MATH3012 – Statistical Methods II

S-Plus Worksheet 3 – Factorial models

1. **Import the data**

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

   (b) This dataset represent survival times (in 10 hour units) of 48 animals each allocated to one of 12 combinations of 4 treatments and 3 poisons. The treatments are in column `t` and the poisons are in column `p`.

   (c) The aim of the experiment is to determine the dependence of survival time on the explanatory variables treatment (T) and poison (P).

   (d) The values of these explanatory variables are categories. We refer to categorical explanatory variables as *factors*. The values assigned to the variables are just labels distinguishing between categories. Even if we used numerical labels, their values would have no intrinsic meaning. Therefore, we cannot include such variables directly in a linear model. We need to convert the two explanatory variables into factors.

   (e) Open the survival data using the object explorer. On the righthand pane of the object explorer you will see three column names, `y, t, p` and their data classes. Here `y` should be numeric. We make `t` and `p` as factors by first highlighting the two columns in the data spreadsheet and then following the path **Data** → **Change Data Type**. Now select `factor` in the `New Type` dialog and press OK. This should do it. Check the object explorer.

2. **Modelling with factors**

   (a) To model the dependence of the response on an explanatory factor, we include a model coefficient for each level of the factor. For example, a model which allows survival time (Y) to depend on P might be written as

$$Y_i = \mu + \alpha(p_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{1}$$

   where p_i is the level of P for the ith observation. Here α takes different values depending on the level of the factor P.

   (b) Note that this model is overparameterised, in that it would be possible to add some fixed constant to each $\alpha(p_i)$ and subtract the same constant from μ and end up with exactly the same dependence of Y on P. Therefore, in practice, we impose a constraint on α. The easiest to interpret is to constrain $\alpha(p)$ to be equal to 0 at the first level of P. In S-Plus we do this by

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

3. **Models**

   (a) The model (1) is then fitted by `survival.lm1 <- lm(y~p,data=survival)`

(b) Factors are included in the standard linear model structure by creating a dummy variable for every level of the factor. 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 dummy variables are incorporated in the design matrix in the usual way. The model coefficient $\alpha(f)$ corresponds to the dummy variable at level f.

(c) Factorial models may include more than one factor. For example, we can model the dependence of Y on T and P using the linear model

$$Y_i = \mu + \alpha(p_i) + \beta(t_i) + \epsilon_i \qquad i = 1, \dots, n$$

fitted in S-Plus by `survival.lm2 <- lm(y~p+t,data=survival)`

(d) This is an additive model, which assumes that the effect of a treatment is the same for every poison, and conversely that the effects of a poison are the same for every treatment. These effects of poison and treatment are called *main effects.*

4. **Models with interactions**

(a) The additive model may not be appropriate. Different poisons may respond differently in combination with different treatments. We call this *interaction.* A model with interaction is

$$Y_i = \mu + \alpha(p_i) + \beta(t_i) + \gamma(p_i, t_i) + \epsilon_i \qquad i = 1, \dots, n$$

We require constraints on the interaction parameters γ, similar to those on the main effects. Usually, we constrain $\gamma(p_i, t_i)$ to be equal to 0 for all combinations where either P or T are at their first level.

(b) The model with interaction can be fitted using

```
survival.lm3 <- lm(y~p+t+p:t,data=survival)
```

where `p:t` indicates an interaction between P and T. Do not attempt to include an interaction term without the corresponding main effects (the model is meaningless).

(c) Note that in factorial models it rarely makes sense to model some of the levels of a factor but not others. Therefore, when we add an explanatory factor F into the model, we are adding $l_F - 1$ extra coefficients, where l_F is the number of levels of the factor concerned. As usual, we can use an F test to determine whether adding a factor leads to a significant improvement in fit.

5. **Exercises**

(a) For this dataset, fit factorial models which include T, P, both, neither, and the model with interaction. By comparing models using F tests (remember the `anova` command) determine which model you prefer. Is it a good model?

(b) Now transform the response to death rate (reciprocal of survival time). Try modelling death rate using factorial models. Which model do you prefer now?

(c) In each case, write down your model and the corresponding parameter estimates. Overall, what is your preferred model. What conclusions do you draw about the effects of the different poisons and treatments?

Some notes

This dataset represent survival times (in 10 hour units) of 48 animals each allocated to one of 12 combinations of 4 treatments and 3 poisons. The aim of the experiment is to determine the dependence of survival time on the explanatory variables Treatment (T) and Poison (P).

