Exploratory model analysis

with R and GGobi

Hadley Wickham

2006-12-18

1 Introduction

Why do we build models? There are two basic reasons: explanation or prediction [Ripley, 2004]. Using large ensembles of models for prediction is commonplace, but is rarely used for explanation, where we typically choose one "best" model. When there are several equally good models, it is common sense to look at them too, but can the "bad" models tell us something as well?

This paper describes exploratory model analysis for ensembles of linear models, where we look at all possible main effects models for a given dataset (or a large subset of these models). This gives greater insight than looking at any small set of best models alone: an ensemble of many models can tell us more about the underlying data than any individual model alone.

This paper builds heavily on exploratory modelling analysis as introduced by Unwin et al. [2003], but rather than describing different types of useful plots it is organised around different levels of data. We will assume we have m models describing a data set with n observations and p variables. If all possible main effects models are fit, there will be $2^p - 1$ models in the ensemble. After fitting the models, we compute summary statistics on four levels:

- Model level: model fit statistics. m observations.
- Coefficient level: coefficient estimates on various scales. $m \times p$ observations.
- Residual level: residuals and influence measures. $m \times n$ observations.
- Observation level: the original data, plus summaries of residual behaviour. n observations.

These ideas are implemented in R package, `meifly`, described briefly in appendix A. Meifly (**m**odels **e**xplored **i**nteractively) uses R to fit models, and then displays them using GGobi [Swayne et al., 2003]. This paper presents a combination of static graphics produced by `ggplot` [Wickham, 2006], and screenshots from GGobi, but my exploration was performed almost entirely in GGobi. The static graphics are attractive and easy to supplement with extra information, but do not allow the rich exploration of structure that interactive graphics do.

To illustrate these ideas we will use a data set on fertility in French-speaking Swiss provinces in the late 1800's [Mosteller and Tukey, 1977]. We are interested in predicting fertility based on proportional of agricultural workers, average performance on an army examination, amount of higher education, proportion of Catholics and infant mortality. There are 47 observations and six predictor variables, giving an ensemble containing 31 ($2^5 - 1$) models. The data itself is rather irrelevant but I have not yet found a dataset which presents so many interesting features so obviously.

2 Model level

Our summaries start with the models themselves. For each model we record:

- Degrees of freedom.
- R^2 and $AdjR^2$.
- Log-likelihood, AIC and BIC (defined as log-likelihood minus the correction for number of parameters, so as to be in same direction as the other statistics)

These statistics let us investigate model fit. While theory is available for testing differences between nested models, here we are using these statistics heuristically, looking at how increasing degrees of freedom improves model fit (beyond what we might expect from an adding an unrelated variable). For this purpose, it is not necessary to look at all five statistics, but it may be useful pedagogically.

Figure 1 shows the how model fit varies across models for the Swiss dataset. The pattern is very similar across all statistics, and suggests that either a four or five variable model is the "best".

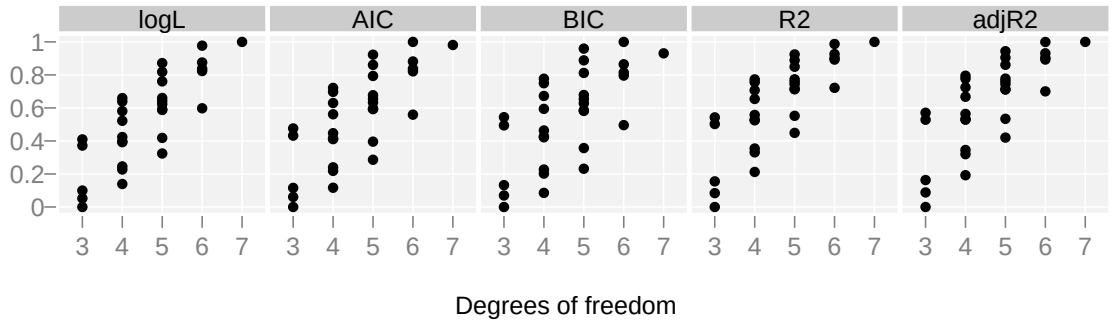

Figure 1: Model summary statistics, each scaled to $[0, 1]$ to aid comparison. Degrees of freedom includes one df for estimating $\hat{\sigma}^2$. The intercept only model is not displayed.

