Validating Models

CAS Special Interest Seminar on Predictive Modeling

Christopher Monsour, FCAS, MAAA

October 5, 2006

Why?

- Different types of validation
  - Is it the right model?
    - Is this predictor worth including?
  - How will it perform?
  - What to do about overfit models?
  - What should my tuning parameter be?

Right Model?

- Validation (in a broad sense) as part of model search
  - Supplement to significance testing
  - Handles difficult comparisons between models

Recall What a GLM Is

- $E(y|x)=f(\beta_0+\Sigma\beta_i x_i)$
- $Var(y|x)=\phi V(E(y|x))=\phi V(f(\beta_0+\Sigma\beta_i x_i))$

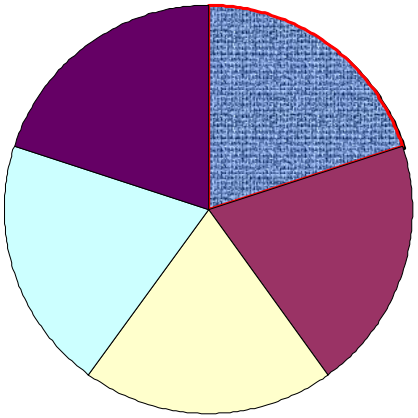
- The β_i are estimated by maximum likelihood or maximum quasilikelihood, and do not depend on the estimate of ϕ
- Once the β_i are estimated, ϕ is estimated by one of a number of methods

Whether to Include a Predictor in a GLM?

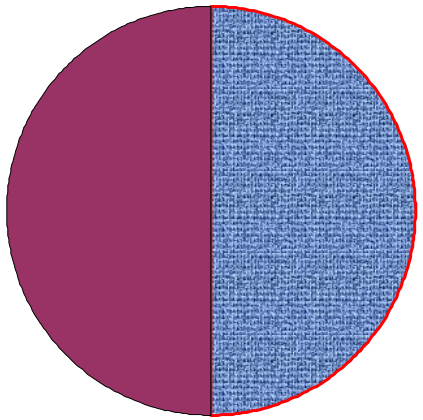
- p-value of the parameter estimate not always very useful in an absolute sense
  - Inaccurate because assumptions violated
    - Var(y|x)=ϕV(μ) usually not entirely accurate
      - But the hypothesis test depends on this
    - Model not fully specified
      - Rare that a model would be fully specified
    - Similar difficulties with other types of models…not just a GLM problem
  - Not perfect in a relative sense, but even less accurate in an absolute sense, since absolute p-values very sensitive to estimate of ϕ, the dispersion parameter
  - So what to do?

Whether to Include a Predictor in a GLM?

- Compare many models that are a little different
  - Losses capped at different amounts
  - Different pieces of data taken
    - — At random (e.g., ordinary validation and cross-validation set-ups
    - — By design (e.g., by policy year or coverage choice)



Can use a "voting procedure" to determine whether to include a predictor



Whether to Include a Predictor in a GLM?

- Look at which parameter estimates are relatively stable as data perturbed
  - Also a good way to find interactions
    - when the divisions in the data are not random
  - Sometimes easier just to test an interaction directly
    - But if you have a stepwise algorithm set up, may be worth running that separately on different pieces of the data
- For other types of models, similar ideas apply. For hierarchical models, you can perturb the model's hierarchy too
  - For example, trees using the 2nd or 3rd best main split

Whether to Include a Predictor in a GLM?

- Another approach—break the dependent variable into its pieces
  - For example
    - — Frequency and severity by peril rather than loss ratio for all perils combined
    - — Lost-time and medical-only losses separately
  - Reduces but does not eliminate need to “perturb” data to validate variable selection in each model
    - — Reduces the need because model closer to being fully specified and variance assumption closer to being true
    - — And because “outliers” have been relegated to specific sub-models (e.g., fire losses or hurricane losses)

Whether to Include a Predictor in a GLM?

- Yet a third approach:
  - Performance on test data with and without the predictor
  - Of course, you don't just pull the predictor out of the model, you also have to re-estimate the other parameters when you do so
  - When the parameter estimate across subsets of data is stable, the parameter estimated on one subset will improve prediction on the other

Other Types of Comparisons

- How many degrees of freedom to give a predictor
  - AIC, Schwarz's Bayesian Criterion, etc., not always the best way to evaluate this
    - — Depend on GLM assumptions
    - — Depend on the estimate of ϕ since the (quasi)likelihood depends on the estimate of ϕ
  - Can look at performance on test data with the more and less complex treatments
    - — E.g., 2 versus 3 parameters for an amount of insurance predictor

How Will it Perform?

