MATH3012 – Statistical Methods II
S-Plus Worksheet 7 – A large factorial example

- The `heartattack` dataset represents the results of a clinical trial to assess the effectiveness of a thrombolytic (clot-busting) treatment for patients who have suffered an acute myocardial infarction (heart attack). The first column `y` is the number of patients who survived the attack for 35 days. The column `n` is the number of people who had the heart attack. There are four categorical explanatory variables, representing the treatment the patient was given (`Rtreat`: active or placebo); whether the patient was already taking Beta-blocker medication prior to the infarction (`Blocker`: yes or no); time between infarction and treatment (`Time`: Le12H (less than 12 hours) or Mo12H (more than 12 hours); site of infarction (`Site`: anterior, inferior or other). We shall abbreviate the factors to be R, B T and S where R =Rtreat, B =Blocker, T =Time and S=Site.

- By fitting logistic regression models to these data, find the model which you feel best explains the dependence of survival (y) on the explanatory variables. You should allow for potential interactions. Is a binomial model valid for these data?

More about factors

1. Recall from Worksheet 3 that we refer to categorical explanatory variables as *factors*, and factorial models can include *main effects* and *interactions*. For example, a logistic regression model which allows survival (Y) to depend on S, R and an interaction between these two factors is

$$Y_i|n_i,p_i \sim \text{ Binomial}(n_i,p_i), \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha+\beta_S(s_i)+\beta_R(r_i)+\gamma_{S,R}(s_i,r_i), \quad i=1,\ldots,24.$$

   where s_i is the level of S, and r_i is the level of R for the ith observation. Here β_S, the main effect of S takes different values depending on the level of S, so in principle β_S takes three values [β_S(anterior), β_S(inferior) and β_S(other)]. Similarly, β_R depends on the level of R (two values) and $\gamma_{S,R}$ depends jointly on the levels of S and R (six values).

2. In practice, by setting

   ```
   options(contrasts=c("contr.treatment", "contr.poly"))
   ```

   in S-Plus,we constrain any main effect to be equal to 0 at the first level of a factor.

3. If the factor levels are not labelled numerically, S-Plus interprets the 'first' level to be the first in alphabetical order. If a factor has been coded using numeric labels, then it needs to declared as a factor in S-Plus using `factor`.

4. The concept of interaction can be extended when we have three or more factors. For example, a *three factor interaction* allows the response to depend jointly on three factors. Hence SBR, the three factor interaction between S B and R, corresponds to coefficients of the form $\gamma_{S,B,R}(s_i,b_i,r_i)$. Setting

   ```
   options(contrasts=c("contr.treatment","contr.poly"))
   ```

constrains any interaction to be equal to 0 for all combinations where any of the factors are at their first level. Hence, where the main effect S involves $l_S - 1$ free coefficients, the interaction SR involves $(l_S - 1)(l_R - 1)$ free coefficients, SBR involves $(l_S - 1)(l_B - 1)(l_R - 1)$ free coefficients, *etc.* where l is the number of levels of each factor ($l_S = 3$, $l_B = 2$ and $l_R = 2$ for the data in page 4).

5. It rarely makes sense to include some coefficients of a main effect or an interaction but not others.

6. Interactions involving more and more factors become progressively more difficult to interpret. A two-factor interaction like SR allows the way in which the response Y depends on S (R) to be different for different levels of R (S), *i.e.* the way in which one factor affects the response depends on the level of the other factor. Presence of the three factor interaction SBR means that the way in which the dependence of Y on S varies with R, depends on the level of B! One rule which must be followed is the *principle of marginality* which states that 'whenever an interaction is present in a model, all *marginal* main effects and interactions (those which correspond to 'subsets') must also be present. For example, if we include the SR interaction, then the main effects of S and R must also be in the model, as above. Similarly if we include the three factor interaction SBR, the main effects S, B, R and the interactions SB, SR, BR must all be present. If the principle of marginality is violated, then factorial models become almost impossible to interpret.

7. In S-Plus, we use the shorthand `S:R`, `S:B:R` *etc.* to denote interaction terms in a model formula. Another useful shorthand permitted by S-Plus is `S*B*R` which represents the interaction `S:B:R` together with all its marginal terms. We shall also adopt this notation. The number of coefficients corresponding to an expression like $S * B * R$ is the product of the number of levels of the factors concerned ($l_S l_B l_R$ for $S * B * R$).

8. If the data set consists of a perfectly structured array, with every combination of the explanatory factors appearing exactly once, then the model containing the highest possible interaction, together with all marginal terms, is the saturated model (scaled deviance 0 on 0 degrees of freedom). Therefore, for the data page 4, the saturated model can be interpreted as $S * T * B * R$. It is often useful to take this model as the starting point of a 'backwards elimination' approach to identifying a suitable model.

