Survival Times of Liver Surgery Patients

Arden Miller

Executive Summary

The following model was identified as being suitable for predicting the survival times of liver surgery patients:

$$\text{time} = \exp\Big(0.6937 + 0.3090\ \text{clot} - 0.0115\ \text{clot}^2 + 0.0214\ \text{prog} + 0.0220\ \text{enz}\Big)$$

This model should have good predictive power over the ranges of clot (2.6 to 11.2), prog (8 to 96), and enz (23 to 119) contained in the data.

1 The Data

The data consists of measurements that were take on a random sample of 79 patients that underwent a particular type of liver surgery. The survival times of the patients (in days) and the values of four indices, which were thought to be possible predictors of survival time, were recorded for each patient.

clot A score that measures blood clotting
prog A prognostic index, that includes age
enz An enzyme function test score
liv A liver function test score

Table 1 gives the five number summary for each of the variables in the data set. The survival times are positively skewed. Three quarters of the patients survived for 216 or less days, but a few patients survived much longer - up to 830 days. The four indices all seem to be distributed in a fairly symmetric manner.

Table 1: Summaries of the variables in the dataset

	clot	prog	enz	liv	time
minimum	2.60	8.0	23	0.74	34.0
lower quartile	5.15	51.5	64	1.94	103.5
median	5.80	61.0	77	2.58	148.0
upper quartile	6.70	74.0	88	3.23	216.0
maximum	11.20	96.0	119	6.40	830.0

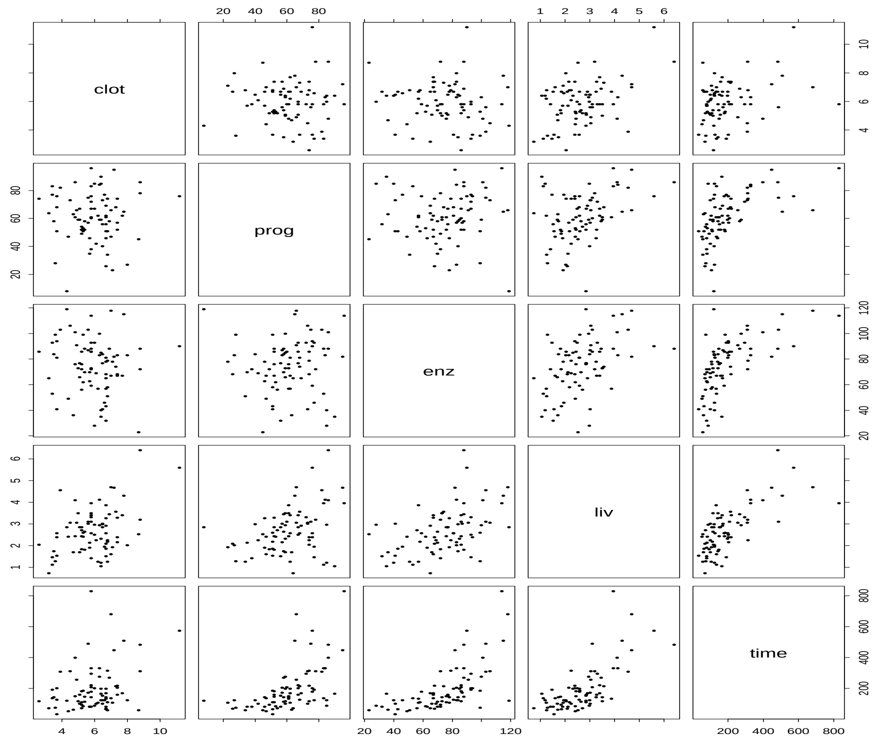

Figure 1: Pairwise scatter plots of the variables

Figure 1 contains pairwise scatter plots of the data. There is a clear relationship between survival times and each of the indices. For each index, longer survival times tend to occur with higher values of the index. This trend is most pronounced for the enzyme test score (enz) and the liver function test score (liv). These plots also indicate that the function test score (liv) is positively correlated with each of the other three indices (clot, prog, and enz). There is no indication that the variables clot, prog, and enz are correlated with each other. These plots also reveal a few observations that have unusual values for the explanatory variables. The plot of clot versus liv indicates two observations that have an unusually large values of both these indices.