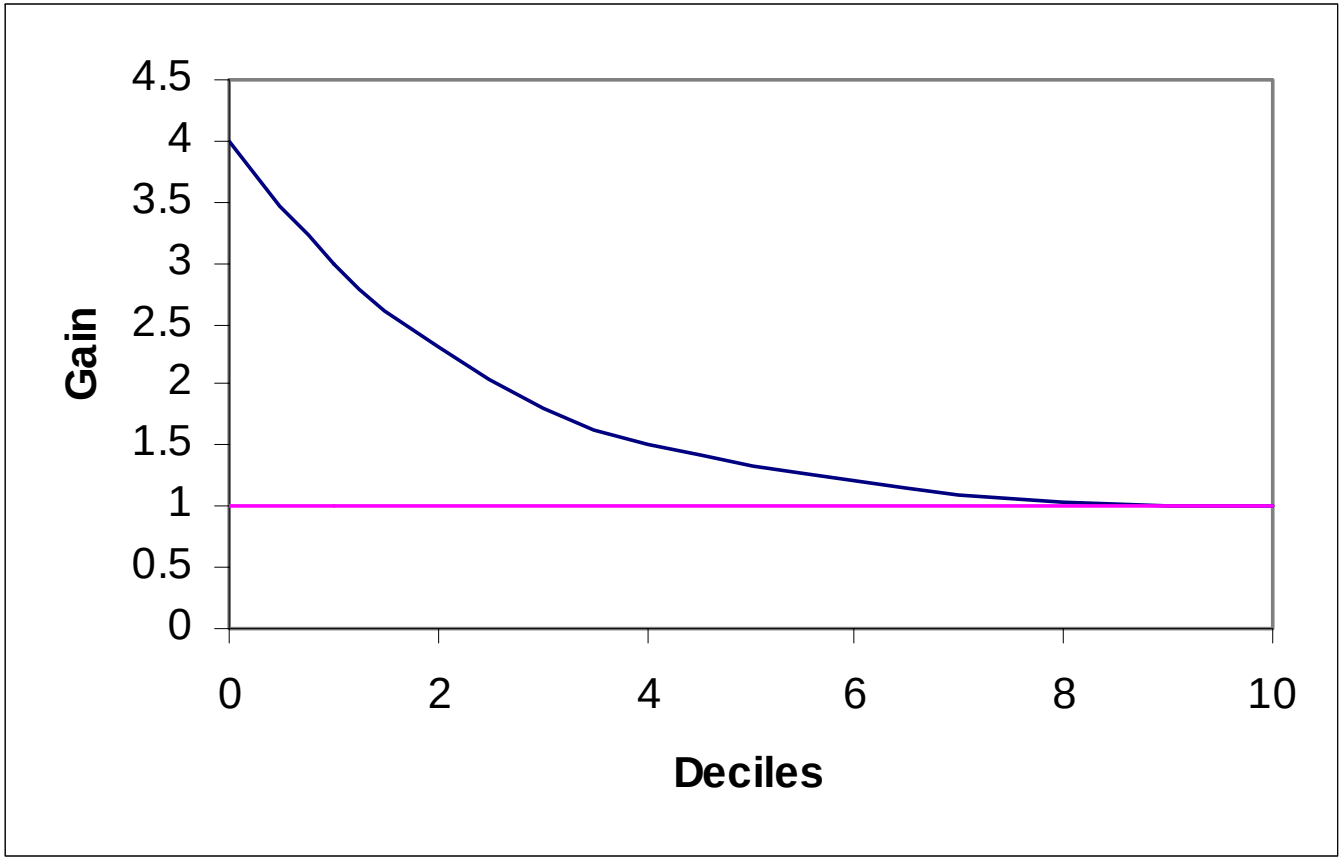
- Management wants to know
- Testing on new data obviously ideal
- Cross-validation often employed because can't hold out data
  - Regulatory reasons
  - Practical reasons
    - Thin data to start with
      - 500,000 exposures may not seem thin, but you may be trying to estimate a parameter for 0.2% of the population, and that group may have a 5% claim frequency
- Goal: Make the validation as accurate and objective as possible
- Will assume here that there is an agreed performance measure on out-of-sample data, whether it is lift or MSE or concordance or classification accuracy or a gains chart or something else

Accuracy and Objectivity

Out of Sample Validation (Hold Out)

- If you don't later re-estimate the model from the full data
  - Accurate and objective but model may be sub-par
- If you do re-estimate from the full data
  - Approximate lower bound for model quality, as long as the same process followed in re-estimating from the full data
- Can't help if a data quirk is common to all the data including the hold-out, but not to the operational data going forward
  - For example, if some characteristic (e.g., divorced status) is more accurately recorded in historical data for policies with claims

