MATH38172 Generalised Linear Models

Computer Lab 3

Fitting GLMs with various response distributions

Recall that the assumptions of a GLM are that

$$
\begin{aligned}
&Y_i \overset{\text{indep}}{\sim} \mathcal{F}(\theta_i, \phi/w_i) && \mathcal{F} \text{ an exponential dispersion family} \\
&\mu_i = \mathrm{E}(Y_i) && \mu \text{ the expected response} \\
&g(\mu_i) = \eta_i && g \text{ the link function} \\
&\eta_i = \beta_1 x_1 + \ldots + \beta_p x_p && \eta \text{ the linear predictor}
\end{aligned}
$$

We have already seen how to fit two types of GLM using the `glm` command: Poisson response models with a log link (using `family=poisson`), and Binomial response models with a logit link (using `family=binomial`).

More generally, we can fit models with other response distribution families, e.g.

- `glm(y~x, family=Gamma)` for a Gamma GLM
- `glm(y~x, family=inverse.gaussian)` for an inverse Gaussian response

The gamma and inverse-Gaussian distributions are commonly used to model positive continuous responses. They may be appropriate if the variance increases with the mean, since $V(\mu) = \mu^2$ for the Gamma distribution and $V(\mu) = \mu^3$ for the inverse-Gaussian.

By default, R uses the canonical link function, or a scalar multiple, e.g.:

Family	Canonical link	Equation	R link
Binomial proportion	logit	$g(\mu) = \log \frac{\mu}{1-\mu}$	
Poisson	log	$g(\mu) = \log \mu$	
Gamma	reciprocal	$g(\mu) = -1/\mu$	$g(\mu) = 1/\mu$
Inverse-Gaussian	inverse-square	$g(\mu) = -1/(2\mu^2)$	$g(\mu) = 1/\mu^2$

Note that, for a gamma response, using R's link rather than the true canonical link corresponds to multiplying the β_j by -1.

Different link functions can be specified using the `link` argument as follows:

- `glm(y~x, family=binomial(link="probit"))`
- `glm(y~x, family=Gamma(link="log"))`
- `glm(y~x, family=Gamma(link="identity"))`

For full details of options, see the R help file `?family`

For Gamma responses, the canonical link has the problem that $\mu = 1/\eta$ is negative for some values of η. As the expectation of a gamma distribution must be positive, this leads to constraints on the parameters and explanatory variables. However, the log link does not suffer from this problem, as $\mu = g^{-1}(\eta) = e^{\eta}$ is always positive. As a result the log link is a common choice for the Gamma distribution.

Model comparison

Hypothesis tests

Recall that if we wish to compare two GLMs with the same response distribution but different (nested) linear predictors, e.g.

$$\begin{aligned} &\text{Model A:} \qquad \eta = \beta_1 x_1 + \cdots + \beta_{p_A} x_{p_A} \\ &\text{Model B:} \qquad \eta = \beta_1 x_1 + \cdots + \beta_{p_A} x_{p_A} + \cdots + \beta_{p_B} x_{p_B} \end{aligned}$$

then this can be done by testing the null hypothesis H_0 : Model A is correct versus H_1 : Model B is correct.

Recall that for some exponential families (e.g. Binomial or Poisson) the dispersion parameter is known (e.g. $\phi = 1$). As we have seen in lectures, in this case, the fit of Model A and Model B can be compared using a likelihood ratio test:

$$\text{Reject } H_0 \text{ if } L = 2\{\ell_B - \ell_A\} = \frac{1}{\phi}[D(\hat{\boldsymbol{\mu}}_A, \mathbf{y}) - D(\hat{\boldsymbol{\mu}}_B, \mathbf{y})] > \chi^2_{\alpha; p_B - p_A}\,.$$

This can easily be done in R using the `anova` command.