2 Predicting the survival time of patients

In order to obtain a linear model it was necessary to model log(time) rather than survival time itself. A suitable model was found to be

$$\log(\text{time}) = 0.6937 + 0.3090\ \text{clot} - 0.0115\ \text{clot}^2 + 0.0214\ \text{prog} + 0.0220\ \text{enz}$$

Over 95% of the variability in the log survival times is explained by this model which indicates good predictive power. Notice that this model does not contain the liver function test score (liv). It was found that adding the variable liv to the above model did not appreciably increase its predictive ability. This indicates that the information liv contains about survival times is also contained in the other three explanatory variables.

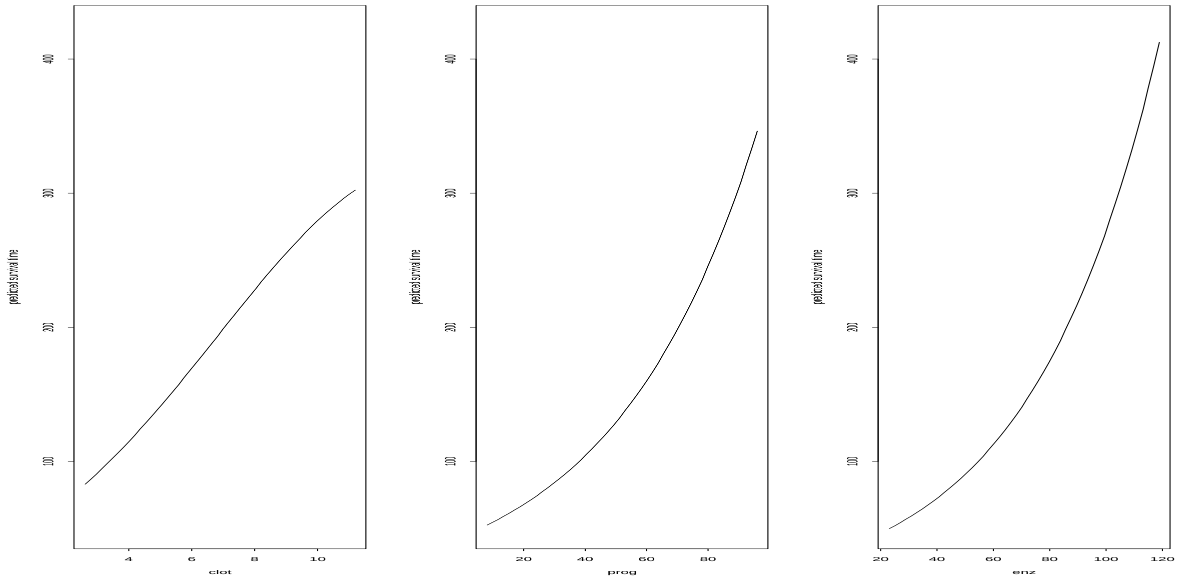

Figure 2: The effects of clot, prog and enz on expected survival time.

To get estimates of survival time, the exponential function is applied to the previous model:

$$\text{time} = \exp\Big(0.6937 + 0.3090 \text{ clot} - 0.0115 \text{ clot}^2 + 0.0214 \text{ prog} + 0.0220 \text{ enz}\Big)$$

The effects of each of the 3 explanatory variables on the estimated survival time of patients are summarised in Figure 2. In each plot the values of one of the explanatory variables is varied through its range while the other 2 regressors are set to their median values. The effect of clot on expected survival time is close to linear while the effects of prog and enz show exponential relationships. For all 3 regressors the expected survival time increases as the value of the regressor increases. It appears that the enzyme function test score (enz) has a slightly larger effect on survival time than the prognostic index (prog) which, in turn, has a slightly larger effect than the blood clotting score (clot).

This predictive model is valid for values of clot, prog and enz within the ranges for the data (see Table 1). The precision of the estimated survival times generated by this model will depend on the values of clot, prog and enz. In general the estimates will be more precise when the values of clot, prog and enz are close to or smaller than their mean values: mean(clot) = 5.9, mean(prog) = 61.1 and mean(enz) = 74.8. For patients who have clot = 5.9, prog = 61.1 and enz = 74.8, we can say, with 95% confidence, that the mean survival time is between 154 and 165 days. As the values of clot, prog and enz are increased the width of such an interval will increase. If we create such intervals for the 79 patients in our dataset then the mean width of those intervals is 23 days and 90% are less than 43 days wide. However, there is one observation which has an unusually wide interval (220 days). This corresponds to the patient who had a much higher score for clot (11.2) than any other patient. It indicates that the model will not, in general, produce precise predictions for unusually large values of the explanatory variables (clot, prog and enz).