Again, I think the safest way to fit models is to start with the most complex model and try to eliminate explanatory variables. Therefore, I first fitted the model with interaction. However, there is no evidence that interaction is required, as removing the interaction does not significantly increase the deviance, but does considerably reduce the complexity (the interaction term contributes 6 parameters to the model). Formally, the F statistic is 1.874 on $6, 36$ degrees of freedom, leading to a p-value of 0.1123.

Neither P the main effect of Poison ($F = 20.6$ on $2, 42$ d.f. p-value=0.0000) or T the main effect of Treatment ($F = 12.3$ on $3, 42$ d.f. p-value=0.0000) should be removed from the model $T + P$.

The chosen model $T + P$ can be written as

$$Y_i = \mu + \alpha(p_i) + \beta(t_i) + \epsilon_i \qquad i = 1, \ldots, 48 \tag{1}$$

where (p_i, t_i), $i = 1, \ldots, 48$ are the observed categories of P and T, $Y_1, \ldots, Y_{48}$ are the survival times, and $\epsilon_1, \ldots, \epsilon_{48}$ are independent $\mathrm{N}(0, \sigma^2)$ residuals.

Based on the observed data, our least squares estimates, together with their standard errors and 95% confidence intervals are

Parameter	Estimate	Standard error	Confidence interval
$\hat{\mu}$	0.452	0.056	(0.339,0.565)
$\hat{\alpha}$(I)	0		
$\hat{\alpha}$(II)	–0.073	0.056	(–0.186,0.040)
$\hat{\alpha}$(III)	–0.341	0.056	(–0.454,–0.228)
$\hat{\beta}$(A)	0		
$\hat{\beta}$(B)	0.363	0.065	(0.232,0.493)
$\hat{\beta}$(C)	0.078	0.065	(–0.052,0.209)
$\hat{\beta}$(D)	0.220	0.065	(0.090,0.350)

The unbiased estimate for σ^2 is $0.1582^2 = 0.02503$.

The consequence of the model is that both P and T have a significant effect on survival time, but they do not interact. In other words, the effect of a particular poison (treatment) does not depend on the treatment (poison) it is combined with. Poisons I and II have similar effects, but Poison III gives an estimated expected survival time 3.4 hours less than Poison I and 2.7 hours less than Poison II. Similarly, the treatments can be ranked BDCA, in order of expected survival time, with the expected difference between B and A being 3.6 hours. We can make these statements comparing treatments or Poisons, because there is no interaction. In the presence of interaction, individual combinations of factors need to be considered separately.

The R^2 coefficient is 65.0%, which indicates that the model is not a very good fit, and that a large proportion of the variability remains after fitting the model. Furthermore, residual plots indicate serious cause for concern. The plot of residuals against fitted values exhibits marked 'funneling', and the normal plot has obvious curvature at its upper end.

With death rate (reciprocal of survival time) as the response variable, we again choose the model $T + P$, as we can eliminate the interaction ($F = 1.09$ on $6, 36$ d.f. $p = 0.3867$) but neither the main effect P ($F = 71.7$ on $2, 42$ d.f. $p = 0.0000$) or T ($F = 28.0$ on $3, 42$ d.f. $p = 0.0000$).

