STAT 538 - Generalized Linear Models (Term 1, 2004-2005)

COURSEWORK

Note: dates are 'posted' dates, not due dates!

SEPT 17 (Straightforward, I think): Try Exercises 3.5, 3.7, and 3.9 from the textbook. For 3.9, also comment on whether this is a sensible model.

SEPT 17 (More open-ended - make what you will of it!): On the theme of 'can we get away with a LM instead of a GLM'...

Say $X > 0$ has some distribution. Say $E(Y|X) = \beta_0 + \beta_1 X$ for some unknown parameters $\beta_0 > 0$ and $\beta_1 > 0$. Consider two variants of this model, one which postulates that $Y|X$ has a normal distribution, the other which postulates that $Y|X$ has a Poisson distribution.

Carry out a small simulation study to compare estimators of β obtained under the two models, *when in fact the Poisson model is the true model.* Does there appear to be a bias in the estimator obtained by assuming the wrong model? Is one estimator more precise than another? Can the estimated precision (i.e., standard error or confidence interval) still be trusted if the wrong model is assumed?

The R functions lm() and glm() can be used to fit the normal and Poisson models respectively.

Now carry out a theoretical study of the large-sample bias and variance of the two estimators (again under the conditions that the Poisson model is the right model). If you want to make your life simpler, restrict consideration to the one-parameter situation where $E(Y|X) = \beta X$. For the normal model you can also get away with pretending that $\sigma^2 = Var(Y|X)$ is known, since the ML (least-squares) estimator of β will be unaffected by this.

There are probably different lines of attack here. One could simply start with expressions for the two estimators and try to work out approximate means and variances. Or one could take a more general approach based on the large-sample theory for maximum likelihood estimators (i.e., Fisher information etc.) The application of this theory in 'right-model' situations is well known, but its application in 'wrong-model' settings is less often discussed in courses. Many discussions of this in books is more complicated than necessary, but came across a nice readable description on the web:

http://www.stat.umn.edu/geyer/5601/examp/sand.pdf

I think I'll review this stuff in class.

SEPT 30: Lately we have been glossing over the fact that many distributions we might want to use for GLMs have an extra nuisance parameter (Normal and Gamma are prime examples). Say you choose to model the response Y with a distribution satisfying $Var(Y_i) = \phi v(\mu_i)$ for some function $v()$, where as usual $\mu_i = E(Y_i)$. Adding a link $g(\mu_i) = x_i^T \beta$ then completes the GLM, yielding a model parameterized by (β, ϕ), where ϕ is called the *dispersion* parameter.

(i) Show that the gamma distribution can be cast in this form.

(ii) Consider treating the dispersion parameter as fixed and known. Argue that the estimate of β will NOT depend on the value of ϕ you choose, but that the standard errors (based on the information matrix) will be affected.

SEPT 30: Exercise 4.2 from the textbook. (It might be helpful to tackle the previous problem before doing this one.) I don't think 'family=Exponential' is supported by the glm function in R, but 'family=Gamma' is.

SEPT 30: Pick a GLM setting and investigate by simulation how large n must be for a confidence interval or hypothesis test to perform well. Recalling that different kinds of asymptotics can be applied (Wald, score, LR). Does changing from one to another make any difference in your context? Lots of open-endedness here...

THIS IS CUTOFF # 1. THE ABOVE PROBLEMS ARE DUE WED. OCT 20 (OR SOONER IS FINE TOO!)

OCT 6: Consider the beetle mortality example in Ch. 7. First, produce a plot of the data (as per Fig. 7.3), but with the fitted probability versus dose curves from the three models superimposed. Is the better fit of the complementary log-log link evident?

Second, consider estimating the *median lethal dose* (LD-50), i.e., the dose of CS_2 (in units of mgl^{-1}) that yields expected mortality of 50%. Give both a point estimate and confidence interval based on each model. Do the three sets of point/interval estimates agree?

What about LD-5? Make some qualitative comment about your findings.

HINT: use what you know about estimating *functions* of parameters.

OCT 18: The text says that the approximate null distribution of the Hosmer-Lemeshow statistic has been found *by simulation* to be χ^2_{g-2}, where g is the number of groups into which the fitted probabilities are divided. Pick your favourite GLM scenario and carry out a simulation to see if you agree with this.

OCT 18: In class today we discussed the latent variable representation for the probit regression model, and the EM algorithm for finding maximum likelihood estimates by taking advantage of the form of the 'complete-data' likelihood.

Say $Z \sim N(\mu, \sigma^2)$. Determine expressions (involving the standard normal density and/or distribution functions) for $E(Z|Z > c)$ and $E(Z|Z < c)$.

Implement the EM algorithm for fitting a probit regression model. How does this algorithm seem to compare with the usual IRWLS algorithm? Obvious comparisons would involve whether the algorithms converge from various starting values, and how many iterations each takes.

Note: if you look at the R documentation for glm(), it is possible for the user to specify starting values.

NOV 5: Exercises 8.2 and 8.3 from the text.

COMPUTING: As mentioned in class, glm() in R does not support multinomial models. However multinom() in library(nnet) and polr() in library(MASS) can be used for nominal and ordinal (proportional odds) models. Alternatively there is a more general package for categorical data glms available on the web – see http://www.stat.auckland.ac.nz/ yeee/VGAM/. I got as far as installing it but haven't tried it out yet. You should be able to try it on unixlab with

> libary(VGAM, loc.lib=”/home/faculty/gustaf/myRlibs”)

if you don't want to install it yourself. In particular, the examples under

> help(vglm)

point one toward fitting ordinal and nominal multinomial models.

NOV 8: Exercise 9.2 from the text.

THIS IS CUTOFF # 2. THE ABOVE PROBLEMS ARE DUE WED. NOV. 24 (OR SOONER IS FINE TOO!)

The third and final set of problems will be based on the 'special topics' portion of the course, and will be due near the end of the exam period.

SPECIFICALLY, THE FOLLOWING TWO PROBLEMS ARE DUE BY THE END OF THE EXAM PERIOD.

NOV 30: Do some empirical (i.e., simulation) work to compare random effect models (i.e, GLMMs) and marginal models. There is lots of flexibility in how to do this, and you can use some of the R examples from class as a starting point. Based on your empirical work, do you agree that getting the correlation structure 'wrong' is 'no big deal' when fitting marginal models with the gee/sandwich estimation method. What about the somewhat analogous question of getting the distribution of the random effects wrong in fitting a glmm? Another question of interest might be whether fitted values or predicted values from the two kinds of models seem similar for the same dataset.

NOV 30: Construct a fairly simple GLM scenario with some missing data and implement an EM-algorithm approach to model fitting which adjusts for the missing data. Again you can use the in-class example as a starting point, but you are likely to get more credit if you don't simply 'clone' that example.

While we talked about it in general, we didn't get to the specifics of SEs which are adjusted for missing data, so you can just consider point estimation. (If you are brave enough to take on the SE problem, you may get some extra credit.) Look specifically at whether the EM approach tends to improve things in the sense of providing point estimates that move away from the complete-case estimates and toward the full-data estimates. (Of course, we can only get full-data estimates in simulated settings where we can access the 'unobserved' data.)

CONGRATULATIONS - YOU HAVE MADE IT THROUGH STAT 538. CONSIDER YOURSELF A GENERALIZED LINEAR MODELLER!