Model comparison

Patrick Breheny

March 28

Wells in Bangladesh

- In this lecture and the next, we will consider a data set involving modeling the decisions of households in Bangladesh about whether to change their source of drinking water[1]
- Many of the wells used for drinking water in Bangladesh and other South Asian are contaminated with naturally occurring arsenic, affecting an estimated 100 million people
- Arsenic is a cumulative poison, with risks of cancer and other diseases thought to be proportional to exposure

[1] This data set comes from Gelman & Hill (2007), "Data Analysis Using Regression and Multilevel/Hierarchical Models"

Switching

- A research team from the United States and Bangladesh measured arsenic levels for all wells in a certain area, labeled the well with its arsenic level, and encouraged households drinking from unsafe wells ($> 0.5 \mu g/L$) to switch to a safer well
- A few years later, the researchers returned to find out who had switched wells `Switch=1` and who had not `Switch=0`
- The file `wells.txt` contains information on well switching for 3,020 households

Explanatory variables

We consider the following explanatory variables:

- `Arsenic`: The arsenic level of the household's well
- `Dist`: The distance to the nearest safe well
- `Community`: Whether any members of the household are active in community organizations
- `Education`: Years of education of the head of the household

R^2 for logistic regression?

- In linear regression, R^2 is a very useful quantity, describing the fraction of the variability in the response that the explanatory variables can explain
- There are a number of ways one can define an analog to R^2 in the logistic regression case, but none of them are as widely useful as R^2 in linear regression

Correlation approach

- One approach is to compute the correlation r between the observed outcomes $\{y_i\}$ and the fitted values $\{\hat{\pi}_i\}$
- In linear regression, the square of this correlation is R^2
- Thus, one reasonable way to define an R^2 for logistic regression is to square r, the Pearson correlation between the observed and fitted values

Squared error approach

- Another approach is to measure the reduction in squared error:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- This approach has the advantage that it looks exactly like R^2 for linear regression, and we can therefore easily adjust for the number of parameters:

$$R^2_{\text{adj}} = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2/(n-p)}{\sum_i (y_i - \bar{y})^2/(n-1)}$$

A closer look at squared error assumptions

- These two preceding measures have the advantage of working on the scale of the original variable and being easy to interpret
- However, one might question the logic of treating all $(y_i - \hat{\pi}_i)$ differences equally
- Compare $\hat{\pi}_i = .9$ with $\hat{\pi}_i = .99$ for an observation with $y_i = 0$
- The squared differences are similar ($0.99^2 = 0.9801$, $0.9^2 = 0.81$) despite the fact that $\Pr(y_i = 0)$ differs by a factor of 10 for the two models

Deviance vs. squared error

- This is the rationale behind considering differences on the likelihood scale (*i.e.*, instead of looking at the reduction in squared error, we look at the reduction in deviance)
- In our example, the contribution to the deviance by the two estimates are

$$-2\log(.1) = 4.6$$
$$-2\log(.01) = 9.2,$$

a two-fold difference, as opposed to the 20% difference as measured by squared error

Explained deviance

- Letting D_0 denote the null deviance (*i.e.*, the deviance of the intercept-only, or simple binomial, model), another attempt at an R^2-like measure is

$$\frac{D_0 - D}{D_0} = 1 - \frac{D}{D_0},$$

  the *explained deviance* (often reported as a percentage)
- Because deviance roughly follows a χ^2_{n-p} distribution, it can also be adjusted for number of parameters:

$$1 - \frac{D/(n-p)}{D_0/(n-1)}$$

Other approaches

- Other approaches involve looking at all pairs for which $\hat{\pi}_i > \hat{\pi}_j$ and recording whether or not y_i and y_j differ
- If $y_i = 1$ and $y_j = 0$, then our model gets a point; if $y_i = 0$ and $y_j = 1$, then our model loses a point (nothing happens if y_i and y_j are the same)
- This is the idea behind *Kendall's* τ, *Somer's* D, and *Goodman and Kruskal's* γ
- There are several other approaches too, so almost a dozen altogether (thankfully, they all have the property that they lie between 0 and 1, with 1 being the best)

Well-switching example

To get a sense of how these measures look, let's compare three models:

Model 1: $\eta = \beta_0 + \beta_1 \texttt{Distance}$

Model 2: $\eta = \beta_0 + \beta_1 \texttt{Distance} + \beta_2 \texttt{Arsenic}$

Model 3: $\eta = \beta_0 +$ all explanatory variables

Well-switching example (cont'd)