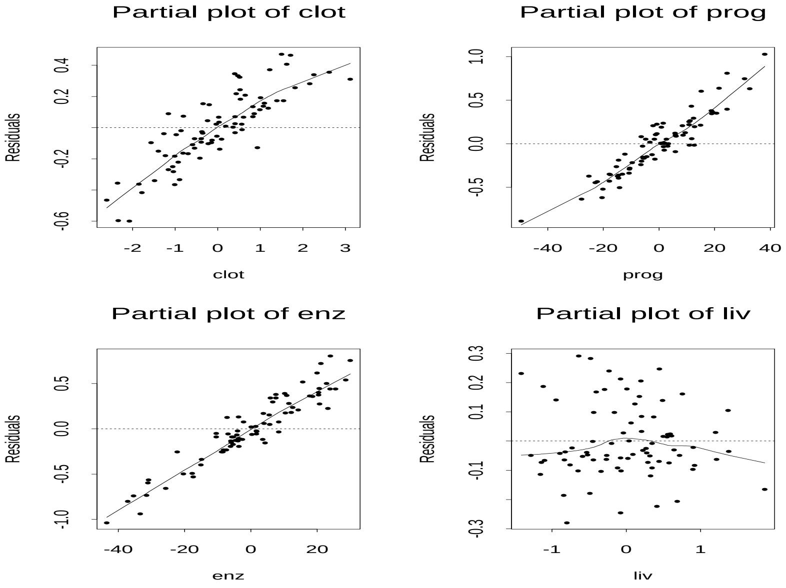

Figure 3: Partial regression plot using log(time) as the response

Statistical Appendix

The main goal of this assignment was to identify a suitable model for predicting the survival time of patients. I started by fitting the model

$$\mathsf{time} = \beta_0 + \beta_1 \mathsf{clot} + \beta_2 \mathsf{prog} + \beta_3 \mathsf{enz} + \beta_4 \mathsf{liv}.$$

Diagnostic plots indicated some problems with this model. Most importantly, the plot of residuals versus fitted values indicated that the response surface was not linear. This plot also indicated a "funnel effect" and the normal probability plot of residuals indicated a problem with the normality assumption. I decided to try transforming the response since this will affect linearity, constant variance, and normality. From the plot of residuals versus fitted values it is apparent that the power of Y needs to be reduced to achieve linearity.

I tried several different power transformations and chose $\log(Y)$ since it did a good job of producing a linear regression surface. Partial regression plots and ace plots were used to determine if the all explanatory variables are needed and whether they need to be transformed (see Figures 3 and 4). These plots suggest that liv is not needed in the model and that a transformation of clot and possibly prog should be considered. I tried fitting the model for log(time) that contained quadratic functions of clot and prog as well as enz and liv. The t-test's for this model indicate that the squared term is needed for clot but not for prog and that liv is not needed in the model. Therefore, I tried the model

$$\log(\mathsf{time}) = \beta_0 + \beta_1 \mathsf{clot} + \beta_{11} \mathsf{clot}^2 + \beta_2 \mathsf{prog} + \beta_3 \mathsf{enz}$$

The diagnostic plots for the fitted model are given in Figure 5. These indicates that for this model the assumptions of a planar regression surface and constant variance are reasonable. The normal probability plot of residuals indicates that the normal assumption is not satisfied. However, I could not find an alternative model that satisfied all of the assumptions (linearity, constant variance, and normality). Since Normality is the least important assumption I decided