Example For the Small Business Administration data (SBA.csv), compare the binary-response logistic regression models with the following two linear predictors:

$$\begin{aligned} &\text{Model A:} \qquad \eta \sim \texttt{Portion} \\ &\text{Model B:} \qquad \eta \sim \texttt{Portion + Recession + RealEstate} \end{aligned}$$

```
SBA <- read.csv("SBA.csv")
fitA <- glm(Default~Portion, family=binomial, data=SBA)
fitB <- update(fitA, ~.+ Recession +RealEstate)
```

Note the use of the `update` command to include additional variables in the linear predictor of Model A. Then conduct the test:

```
anova(fitA,fitB,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Default ~ Portion
## Model 2: Default ~ Portion + Recession + RealEstate
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2100     2340.7
## 2      2098     2222.9  2   117.79 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the additional terms in the linear predictor significantly improve the model; the p-value of the test is $< 2.2 \times 10^{-16}$, so we would reject H_0 at any reasonable significance level.

It can sometimes be useful to directly access the deviance of a given model, and/or its residual degrees of freedom $n - p$. This can be done via

```
deviance(fitA); df.residual(fitA)
```

```
## [1] 2340.664
```

```
## [1] 2100
```

For other families (e.g. Gamma/inverse-Gaussian), ϕ is unknown. As we have seen in lectures, in this case, the fit of Model A and Model B can be compared using analysis of deviance, i.e.

$$\text{Reject } H_0 \text{ if } F = \frac{[D(\hat{\boldsymbol{\mu}}_A, \mathbf{y}) - D(\hat{\boldsymbol{\mu}}_B, \mathbf{y})]/(p_B - p_A)}{\hat{\phi}_B} > F_{\alpha;(p_B-p_A),(n-p_B)} \,,$$

where $\hat{\phi}_B$ is an estimate of ϕ computed using Model B, usually the Pearson estimate. Again this can be done easily in R using the `anova` command in R.

Example For the house price data (Houses.csv), using a Gamma response GLM with `price` as the response and identity link, compare the following two linear predictors:

$$\begin{array}{ll} \text{Model A} & \eta \sim \texttt{size} \\ \text{Model B} & \eta \sim \texttt{size} + \texttt{new} + \texttt{size:new} \end{array}$$

```r
Houses <- read.csv("Houses.csv")
fitA <- glm(price~size, family=Gamma(link="identity"), data=Houses)
fitB <- glm(price~size+new+new:size, family=Gamma(link="identity"), data=Houses)
anova(fitA,fitB,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: price ~ size
## Model 2: price ~ size + new + new:size
##   Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
## 1        98     11.283
## 2        96     10.563  2  0.72071 3.2698 0.04229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus the inclusion of the main effect of `new` and the interaction term `new:size` significantly improves the model at the 5% significance level.

Exercises 1

1. For the SBA data:

    a. Assess whether inclusion of the Portion:RealEstate and Portion:Recession interactions significantly improves Model B.
    b. Interpret the resulting model.
    c. Try fitting a model that includes the main effects and two-factor interactions between Portion, RealEstate, and Recession. What do you notice about the parameter estimates?
    d. Why does this happen? [Hint: tabulate the combinations of values of the variables Recession and RealEstate that are present in the data using `table(SBA$Recession, SBA$RealEstate)`]

2. Suppose that $n = 300$ patients are treated, with $m = 100$ patients allocated to each of three treatments. Each patient either recovers or not. The sample recovery rates are 87%, 83%, and 78%, for Treatments A, B, and C respectively. Using R, assess if there is significant evidence at the 5% significance level that the probability of recovery depends on which treatment is applied.

3. The file `nambeware.csv` contains data about items produced by Nambe Mills, a tableware manufacturer. The following variables are included: `Type` (item type), `Diam` (item diameter), `Time` (item polishing and grinding time) and `Price` (item price).

    a. Plot item price versus diameter and explain why a Gamma GLM may be suitable.
    b. Using Gamma GLMs with a log link, assess whether there is significant evidence that the item price depends on the item type, adjusting for item diameter and polishing and grinding time.