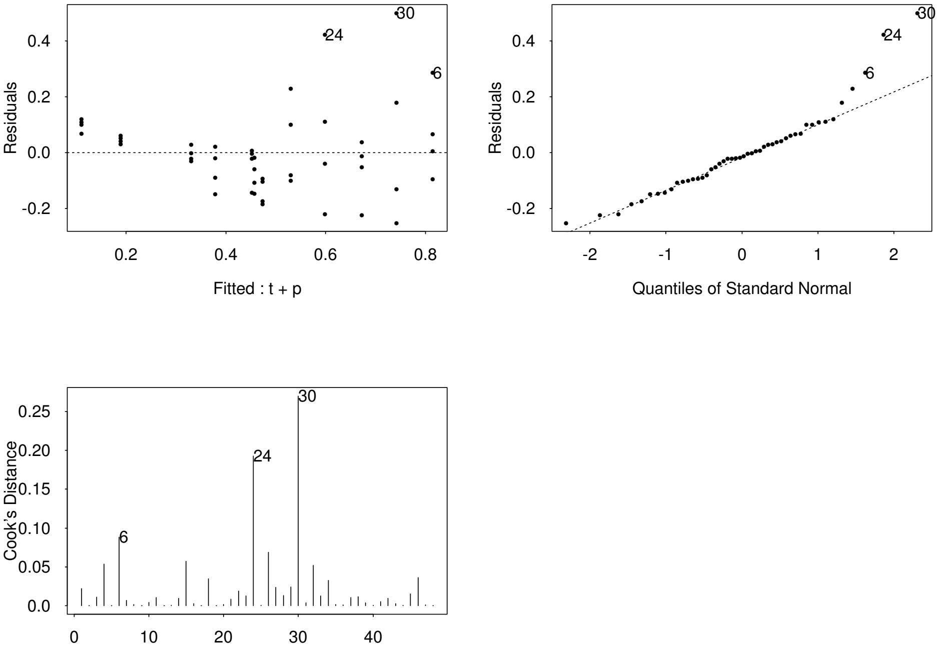

Figure 1: Residual plots for the response survival times

The chosen model $T + P$ is now written as

$$\frac{1}{Y_i} = \mu + \alpha(p_i) + \beta(t_i) + \epsilon_i \qquad i = 1, \ldots, 48. \tag{2}$$

Based on the observed data, our least squares estimates, together with their standard errors and 95% confidence intervals are

Parameter	Estimate	Standard error	Confidence interval
$\hat{\mu}$	2.698	0.174	(2.346,3.050)
$\hat{\alpha}$(I)	0		
$\hat{\alpha}$(II)	–0.469	0.174	(0.117,0.820)
$\hat{\alpha}$(III)	–1.996	0.174	(1.645,2.348)
$\hat{\beta}$(A)	0		
$\hat{\beta}$(B)	–1.657	0.201	(–2.064,–1.251)
$\hat{\beta}$(C)	–0.572	0.201	(–0.978,–0.166)
$\hat{\beta}$(D)	–1.358	0.201	(–1.765,–0.952)

The unbiased estimate for σ^2 is $0.4931^2 = 0.2431$.

Qualitatively this model is much the same. P and T effect survival time independently (no interaction). There is now a more marked distinction between Poisons I and II, with Poison II having an estimated expected death rate 0.47 deaths/10 hours greater than Poison I. Poison

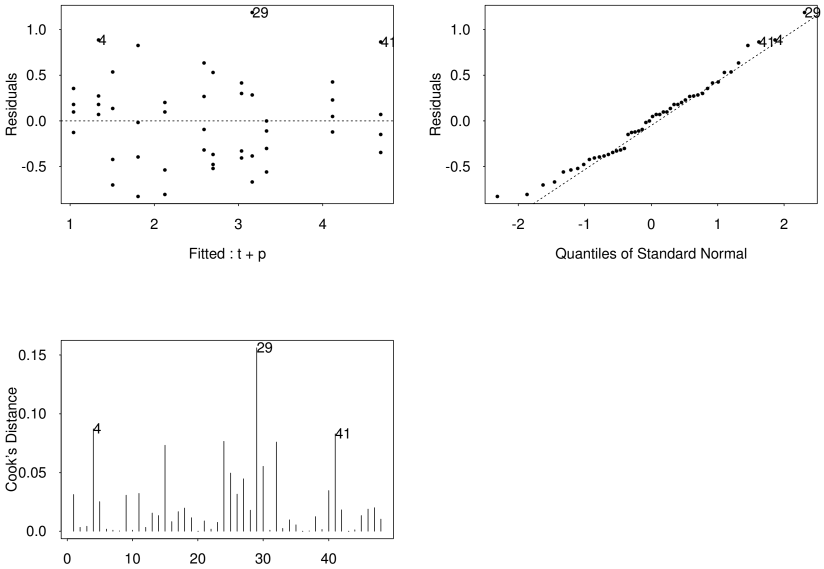

Figure 2: Residual plots for the death rate response.

III remains the most lethal with estimated expected death rate 2.00 greater than Poison I and 1.53 greater than Poison II. The treatments can still be ranked BDCA, in order of survival time (increasing death rate), with the expected difference between B and A being 1.66 deaths/10 hours.

The R^2 coefficient is a much more satisfactory 84.4%, and the residual plots look fine. Altogether model (2) seems much better than model (1).