiaLogReg1)
```

We have to specify the response distribution which is (scaled) binomial. The link is taken as the canonical logistic link unless we specify otherwise. We see the output from `summary` applied to a `glm` object is very similar to that obtained when it is applied to an `lm` object. In the former case, the function `summary.glm` is called (in the latter case it is `summary.lm` that is used). You can find out more about `summary.glm` by accessing the help with `?summary.glm`.
You should be able to understand most of the output produced. Why are there no coefficients estimates for `mumMypoicNo`? The standard errors are given by the square roots of the diagonal entries of the inverse of the Fisher information matrix evaluated at the maximum likelihood estimator. Recall that the m.l.e. $\hat{\beta}$ satisfies $\hat{\beta} \sim AN_p(\beta, i^{-1}(\hat{\beta}))$. The z values are then

$$\frac{\hat{\beta}_j}{\sqrt{\{i^{-1}(\hat{\beta})\}_{jj}}}.$$

Under the null hypothesis that $\beta_j = 0$, this should approximately have a $N(0,1)$ distribution. The final column of the output gives the (approximate) p-values for each of the hypotheses $\beta_j = 0$ for $j = 1, \ldots, p$, using this approximation. Which variables seem to be statistically significant?
At the bottom of the output, we are told the dispersion parameter σ^2 for the binomial family is taken to be 1, as we expect. The null deviance is the deviance of an intercept only model; the residual deviance is what we have called the deviance of the model in lectures.
To test the null hypothesis that the intercept only model is correct, against the alternative of the model we have fitted, we can perform a likelihood ratio test with the following:

```
> pchisq(480.08-439.60, df = 617-609, lower.tail = FALSE)
[1] 2.607117e-06
```

Should we reject the null hypothesis?
Rather than performing the test manually, we can let R do the work using the `anova` function. We have already seen it in action for `lm` objects. With `glm` objects it performs what is sometimes known as an *analysis of deviance*.

```
> MyopiaLogReg0 <- glm(myopic ~ 1, family = binomial)
> anova(MyopiaLogReg0, MyopiaLogReg1, test = "LR")
Analysis of Deviance Table

Model 1: myopic ~ 1
Model 2: myopic ~ gender + sportHR + readHR + compHR + studyHR + TVHR +
    mumMyopic + dadMyopic
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       617     480.08
2       609     439.60  8   40.478 2.61e-06 ***
```

In contrast to the `lm` case, here we need to specify that we want to perform a likelihood ratio test (`test = "LR"`) in order for the final column containing the p-value to be given. Make sure you understand where the entries in the table above are coming from.
Let us also fit a model that additionally includes interactions between the categorical variables `mumMyopic`, `dadMyopic`. We do this with the following:

```
> MyopiaLogReg2 <- glm(myopic ~ . + mumMyopic:dadMyopic, data = Myopia, family = binomial)
> summary(MyopiaLogReg2)
```

Exercises

1. Test the null hypothesis that the model in `MyopiaLogReg1` is correct against the more complex alternative model `MyopiaLogReg2`. Which model should you prefer?

2. Now fit a model `MyopiaLogReg3` with all the variables as in `MyopiaLogReg1` but without the `compHR` and `TVHR` variables. Use this to test whether hours spent watching TV and hours spent playing computer games are collectively significant.

3. Now we wish to test whether the coefficients for `mumMyopic` and dadMyopic are the same (i.e. we want to see whether the mother's or the father's myopia contributes more to predicting that their child is myopic). To do this, you'll need to create another variable `mumORdadMyopic`:

   ```
   > mumORdadMyopic <- (dadMyopic == "Yes") | (mumMyopic == "Yes")
   > mumORdadMyopic[1:10]
    [1]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
   > mumORdadMyopic <- factor(mumORdadMyopic, labels=c("No", "Yes"))
    [1] Yes Yes No  Yes Yes Yes Yes No  No  No
   Levels: No Yes
   ```

   Each of `(dadMyopic == "Yes")` and `(mumMyopic == "Yes")` are logical vectors and `|` performs an OR operation. The final line is not really necessary but simply makes `mumORdadMyopic` a factor with levels "Yes" and "No" so it is similar to the other variables. Fit a model similar to the model you have fitted in the previous question, but swapping `mumMyopic` for the new variable `mumORdadMyopic`. What can you conclude?

Binomial responses with multiple trials

Download the `Smoking` data from the course webpage with