9. Factors are included in the linear predictor by creating a dummy variable for every level of the factor other than the first. The dummy variable at level f of a factor F takes the value 1 for every observation where $F = f$ and 0 for all other observations. The model coefficient $\beta(f)$ corresponds to the level f dummy variable for F. Similarly, interactions correspond to products of dummy variables. For example, the interaction parameter $\gamma(f, g)$ for factor F at level f and factor G at level g corresponds to an explanatory variable created by multiplying together the corresponding dummy variables.

S-Plus instructions

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 6.

2. **Import the data:** To do this you just install the MATH3012 files and click on the `heartattack` icon with the `S` symbol on my computer window.

3. By a process of backward elimination, all of the interaction terms can be removed without a significant increase in deviance. [You can see this yourself. I am not giving the detailed instructions.] The most marginal decision concerns ST; the log likelihood ratio statistic for the test comparing models $S+T+B+R$ and $B+R+S*T$ is 5.27 on 2 degrees of freedom, leading to a p-value of 0.0716.

   (a) `h.glm <- glm(y/n ~ Rtreat + Blocker + Site * Time, family=binomial, weights=n, data=heartattack)`

   (b) `anova(h.glm, test="Chisq")`

4. Having removed the ST interaction, it is then reasonable to remove the main effect T (log likelihood ratio = 1.95 on 1 df, p-value = 0.1627). We cannot remove any further main effects without significantly increasing the scaled deviance. The p-values for the tests are all significantly small.

5. Therefore, our preferred model is

$$Y_i|n_i,p_i \sim \text{ Binomial}(n_i,p_i), \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_S(s_i) + \beta_B(b_i) + \beta_R(r_i), \quad i = 1,\ldots,24.$$

   where s_i is the level of S, b_i is the level of B, and r_i is the level of R for the ith observation.

   (a) `hglm.final <- glm(y/n ~ Rtreat + Blocker + Site , family=binomial, weights=n, data=heartattack)`

   (b) `summary(hglm.final)`

   (c) `anova(hglm.final)`

6. The model is a good fit. Its scaled deviance is 15.86 on 19 degrees of freedom. We would only have real cause for concern if the deviance exceeded 30.14, the 95% point of χ^2_{19}.

7. Based on the observed data, our parameter estimates, together with their standard errors and 95% confidence intervals are

Parameter	Estimate	Standard error	Confidence interval
$\hat{\alpha}$	2.168	0.114	(1.944,2.391)
$\hat{\beta}_{Site}$(anterior)	0		
$\hat{\beta}_{Site}$(inferior)	0.550	0.141	(0.275,0.826)
$\hat{\beta}_{Site}$(other)	—0.152	0.155	(–0.456,0.151)
$\hat{\beta}_{Blocker}$(no)	0		
$\hat{\beta}_{Blocker}$(yes)	–0.378	0.146	(–0.665,–0.091)
$\hat{\beta}_{Rtreat}$(active)	0		
$\hat{\beta}_{Rtreat}$(placebo)	–0.283	0.121	(–0.520,–0.045)

8. Qualitatively this model suggests that probability of 35 day survival is enhanced by the thrombolytic treatment (logit(p) and therefore p is significantly lower for the placebo). Probability of survival is significantly higher for those whose site of infarction is 'inferior'. This is the most pronounced effect. There is no real significant difference between the other sites. Patients who were on prior beta blocker medication also have a lower probability of survival.

9. We investigate the residuals as follows.

   (a) `plot(hglm.final)`

   (b) `u <- resid(hglm.final, type="pearson") # The Pearson residuals are saved in u`

   (c) `v <- resid(hglm.final, type="deviance") # The deviance residuals are saved in v`

   (d) `sum(u^2) # The result is the Pearson X^2 statistic`

   (e) `sum(v^2) # The result is the scaled deviance`

S	T	B	R	Survived? Yes	Survived? No
Anterior	≤ 12 hours	Yes	Active	53	6
			Placebo	42	7
		No	Active	207	20
			Placebo	220	42
	> 12 hours	Yes	Active	50	8
			Placebo	44	12
		No	Active	241	29
			Placebo	257	36
Inferior	≤ 12 hours	Yes	Active	41	7
			Placebo	32	5
		No	Active	223	22
			Placebo	210	20
	> 12 hours	Yes	Active	40	4
			Placebo	50	4
		No	Active	226	11
			Placebo	226	13
Other	≤ 12 hours	Yes	Active	12	2
			Placebo	20	8
		No	Active	73	9
			Placebo	83	13
	> 12 hours	Yes	Active	18	2
			Placebo	17	5
		No	Active	90	13
			Placebo	102	18