MATH3012 – Statistical Methods II
S-Plus Worksheet 5 – Generalised linear models

Fitting generalised linear models in S-Plus is very similar to fitting linear models. You replace the command `lm` with the command `glm`. The distribution and link function which you require for your model is specified by the argument `family` which is supplied to the model. For example `family=binomial(logit)` fits a generalised linear model with binomial distribution and logistic link. In fact `family=binomial` does the same, as the default is the canonical link. The form of the linear predictor is specified in the model formula for `glm` in the same way as for `lm`.

The result of the `glm` command is a generalised linear model object which can be used within many of the same S-Plus commands as a linear model object. Useful ones are `resid`, `coef`, `deviance`, `fitted`, `plot`, `print`, `summary`, `update` and `anova`. However, the commands `add1` and `drop1` do not produce useful output for generalised linear models.

1. **Obtain the source**. Fire up the internet explorer. Go to the course webpage `http://www.maths.soton.ac.uk/staff/Sahu/teach/math3012` or otherwise. Click on the source file for worksheet 4.

2. **Import the data**

   (a) To do this you just install the MATH3012 files and click on the `beetle` icon with the `S` symbol on my computer window.

   (b) This dataset represents the number of beetles exposed (n) and number killed (y) in eight groups exposed to different doses (x) of a particular insecticide. Interest is focussed on how mortality is related to dose. It seems sensible to model the proportion of beetles killed in each group as a binomial random variable with probability of death depending on dose.

3. Pretend that we do not know the GLM theory and are just going to use the normal linear models. We issue the following commands.

   ```
   beet.lm <- lm(y/n ~ x , data=beetle, weights=n)
   summary(beet.lm)
   predict(beet.lm)
   ```

   The predicted probability for dose 1.88 is 1.085304! This is not the work of a good statistician! Probability cannot be bigger than 1.

4. We should fit the logistic regression model. A *logistic regression model* is

$$Y_i|n_i, p_i \sim \text{Binomial}(n_i, p_i), \qquad \log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 x_i, \qquad i = 1, \ldots, 8.$$

   This is a generalised linear model with a binomial distribution for the response and logistic link function.

5. To illustrate the effect the logit link function has, it is useful to plot y/n against x and compare with a plot of the *empirical logit*, $\text{logit}(y/n)$ against x. In fact, because there are extreme values $y/n = 1$ in the data, we use the *modified* empirical logit, $\text{logit}([y + \frac{1}{2}]/[n + 1])$ in the second plot. Does it look as if a linear logistic model will fit well? The commands are:

(a) `plot(beetle$x, beetle$y/beetle$n)`

(b) `p <- (beetle$y+0.5)/(beetle$n+1)`

(c) `plot(beetle$x, log(p/(1-p)))`

6. To fit the linear logistic regression model in S-Plus, we use the command

```
beet.glm <- glm(y/n~ x, data=beetle, family=binomial, weights=n)
```

The response variable for a binomial GLM can be given in one of several formats. It can just be a factor with two levels (one for success and the other for failure). Here we have y out of total n, hence we have used this format. See `?glm`.

7. **Interpreting the output**

(a) Issue the command `summary(beet.glm)`. It provides similar information to the equivalent command for a linear model.

(b) `Call` confirms the model which was fitted.

(c) `Residuals` gives a summary of the distribution of the *deviance* residuals. We will discuss these later.

(d) `Coefficients` provide the maximum likelihood estimates, together with their standard errors. The column t-value should be compared to the cut-off point from the standard normal distribution, e.g. 1.96 at the 5% level of significance. Using the asymptotic normality of the mle's we can obtain the confidence intervals as well. We interpret the coefficient for x as follows. We fitted the model:

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_1 + \beta_2\, x_i = -60.72 + 34.27 x_i$$

The maximum likelihood (ML) estimate of β_1 is $\hat{\beta}_1 = -60.72$ with s.e. (asymptotic standard error) 5.18. From this we can test $H_0 : \beta_1 = 0$ by performing a normal test. We will reject H_0 if $\left|\hat{\beta}_1/s.e.\right| > 1.96$. In this case we do reject H_0 at 5% level of significance.

The ML estimate of β is $\hat{\beta} = 34.27$ with s.e. 2.91. We interpret this as follows. For a unit change in x (dose), the estimated odds of beetle being killed are multiplied by `exp(34.27)`. For example, consider two beetles with $x = 1.7$ and 1.8 respectively. The odds of the second beetle being killed is `exp(34.27 × 0.1 )` times the odds of the first beetle being killed. This means that the second beetle with a higher dose is much more likely to be killed than the first beetle.

For β_2 also we can perform $H_0 : \beta_2 = 0$ and form 95% confidence interval given by $\hat{\beta}_2 \pm 1.96 \times s.e.$

(e) `Residual Deviance` is the scaled deviance for the model, which can be compared with a chi-squared distribution to assess the goodness of fit of the model.

In our example, it is 11.23 on 6 degrees of freedom. This means that a deviance value of only 11.23 is NOT explained by the fitted model. Put this on a theoretical χ^2 distribution

of 6 df. The residual is not large, since `1-pchisq(11.23, df=6)` is 0.082. Thus the residual deviance did not fall above the 5% critical value.

The deviance of a model provides an overall measure of goodness of fit which can be calibrated. Generally, if the model is an acceptable fit then the residual deviance of the model will be an observation from a chi-squared distribution whose degrees of freedom are the same as the residual degrees of freedom of the model. If the model is a poor fit then the deviance will be larger than would be predicted by the relevant chi-squared distribution.

(f) `Null Deviance` is also provided, so that the model may be compared with the null model (without the covariate x), again using an appropriate chi-squared distribution. This is the residual deviance for the model with intercept only.

(g) `Number of Fisher Scoring Iterations` tells us how quickly the Fisher scoring algorithm converged to the maximum likelihood estimate.

(h) `Correlation of Coefficients` is the estimated correlation of the estimates of the coefficients. If correlations are high, then removing terms from the model may result in other coefficients changing their values considerably.

(i) Issue the command `anova(beet.glm, test="Chisq")`. It provides similar information to the equivalent command for a linear model. It tests whether the it is worth including x in the model.

8. Let us see the fitted probabilities. We issue `predict(beet.glm, type="response")` All probabilities are between 0 and 1.

9. **Other link functions:** S-Plus allows us to use link functions other than the canonical link. For example, for binomial data, we can use

$$g(\mu) = \Phi^{-1}(\mu)$$

where Φ is the standard normal distribution function, so $\Phi(z) = P(Z \leq z)$ where Z is a standard normal random variable. This is the *probit* link function. Alternatively

$$g(\mu) = \log[-\log(1-\mu)]$$

is called the *complementary log-log link function.* Note that both of these links map $(0, 1)$ on to $\mathcal{R}$.

10. Use the function `comparelinks()` to see the differences between these link functions. You have to compile this function previously downloaded from the website.

11. Try fitting the models using these alternative link functions. For example,

```
beet.prlink <- glm(y/n~ x, data=beetle, family=binomial(probit), weights=n)
beet.cloglink <- glm(y/n ~ x, data=beetle, family=binomial(cloglog), weights=n)
summary( beet.prlink)
summary( beet.cloglink)
summary(beet.glm)
```