3 Coefficient level

Each model contains between 1 and p variables, and for each variable we calculate the following information:

- Raw coefficient. Useful when all covariates are on the same scale.
- Standardised coefficient (from model fit to data standardised to mean 0, standard deviation 1). Useful as measure of relationship strength; comparable as on a common scale.
- t-value, and absolute t-value. Allow us to assess the significance and direction of the relationship between the response and predictor.

One simple way to use this data is compare the estimates given by different models, as in Figure 2. This lets us see the distribution of individual estimates, but we don't see how they

vary together for a given model. We can do this in two ways: dynamically, using linking in GGobi; or statically, by connecting all observation from a single model. These two techniques are illustrated in Figures 3 and 4.

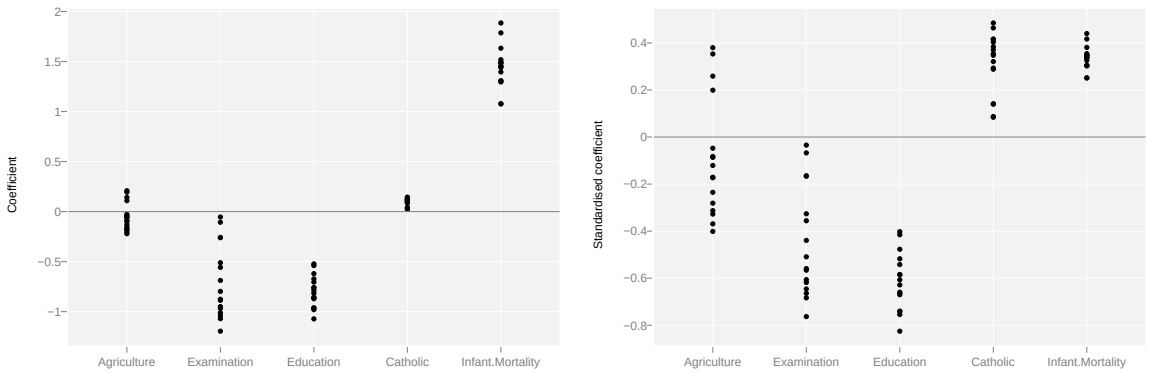

Figure 2: (Left) Raw coefficients are only useful when variables are measured on a common scale. (Right) In this case, as with most data sets, looking at the standardised coefficients is more informative, as we can judge relative strength on a common scale.

We can also use this graphic in another way: to explore the variance-covariance matrix of the predictors. There are two interesting examples of this in Figure 3. Firstly, the t scores for infant mortality are very similar over all models. This indicates that this variable is largely independent of the other explanatory variables.

Another interesting phenomenon is the behaviour of the agriculture variable. For four of the models the relationship between fertility and agriculture is positive, while for all others it is negative. Using GGobi (not shown) we can brush these four models and compare them to the others to determine what covariate affects this relationship: it is education.

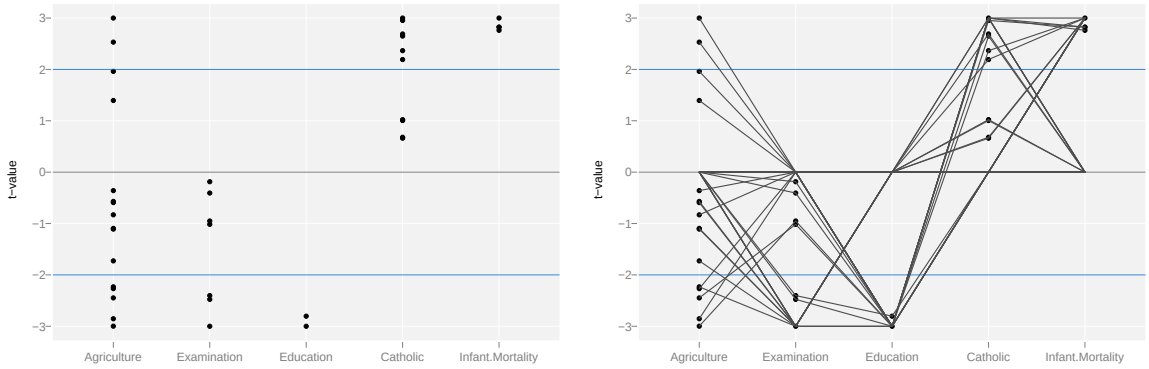

Figure 3: (Left) Scatterplot of t-value vs variable. t-values greater than three are truncated to three. (Right) We turn this into a type of parallel coordinates plot by connecting all coefficients from the same model with a line. If the variable is not included in a model, its value is set to 0 for all coefficient statistics.

