STAT 545 - Data Analysis (Term 1, 2005-06)

ASSIGNMENT

NOTE: More questions will be added as we cover material.

NOTE: Most of the problems are deliberately open-ended, and give you the opportunity to investigate as you best see fit. One tradeoff is as follows. I will not assign a large number of problems, but in return I will be looking for well-documented solutions. At a minimum, you should be reporting your findings in complete sentences/paragraphs (not computer-code fragments!), and using tables and/or figures to summarize empirical work as appropriate. One guideline that cuts across all scientific work is that you should provide enough detail so that an interested reader could replicate what you have done. In terms of using mathematical notation versus words, I'm not fussed. That is, some of you will be more comfortable using less/more mathematical notation than others, and that's fine. Clarity can be achieved either way!

1. We discussed kernel density estimation (KDE) in class. One situation where kernel density estimation may not be so straightforward is when the datapoints $y_1, \ldots, y_n$ are measurements of an inherently *positive* quantity, but some of the measured values are very close to zero. Three strategies for such a situation are

- (i) Do nothing out of the ordinary. Apply KDE to $y_1, \ldots, y_n$ in the usual way.
- (ii) Apply usual KDE to the sample of size $2n$ consisting of $-y_1, \ldots, -y_n$ as well as $y_1, \ldots, y_n$. Then take only the postive half of the density estimate, suitably rescaled.
- (iii) Apply a transformation (log is the obvious candidate) to the data. Apply KDE to the transformed data. Take the density estimate and transform it back to the original scale.

Briefy investigate the performance of these three strategies. Construct some datasets where the values are inherently positive but some are close to zero (for instance, simulate from some appropriate distributions). Display the density estimates arising from the three strategies. Comment briefly on what you find, particularly in terms of which strategies seems to work well or poorly.

2. In the linear regression context we discussed the informal classification of a data point as having 'high leverage' if the corresponding diagonal element of the 'hat' matrix exceeded $3(p/n)$, where the design matrix is $n \times p$. One might also use an informal classification of a data point being an 'outlier' if the magnitude of its studentized or standardized residual exceeds some value.

Carry out a small *simulation study* to study the proportion of points which are classified as high-leverage, outliers, or both. In particular, assess how these proportions vary with (i) the shape of the distribution of the (one or more) predictor variables, and (ii), the shape of the distribution of the error terms (i.e., the distribution of $Y|X$). Are your results in accord with your intuition?

Please be clear in the description of how you conducted your simulations. A rule-of-thumb is to give enough detail that an interested reader could reproduce your simulations from your descriptions.

3. Take a look at the "Scottish hill races" example in Sec. 6.3 of the text. The authors use weighting to reflect the fact that the times for shorter races ought to be more predictable, but they comment that transformations might also be used to acheive this end. Fit an appropriate linear model that uses transformations (but not weighting) to relate *time* to *dist* and *climb*. Does your model seem to fit the data better or worse than the author's model?

4. [Oct. 3] Consider simulating data consisting of n realizations of $(Y, X_1, \ldots, X_p)$ from a logistic regression model, where Y is binary and each X_j is either binary or categorical. As we discussed in class, we could also view such a dataset as comprised of m binomial responses, where m is the number of distinct X patterns manifested in the data. Set-up your data-generating mechanism such that m is (likely to be) much smaller than n.

(a) Try fitting this glm to your simulated dataset, first as n binary responses and then as m binomial responses. Check that you get exactly the same estimated regression coefficients and standard errors, but that the deviance, and the degrees of freedom for the deviance, are different.

(b) Which approximation is better, the χ^2_{n-p} approximation to the deviance based on binary responses or the χ^2_{m-p} approximation to the deviance based on binomial responses? Carry out a simulation study (i.e., repeatedly generate datasets as above) to address this question.

NOTE: STAT students must do 5a and omit 5b. NON-STAT students may elect to do 5a or 5b (but not both!).

5a. [Oct. 5]

(i) Verify that if we have an exponential family distribution and a link function, and proceed to use the Newton-Raphson algorithm (with the Fisher scoring modification discussed in class) to find the MLE, that the algorithm is actually equivalent to a sequence of weighted least-squares fits. Make it clear what the form of the weights and the form of the "response" is at each iteration of the algorithm. Make it clear that the algorithm depends only on the data, the choice of link function, and the variance function (up to a constant of proportionality).

(ii) Verify (as mentioned in class) that if random variable X takes values in $\{0, \ldots, n\}$ and has mean np, then its variance cannot exceed $n^2p(1-p)$.

5b. [Oct. 5] Find an application of generalized linear modelling in your subject-area literature. Briefly describe the application and the modelling carried out. Comment on (i) the advantages (if any) of using GLM rather than LM in this particular problem, (ii) what features of GLM are used (what estimation, testing, and/or residual analysis is carried out), and (iii) whether the modelling and fitting seems to be statistically sound, and (iv) anything you find particularly interesting about the statistical modelling and fitting.

THIS IS CUTOFF #1. Problems 1-5 are due by Wednesday Oct. 19 at the latest. By the way, this is not an invitation to rest on your laurels! New problems will continue to be posted as we cover material.

6. [Oct. 27] Consider the log-linear modelling example in the text (and discussed in class) with the Copenhagen housing data. Recall that interest focusses on how *satisfaction* (a categorical variable with levels low, medium, high) depends on three other categorical variables. Recall that the Poisson modelling of frequencies implicitly fits a model for *satisfaction* as a trinomial (trinary?) response.

As an alternative analysis, try creating a binary reponse variable by combining the low and medium levels of satisfaction together, and then use logistic regression to explain this response in terms of the three explanatory variables. Do the results of this analysis seem to be qualitatively consistent with that of the trinomial analysis. Do the estimated effects from the binary reponse model seem to match those from the trinary response model in an intuitively sensible way?