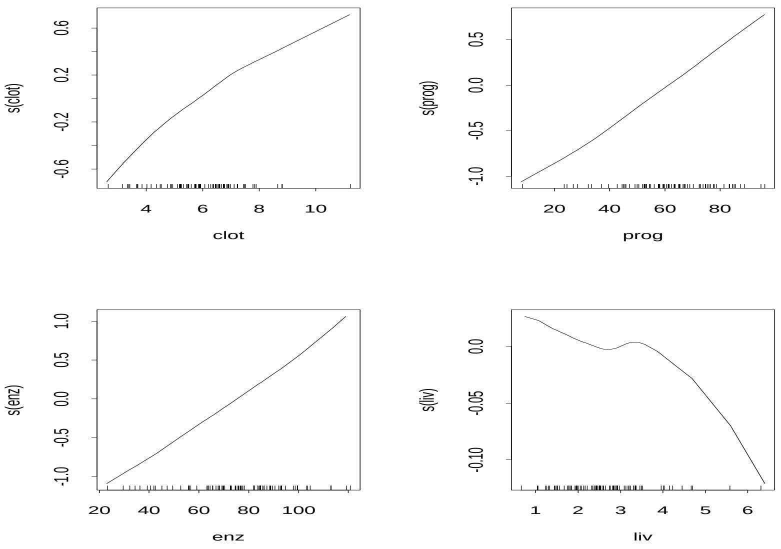

Figure 4: Ace plots using log(time) as the response

to use this model. The lack of Normality may affect the validity of prediction intervals for individual observations but the confidence intervals for the mean response should be okay.

Note, for the model I put in my report I used the actual values of clot and clot^2 and not the scaled versions generated by the poly function. This was done to make it easier to explain the model to the hospital administrator. Given that clot only ranges from 2.6 to 11.2 there should not be any difficulty with this.

There were several points that were flagged as being influential.

```
> infl[c(9,22,28,32,38),]
   (Intercept)       clot I(clot^2)         prog         enz     dffits
 9   0.2864106 -0.2324582 0.2281984 -0.128144345 -0.29352086 -0.4826888
22   0.7833969 -0.7430520 0.6014595  0.457651265 -0.58622668  1.2522716
28   0.3142485 -0.4692440 0.5478061 -0.001266616  0.03658804  0.6567453
32   0.1533856 -0.1050876 0.1372062 -0.096377535 -0.21168507  0.3185533
38   0.2465110 -0.1557852 0.1210066 -0.705777789  0.44122187  0.8943759
   cov.ratio    cooks.d      hats
 9 0.6801474 0.04287476 0.0304217
22 0.5923790 0.27522359 0.1215982
28 3.2050598 0.08719177 0.6708289
32 1.2339026 0.02042729 0.1628529
38 1.1008266 0.15566587 0.2076860
```

However, deleting these points does not appear to have a large effect on the fitted model. Figure 6 indicates that for the most part the fitted values are nearly the same. The biggest departure is noted for observation 28 which corresponds to a patient who has an usually large value for clot. Since deleting these observations doesn't have a large effect on the predicted values, I decided to only report results for the model based on the full data set.

