Chapter 11

Statistical Strategy and Model Uncertainty

11.1 Strategy

Thus far we have learnt various tactics

1. *Diagnostics:* Checking of assumptions: constant variance, linearity, normality, outliers, influential points, serial correlation and collinearity.

2. *Transformation:* Transforming the response — Box-Cox, transforming the predictors — tests and polynomial regression.

3. *Variable selection:* Stepwise and criterion based methods

What order should these be done in? Should procedures be repeated at later stages? When should we stop?

I would recommend *Diagnostics* $\rightarrow$ *Transformation* $\rightarrow$ *Variable Selection* $\rightarrow$ *Diagnostics* as a rudimentary strategy. However, regression analysis is a search for structure in data and there are no hard-and-fast rules about how it should be done. Regression analysis requires some skill. You must be alert to unexpected structure in the data. Thus far, no one has implemented a computer program for conducting a complete analysis. Because of the difficulties in automating the assessment of regression graphics in an intelligent manner, I do not expect that this will accomplished soon. The human analyst has the ability to assess plots in light of contextual information about the data.

There is a danger of doing too much analysis. The more transformations and permutations of leaving out influential points you do, the better fitting model you will find. Torture the data long enough, and sooner or later it will confess. Remember that fitting the data well is no guarantee of good predictive performance or that the model is a good representation of the underlying population. So

1. Avoid complex models for small datasets.

2. Try to obtain new data to validate your proposed model. Some people set aside some of their existing data for this purpose.

3. Use past experience with similar data to guide the choice of model.

Data analysis is not an automatic process. Analysts have personal preferences in their choices of methodology, use software with varying capabilities and will interpret the same graphical display differently. In comparing the competing analyses of two statisticians, it may sometimes be possible to determine that one

analysis is clearly superior. However, in most cases, particularly when the analysts are experienced and professionally trained, a universally acceptable judgment of superiority will not be possible.

The same data may support different models. Conclusions drawn from the models may differ quantitatively and qualitatively. However, except for those well-known datasets that circulate endlessly through textbooks and research articles, most data is only analyzed once. The analyst may be unaware that a second independent look at the data may result in quite different conclusions. We call this problem *model multiplicity*. In the next section, we describe an experiment illustrating the depth of this problem.

11.2 Experiment

In Fall 1996, I taught a semester length masters level course in applied regression analysis to 28 students. Towards the end of the semester, I decided to set an assignment to test the students ability in building a regression model for the purposes of prediction. I generated regression data with a response y and five uncorrelated predictors and $n = 50$ from a model known only to me which was:

$$1/y = x_1 + 0.57x_1^2 + 4x_1x_2 + 2.1\exp(x_4) + \varepsilon$$

where $x_1 \sim U(0,1)$,$x_2 \sim N(0,1)$,$1/x_3 \sim U(0,1)$,$x_4 \sim N(1,1)$, $x_5 \sim U(1,3)$ and $\varepsilon \sim N(0,1)$.

I asked students to predict the mean response at 10 values of the predictors which I specified. I also asked them to provide a standard error for each of their predictions. The students understood and were reminded of the distinction between the standard error for the mean response and for a future observed value. The students were told that their score on the assignment would depend only on the closeness of their predicted values and the true values and on how closely their standard errors reflected the difference between these two quantities. Students were told to work independently.

For a given student's input, let p_i be their prediction, t_i be the true value and s_i be the standard error where $i = 1,\ldots 10$. To assess their prediction accuracy, I used

$$\sum_{i=1}^{10}\left(\frac{p_i - t_i}{t_i}\right)^2$$

whereas to measure the "honesty" of their standard errors, I used

$$\frac{1}{10}\sum_{i=1}^{10}\left|\frac{p_i - t_i}{s_i}\right|.$$

We'd expect the predicted value to differ from the true value by typically about one standard error if the latter has been correctly estimated. Therefore, the measure of standard error honesty should be around one.

1.12	1.20	1.46	1.46	1.54	1.62	1.69
1.69	1.79	3.14	4.03	4.61	5.04	5.06
5.13	5.60	5.76	5.76	5.94	6.25	6.53
6.53	6.69	10.20	34.45	65.53	674.98	37285.95

Table 11.1: Prediction accuracy

The prediction accuracy scores for the 28 students are shown in Table 11.1. We see that one student did very poorly. An examination of their model and some conversation revealed that they neglected to backtransform their predictions to the original scale when using a model with a transform on the response.

Three pairs of scores are identical in the table but an examination of the models used and more digits revealed that only one pair was due to the students using the same model. This pair of students were known associates. Thus 27 different models were found by 28 students.

The scores for honesty of standard errors are shown in Table 11.2. The order in which scores are shown correspond to that given in Table 11.1.

0.75	7.87	6.71	0.59	4.77	8.20	11.74
10.70	1.04	17.10	3.23	14.10	84.86	15.52
80.63	17.61	14.02	14.02	13.35	16.77	12.15
12.15	12.03	68.89	101.36	18.12	2.24	40.08

Table 11.2: Honesty of standard errors - order of scores corresponds to that in Table 11.1.

We see that the students' standard errors were typically around an order of magnitude smaller than they should have been.

11.3 Discussion

Why was there so much model multiplicity? The students were all in the same class and used the same software but almost everyone chose a different model. The course covered many of the commonly understood techniques for variable selection, transformation and diagnostics including outlier and influential point detection. The students were confronted with the problem of selecting the order in which to apply these methods and choosing from several competing methods for a given purpose.

The reason the models were so different was that students applied the various methods in different orders. Some did variable selection before transformation and others the reverse. Some repeated a method after the model was changed and others did not. I went over the strategies that several of the students used and could not find anything clearly wrong with what they had done. One student made a mistake in computing their predicted values but there was nothing obviously wrong in the remainder. The performance on this assignment did not show any relationship with that in the exams.

The implications for statistical practice are profound. Often a dataset is analyzed by a single analyst who comes up with a single model. Predictions and inferences are based on this single model. The analyst may be unaware that the data support quite different models which may lead to very different conclusions. Clearly one won't always have a stable of 28 independent analysts to search for alternatives, but it does point to the value of a second or third independent analysis. It may also be possible to automate the components of the analysis to some extent as in Faraway (1994) to see whether changes in the order of analysis might result in a different model.

Another issue is raised by the standard error results. Often we use the data to help determine the model. Once a model is built or selected, inferences and predictions may be made. Usually inferences are based on the assumption that the selected model was fixed in advance and so only reflect uncertainty concerning the parameters of that model. Students took that approach here. Because the uncertainty concerning the model itself is not allowed for, these inferences tend to be overly optimistic leading to unrealistically small standard errors. Methods for realistic inference when the data is used to select the model have come under the heading of *Model Uncertainty* — see Chatfield (1995) for a review. The effects of model uncertainty often overshadow the parametric uncertainty and the standard errors need to be inflated to reflect this. Faraway (1992) developed a bootstrap approach to compute these standard errors while Draper (1995) is an example of a Bayesian approach. These methods are a step in the right direction in that they reflect the uncertainty in

model selection. Nevertheless, they do not address the problem of model multiplicity since they proscribe a particular method of analysis that does not allow for differences between human analysts.

Sometimes the data speak with a clear and unanimous voice — the conclusions are incontestable. Other times, differing conclusions may be drawn depending on the model chosen. We should acknowledge the possibility of alternative conflicting models and seek them out.