7. [Oct. 27] Consider the in-class example on bootstrapping to acknowledge the uncertainty in selecting a subset of explanatory variables. There we suggested that (i) applying a model selection scheme to remove some predictors and (ii) refitting to the remaining predictors only gives standard errors [from the "refit"] which are too small.

Check this out empirically and see if the SEs really are too small. That is, repeatedly simulate datasets, apply some model selection scheme, re-fit to included predictors only, and form simple $(\hat{\beta} \pm 1.96 SE[\hat{\beta}])$ confidence intervals from the re-fit. (To be clear here, I'm talking about the regular old likelihood-based SE that the computer spits out - no bootstrapping or anything fancy.) What percentage of the simulated datasets give an interval containing the true parameter value. Is it close to the desired 95%?

NOTE: you will likely want to try several sets of true parameter values. In particular, look at the behaviour of an interval when (i) the true parameter value is zero, and (ii) when it is not zero.

NOTE: you will need to decide what to do when the predictor in question is not selected. Presumably if the true value of β_j is non-zero, but X_j is not selected, then that counts as an instance of the confidence interval (a single point at zero) missing the true parameter value.

NOTE: the following problems are not in the same order as the corresponding material was covered in class.

8. [Nov. 17] Consider the Sitka data (in the "MASS" library and discussed in V & R) that was used in an in-class example (recall this is data on the size of 79 trees at five timepoints). Consider fitting models of the form

$$Size_{it} = \alpha_i + \beta_i t + \gamma_i t^2 + \epsilon_{it}$$

where i indexes tree and t indexes timepoint (because the timepoints are evenly spaced and common across trees, without loss of generality we can 'code' t as $0, 1, \ldots, 4$).

To see if there is a very demonstrable benefit to "borrowing strength" try fitting such a model **using only the data for the first four timepoints**. Do this both in a *separate model* manner where $(\alpha_i, \beta_i, \gamma_i)$ are fixed effects associated with the i-th tree, and in a *hierarchical / mixed / random coefficient* manner where $(\alpha_i, \beta_i, \gamma_i)$ are random effects associated with the i-th tree.

For the two fitted models, compare predictions of tree sizes at the fifth timepoint to the actual sizes. Is there evidence of a tangible benefit of "borrowing strength"? In addition to giving a numerical comparison of predictive performance for the two models, give some graphical summaries, such as plots of predicted versus actual values, and perhaps fitted size versus time relationships for a few selected trees. Also, comment on where most of the "borrowing" seems to be going on - in estimating α's or β's or γ's?

9. [Nov. 28] Consider the South African Heart Disease data available from *www-stat.stanford.edu/∼tibs/ElemStatLearn*

There is a binary response (congestive heart disease) and nine potential explanatory variables. (As an aside, these data were actually collected via a case-control design, but there is a theoretical justification for fiting a model for Y given X, despite the sampling scheme involving X give Y.) As a couple of R hints to get started, read.csv() is better than read.table() for getting a comma-seperated data file into a data frame. Also, you may have to remove the case index (1:462) as a variable from the data frame.

Randomly break the 462 cases into equal-sized training and validation samples. In an attempt to keep us all in sync, let's all use the same split. That is, use

```
set.seed(13)
splt <- sample(c(rep(F, 231), rep(T, 231)))
```

and then let TRUE indicate training set cases, FALSE indicate validation set.

In a somewhat similar spirit to the low-birth-weight example discussed in the text/class, use stepAIC() to select a model for the training data. In particular, start with the model having all nine predictors, then use the stepwise procedure to (perhaps) remove some predictors. Then use the stepwise procedure a second time to see if any interactions between the selected variables should be added. Summarize the fit of the final model to the training data by giving a 2×2 table describing the fitted versus actual Y. Also in the spirit of cross-validation give a comparable table describing how well the fitted model predicts Y from X in the validation data.

Now do the same fitting and prediction but using BIC rather than AIC as the model selection criterion (recall from our earlier example that stepAIC can be adapted to use BIC). Again give 2×2 tables summarizing the fit of the selected model to the training data, and the predictive ability of the fitted model on the validation data.

Comment on your findings, i.e., performance on training data versus validation data, use of AIC versus BIC.

10. [Nov. 28] Consider kernel smoothing of the "motorcycle" data, as we discussed in class. Presuming we have chosen a bandwidth, we might ask the question how precise is the fitted curve as an estimate of the 'real' curve $f(x) = E(Y|X = x)$. The purpose of this question is to check if the bootstrap can help with this. [I confess I haven't tried this, so I will be curious to get your impression of how well this works!]

(a) Try fitting a kernel smooth to these data with different bandwidths, until you get a tradeoff between fit and smoothness that you are happy with, at least 'by eye'.

(b) Generate a bootstrap (X,Y) sample, using an idea for bootstrapping with regression data that we discussed – keep the original X values fixed but resample the residuals to generate new Y values. One can then apply the kernel smoother to the bootstap sample, to get a replicated curve.

(c) Generate a number of such replicated curves and plot them on the same set of axes, to give a visual description of the precision in the original estimate.

(d) If you are feeling ambitious, also try to use this approach to compute and display ‘pointwise’ confidence intervals for $f(x)$, i.e., separate confidence intervals for $f(x)$ at different values of x.

11. [Nov. 30] Find a dataset for which you can try fitting both an additive model and a tree model. Do both techniques lead to a similar qualitative impression of the relationship between the response variable and the predictor variables? Is one of the techniques more ‘satisfying’ than the other in any sense? Comment as you see fit.

THE END!

Problems 6-11 are due by noon on Monday December 19. I have no further wiggle-room than this - the registrar’s office will want grades submitted very soon thereafter.