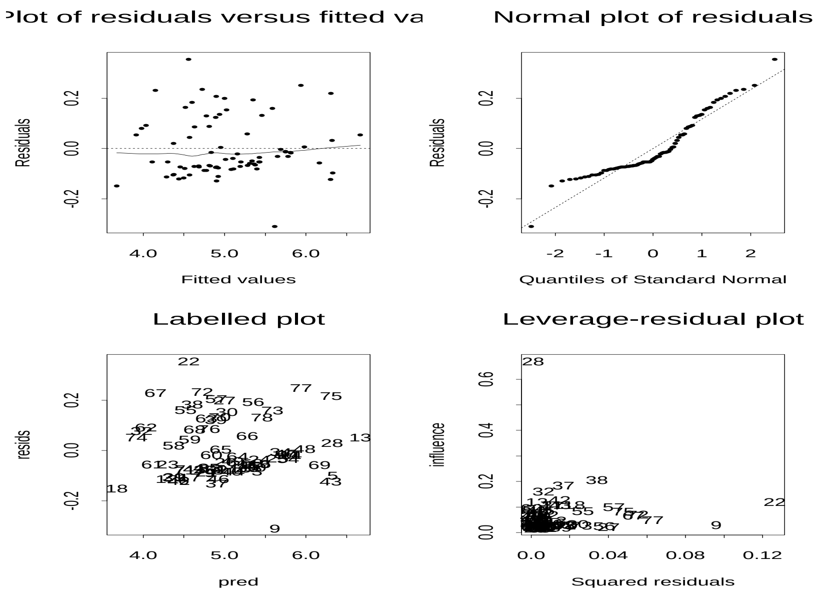

Figure 5: The diagnostic plots for the final model

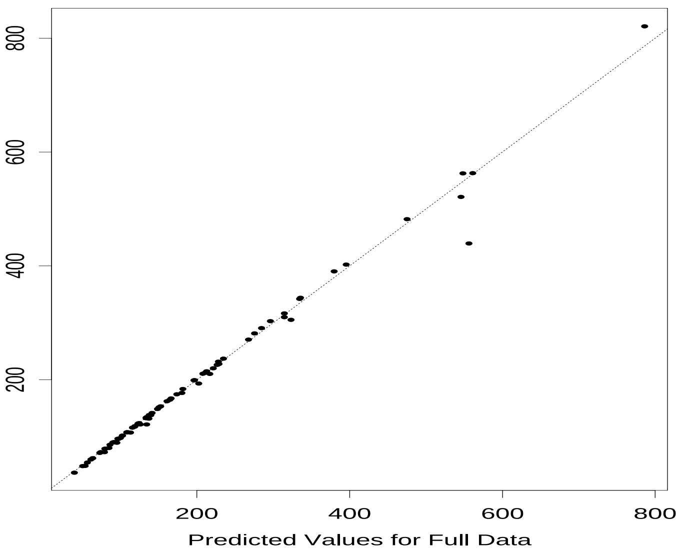

Figure 6: Fitted values: reduced data model versus full data model

475.330 Assignment 2: Marking Guide

This assignment asks the students to analyse data collected on the survival times of patients who received liver operations. The main objective is for them to identify a model that can be used to predict survival times and to explain their model to the hospital administrator.

Report	13 marks
Statistical Appendix	7 marks
Total	20 marks

Report for hospital administrator (13 Marks)

This part of the assignment should describe their model in terms that the hospital administrator can understand. They should discuss the precision of the estimates.

- Presentation - 5 marks for a generally well laid out, coherent report. The reader should not have to search for the important parts among a lot of details. Look for good use of graphs and clear explanations of the model they have chosen. At some point in their report (i.e. an Executive Summary) they should summarise their main findings in a short paragraph. Give: (1) 5 marks for a clear, precise report, that is easy to follow (2) 3 marks if the report is difficult to follow or contains a lot of unnecessary detail, or would be difficult for a non-statistician to understand, (3) 1 mark if it would be very difficult for anyone to understand.
- Content - 8 marks for a report that contains the following
  - A predictive model and a short discussion/explanation of their model.
  - A discussion of the precision of predictions made using their model. To do this properly they will need to make some reference to confidence intervals.
  - A indication of the limitations of their model - ranges for the explanatory variables
  - A discussion of how the explanatory variables are related to the response and each other. I think the main point they should make is that liv clearly has a strong relation to survival time but most people will find that it is not required in their predictive model. They should explain that the information in liv is also contained in the other regressors. Give extra credit if they find other interesting aspects of this data that I haven't.

Statistical Appendix (7 Marks)

This appendix should outline the reasons that they came to the conclusions they presented in the first part of the analysis. They are not required to give a detailed account of everything they did but they should outline why they chose the model they did. Several models are possible for this data. Most people will choose one of the following:

$$\log(\mathsf{time}) = \beta_0 + \beta_1\ \mathsf{clot} + \beta_{11}\ \mathsf{clot}^2 + \beta_2\ \mathsf{prog} + \beta_3\ \mathsf{enz}$$

$$\log(\mathsf{time}) = \beta_0 + \beta_1\ \mathsf{clot} + \beta_2\ \mathsf{prog} + \beta_3\ \mathsf{enz}$$

Some models that use power transformations for the response that are close to 0 (i.e. 1/3, 1/4 ...) produce quite reasonable models as well. Each student should produce some diagnostic plots for the model they have chosen. Make sure that there model does a reasonable job of satisfying linearity and constant variance. Don't be too concerned if there is some evidence of non-Normality. They should also have looked for outliers, high leverage, and influential points (they may have discussed these in their report.

Note:

- Include short comments indicating why a student has lost marks.