We can use GGobi to connect the coefficient level data with the model level data. In Figure 4 we brush the two best models and look at their coefficients. The additional variable in five variable model is examination, and it is very difficult to see any meaningful difference amongst the t-values for other variables. You can also brush in the opposite direction, by selecting groups of coefficients and then seeing which model(s) they correspond to.

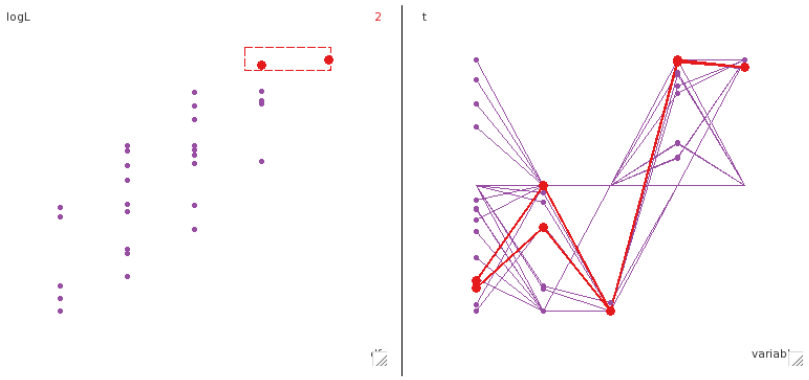

Figure 4: (Left) Scatterplot of log-likelihood vs degrees of freedom. The two best models have been brushed. (Right) Parallel coordinates plot of t-value vs variable. The same two models are highlighted in this plot.

4 Residual level

Whether or not a variable is an "outlier" (or influential, or poorly modelled) is conditional on the model. By looking at the following statistics for each model we can explore how model (mis)specification affects the individual predictions:

- Actual value, fitted value, and residual.
- Studentised residual.
- Influence measures: DFFITS, covariance ratio, Cook's distance and hat values.

Residuals can be linked to both models and observations, so we can use this data to compare model performance or look for observations that are hard to fit.

In Figure 5 we discover two observations with interesting patterns of influence. These observations are explored further in the next section. It would also be interesting to explore how model choice affects the influence measures of these observations. For example, about half the pink points are unusual, probably indicating the inclusion or exclusion of a single variable makes the difference.

5 Observation level

The observation level summaries let us get back to the data summarised by the models, and allows us to put our findings in the context of the original data (important!). The original data is supplemented with residual summary statistics: the mean and standard deviation of the absolute studentised residuals. These provide a quick way to look for observations that are fit poorly on average, or are sensitive to model specification.