	Model		
	1	2	3
r^2	0.014	0.062	0.068
R^2	0.014	0.061	0.068
R^2_{adj}	0.014	0.061	0.066
DE	0.010	0.046	0.051
DE_{adj}	0.010	0.045	0.050
τ	0.050	0.142	0.146
γ	0.104	0.293	0.299
Somer's D	0.103	0.291	0.298

Low values for R^2 and deviance explained are fairly common in health behavior studies such as this one

Classification

- An alternative way of thinking about how well a model fits the data is with respect to *classification*
- This approach forces the model to predict whether $y_i = 0$ or $y_i = 1$ based on $\hat{\pi}_i$
- The obvious approach is to predict $y_i = 1$ if $\hat{\pi}_i > 0.5$, although other cutoffs could be used if, for example, the cost of false positive is larger than the cost of a false negative (or vice versa)

Classification table

For example, let's compare Models 1 and 3:

Model 1		
	No	Yes
$\hat{\pi}_i < 0.5$	194	133
$\hat{\pi}_i \geq 0.5$	1089	1604

Model 3		
	No	Yes
$\hat{\pi}_i < 0.5$	470	346
$\hat{\pi}_i \geq 0.5$	813	1391

Note that we have 1,222 incorrect predictions on the left, and 1,159 on the right

ROC Curves

- However, we can consider varying the cutoff to which $\hat{\pi}_i$ is compared
- As we do so, we will change both the *false positive rate*:

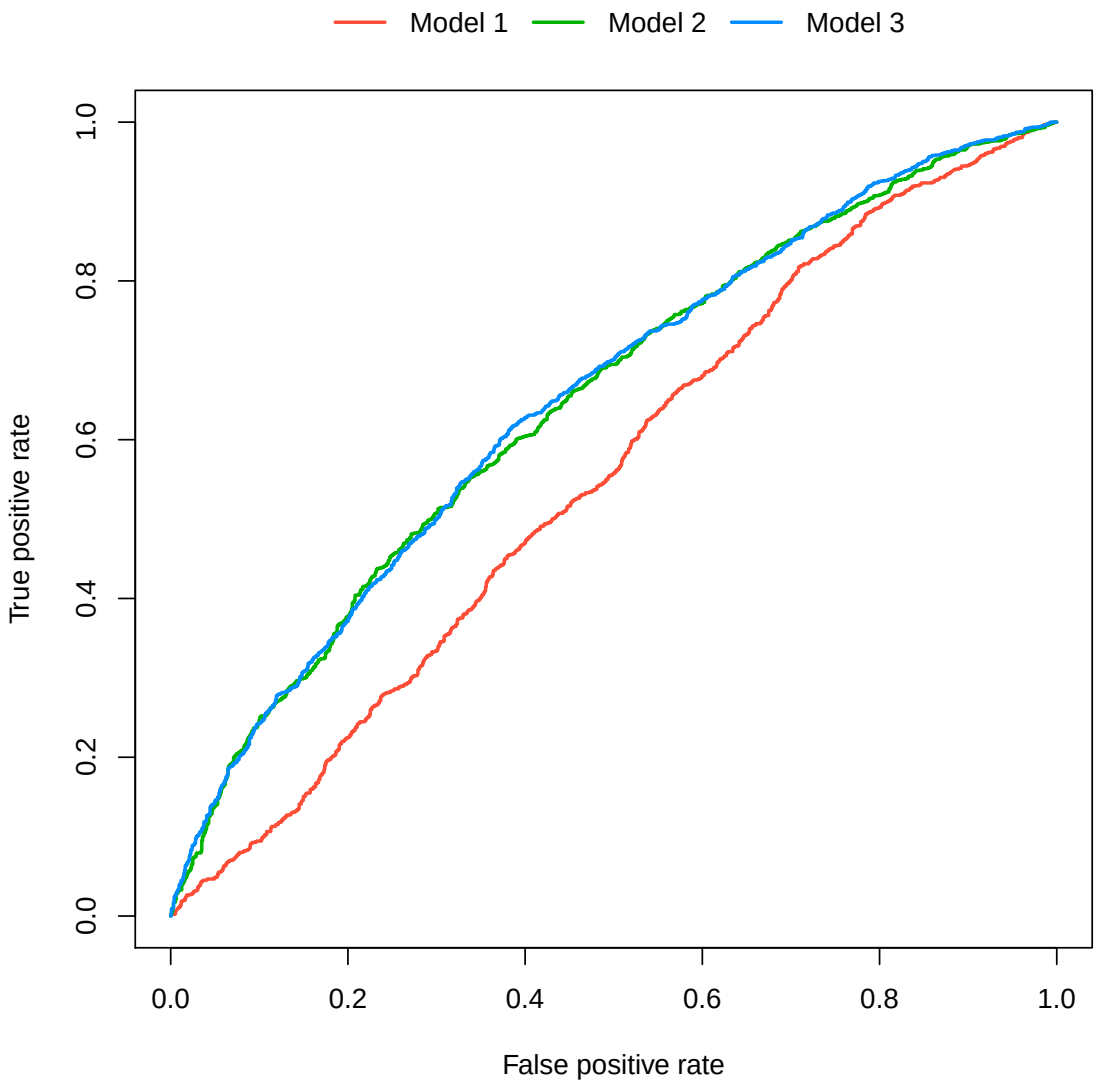
$$\Pr(\hat{y} = 1 | y = 0)$$

and the *true positive rate*:

$$\Pr(\hat{y} = 1 | y = 1)$$

- The true positive rate is also called the *sensitivity* and 1 minus the false positive rate is also called the *specificity*
- As we vary the cutoff from 0 to 1, plotting these two quantities will create a curve known as the *receiver operating characteristic* (ROC) curve

ROC curves for well-switching data



AUC

- For the three models on the previous slide, no matter what the false positive rate, models 2 and 3 had higher true positive rates than model 1
- However, comparing models 2 and 3, either model could be "on top" depending on where we are at on the curve
- A useful summary of the overall quality of the curve is the area under the curve, or AUC (SAS refers to this as "c"; it is located next to γ, τ, and D in the "Association of predicted probabilities..." table):

	Model 1	Model 2	Model 3
AUC	0.55	0.65	0.65

Note that random guessing would yield an AUC of 0.5; perfect classification would yield an AUC of 1

Basic principles of model selection

Let's remind ourselves of the basic principles of model selection that we discussed at the beginning of the course:

- Simple models have low variance, but risk bias
- More complicated models reduce bias and fit the sample data better, but can be highly variable and do not necessarily generalize to the population better
- Model selection criteria can be informative, but should not be applied blindly – there is no substitute for thinking carefully about the scientific meaning and plausibility of the models under consideration

AIC

- Consider the expected prediction accuracy of a model using log-likelihood as a measure of accuracy:

$$\mathrm{E} \sum_i \log \Pr_{\hat{\theta}}(Y_i),$$

where $\hat{\theta}$ is the MLE of the parameters of the distribution function for y and the $\{Y_i\}$ are out-of-sample random variables (*i.e.*, not the $\{y_i\}$ used to fit the model)

- Akaike showed that

$$-2\mathrm{E} \sum_i \log \Pr_{\hat{\theta}}(Y_i) \approx -2\mathrm{E}(\mathrm{loglik}) + 2p,$$

where loglik is the log-likelihood of the fitted model

AIC: Interpretation

- This suggests the following criterion, named the *Akaike information criterion*:

$$\text{AIC} = -2\text{loglik} + 2p = D + 2p$$

- Certainly, a lower AIC is better than a higher AIC (we wouldn't want our expected deviance to be large), but suppose the AIC values for two models differ by, say, 1; is that a meaningful difference?
- A useful rough guide is that AIC differences under 2 are not particularly meaningful, AIC differences of around 5 are fairly convincing, and AIC differences over 10 provide clear support for the model with the lower AIC

BIC

- Another common information model selection criterion for GLMs is called the *Bayesian information criterion*, or BIC
- As you might guess, its derivation is Bayesian and beyond the scope of this course
- However, its form happens to be very similar to AIC:

$$\mathrm{BIC} = -2\mathrm{loglik} + p\log(n) = D + p\log(n)$$

- Note that because $\log(n)$ is bigger than 2 (unless $n < 8$), BIC penalizes model complexity more heavily than AIC, and thus tends to favor more parsimonious models

BIC: Interpretation

BIC has a direct Bayesian interpretation in that it allows you to calculate (approximately, given equal prior probability on each model) the posterior probability of each model under consideration:

$$P(M_j|\mathbf{y}) \approx \frac{\exp(-0.5\text{BIC}_j)}{\sum_k \exp(-0.5\text{BIC}_k)},$$

where the sum is over the models under consideration

AIC and BIC

Applying AIC and BIC to our three models from earlier:

	Model		
	1	2	3
AIC	4080	3937	3918
BIC	4092	3955	3948
$P(M_j\|\mathbf{y})$	0.00	0.03	0.97

Both approaches agree that the most complex model is the best despite its extra parameters, although BIC is less enthusiastic about the difference between models 2 and 3

Summary

- It is important to keep in mind the famous words of George Box:

  All models are wrong, but some are useful.

- Certainly, a useful model should fit the data well, and information criteria are helpful guides here, but other considerations, such as interpretability and scientific justification are also important
- We will continue to look at the well-switching data next time, applying a mix of both statistical and extra-statistical considerations