```
> detach(Myopia)
> (Smoking <- read.csv(paste(file_path, "Smoking.csv", sep ="")))
> attach(Smoking)
```

In 1972–1974 a survey was taken; twenty years later a followup study was conducted, and it was determined if each interviewee was still alive. Among the information obtained originally was whether a person was a smoker or not and their age, divided into seven categories. The `Age.group` variable gives the mid-points of these categories. Let us begin by plotting the data.

```
> total <- Survived + Died
> propDied <- Died / total
> plot(propDied[Smoker == "Yes"] ~ Age.group[Smoker == "Yes"],
+ xlab = "Age group", ylab = "Proportion died")
> points(propDied[Smoker == "No"] ~ Age.group[Smoker == "No"], pch = 4)
```

We can also plot the log odds for the proportion who died in each of the age groups.

```
> logit <- function(p) log(p/(1-p))
> plot(logit(propDied)[Smoker == "Yes"] ~ Age.group[Smoker == "Yes"],
+ xlab = "Age group", ylab = "Empirical logit")
> points(logit(propDied)[Smoker == "No"] ~ Age.group[Smoker == "No"], pch = 4)
```

If a logistic regression model taking as response the proportion of interviewees who died and as explanatory variables the age group and smoking status is correct, we should see that the points fall close to two parallel lines. If there was an interaction between smoking status and age, we would expect two non-parallel lines (why?). The following performs a binomial regression with logit link:

```
> SmokingLogReg1 <- glm(propDied ~ Age.group + Smoker, family = binomial,
+ weights = total)
> summary(SmokingLogReg1)
```

Note we now must specify the `weights` for the GLM. Recalling that the dispersion parameter of the i^{th} observation in a GLM is $\sigma_i^2 = a_i\sigma^2$, the i^{th} *weight* is ${a_i}^{-1}$, which is n_i when the response is $n_i^{-1}\text{Bin}(n_i, \mu_i)$. R uses these weights to form the W_m matrix used in the iterated weighted least squares algorithm.
From the final lines of the `summary` output, we can easily see that we reject the null hypothesis that the intercept only model is correct, against the alternative of the model we have fitted (how can we see this? Recall that $\mathbb{E}(Z) = 2$ when $Z \sim \chi^2_2$).
Exercise: Give an estimate for the multiplicative change in odds of dying for an increase in age of one year.
Since the n_i are quite large in our model, small dispersion asymptotics suggest that the deviance should approximately follow a χ^2_{n-p} $(n-p = 11)$ distribution, if the model is correct. (Recall that the dispersion parameter σ^2 is 1 in our case, so the deviance is the likelihood ratio statistic for testing the null hypothesis that our model is correct against the saturated model with the μ_i unrestricted). The large value of the deviance of 32.572 compared to $\mathbb{E}(Z) = 11$ when $Z \sim \chi^2_{11}$ suggests that the model fit is not too good. Indeed

```
> pchisq(32.572, df = 11, lower.tail = FALSE)
[1] 0.0006170603
```

We can try to improve the model by perhaps introducing a quadratic effect of age, or treating age as a category or factor.

```
> SmokingLogReg2 <- glm(propDied ~ Age.group + I(Age.group^2) + Smoker, family = binomial,
+ weights = total)
> SmokingLogReg3 <- glm(propDied ~ factor(Age.group) + Smoker, family = binomial,
+ weights = total)
```

Note the `I` is needed in the first model fit above so that the `^` operator is treated in the usual way, and it simply squares the age. Also try experimenting with other link functions:

```
> SmokingLogReg4 <- glm(propDied ~ Age.group + Smoker, family = binomial(link=probit),
+ weights = total)
> SmokingLogReg5 <- glm(propDied ~ Age.group + Smoker, family = binomial(link=cloglog),
+ weights = total)
```