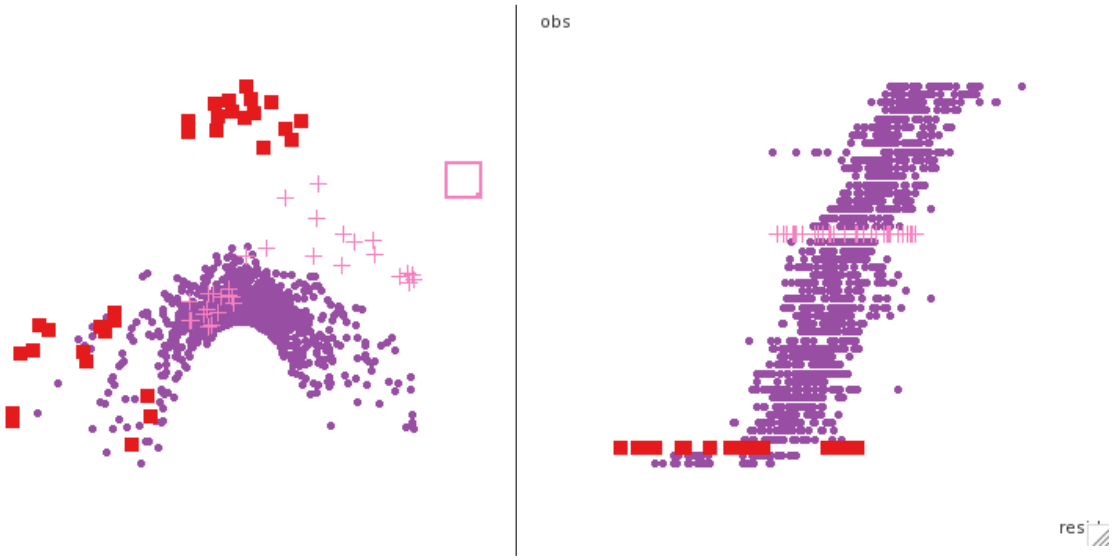

Figure 5: (Left) Scatterplot of covariance ratio vs. DFFITS. (Right) Scatterplot of observation (sorted by mean value) vs residual value. Two unusual groups have been highlighted with red squares and pink plusses. These correspond to two observations.

Figure 6: (Left) Scatterplot of education vs infant morality, points coloured as in Figure 5. (Right) Density plots for each variable reveal that V. De Geneve is has an unusual value on all variables apart from infant morality, while La Vallee is only unusual on the infant mortality variable.

For the unusual points discovered in Figure 5 we can investigate where they appear in the original data space, Figure 6. They appear to be outliers, and it would be wise to refit the model with these observations omitted and investigate how the models change. Providing simple graphical summaries of these changes is an interesting area for further research.

6 Conclusion

This paper has just scratched the surface of exploratory model analysis. I have shown some examples of using the four levels of summary data to gain insight in to a dataset and ensemble of models, and this seems useful for both pedagogy and practical analysis.

There are many possible ways to extend this work. There are other types of models we would like to explore. Generalised linear models and robust linear models fit into to this framework easily. Other families of models, eg. trees or non-linear models, will need a little more effort, particularly at the coefficient level: we can compare a tree-based regression to a linear model at the model level (AIC, BIC), and residual level but there is no meaningful way to compare the coefficients of a linear model to the splits in a regression tree. This raises the more general question of whether or not it is meaningful to compare models across different model classes.

We can also generate ensembles in a different way: by using different datasets with the same model. These datasets could be from bootstraps, or from different subsets of a large dataset, perhaps as a prelude to fitting a mixed model. For these ensembles, the residual and observation level summaries will need to be improved to reflect that an observation might be used zero or many times in a given model.

References

F. Mosteller and J. W. Tukey. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading Mass., 1977.

Brian Ripley. Selecting amongst large classes of models, 2004. Symposium in Honour of David Cox's 80th birthday. http://www.stats.ox.ac.uk/ ripley/Nelder80.pdf.

Deborah F. Swayne, Duncan Temple Lang, Andreas Buja, and Dianne Cook. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444, 2003.

Antony Unwin, Chris Volinsky, and Sylvia Winkler. Parallel coordinates for exploratory modelling analysis. *Computational Statistics & Data Analysis*, 43(4):553–564, 2003. URL `http://rosuda.org/~unwin/AntonyArts/uvwCSDA.pdf`.

Hadley Wickham. *ggplot: An implementation of the Grammar of Graphics in R*, 2006. R package version 0.4.0.

Appendix A: Using the meifly package

The `meifly` package makes it easy to use the techniques described in this paper. Install it using `install.packages("meifly", repos="http://www.ggobi.org/r/")`. There are two ways to create model ensembles: `fitall` which fits all possible model combinations, and `fitbest` which uses the `leaps` package to very rapidly find the n best models for a given number of variables.

Once you you have the ensemble, you can extract the summary statistics at different levels using `coef(x)`, `summary(x)`, `resid(x)`, `summary(resid(x))`. Finally, `ggobi(x)` will load the all model summaries into GGobi, linked together by categorical variables.

A couple of tips when using brushing in GGobi: make sure you're only brushing colour, and remember to change the linking variable to either "model" or "obs".