Response Modeling Example—Gains Chart



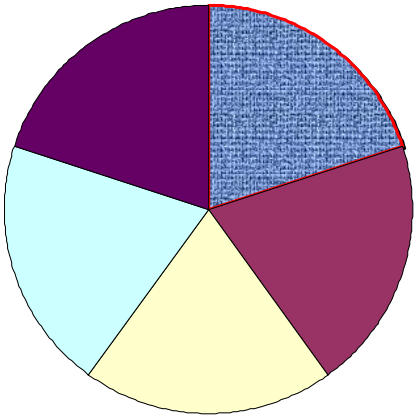
Accuracy and Objectivity

Cross-Validation

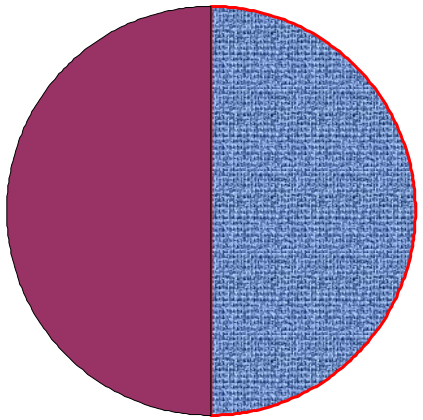
- Estimate model from full data
- Divide data into K groups at random
- Re-estimate model on each collection of K-1 groups
- Ideally, re-estimation means applying the same ***process*** that created the full model
  - What if this involves a lot of human judgment (in selecting the predictors, for example)?
    - — Can impair accuracy of the validation if this isn't applied to each cross-validation model
    - — But does subjective judgment have a place in "validation"?

Accuracy and Objectivity

- Different pieces of data taken
  - At random (e.g., ordinary validation and cross-validation set-ups
  - By design (e.g., by policy year or coverage choice)



Use each slice as test data once, training data all the other times



Cross-Validation

- To re-phrase the issue: If you don't redo the feature selection, then each cross-validation model still depends on a choice that was based on all the data
  - But could a human being ignore that anyway?

Accuracy and Objectivity

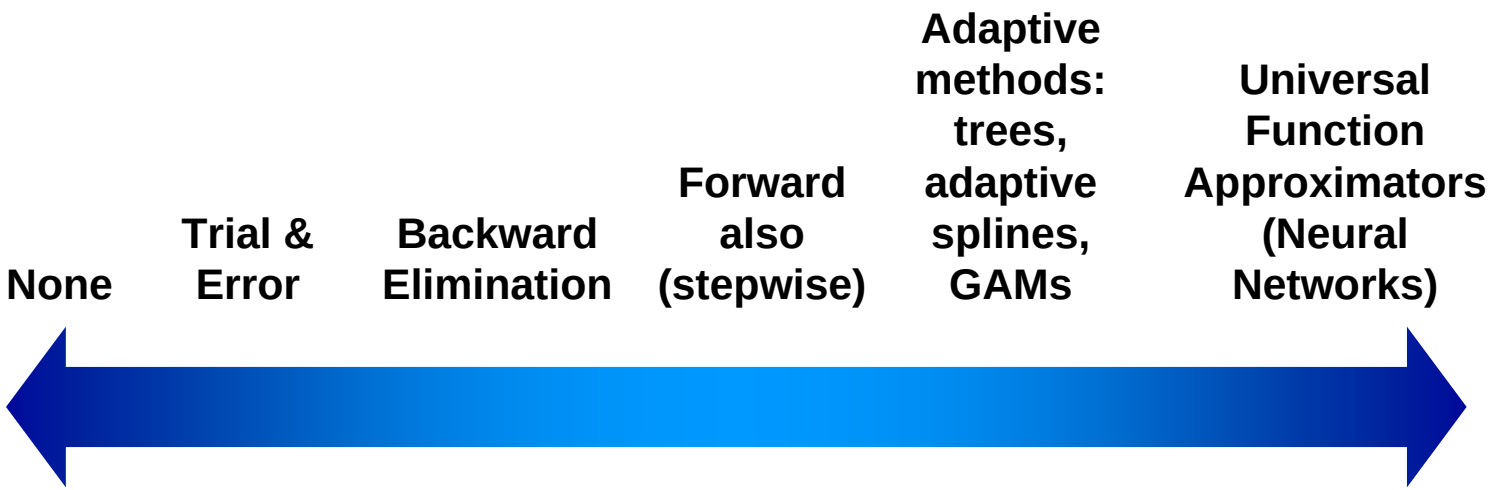
Cross-Validation

- Balance:
  - The amount of optimism depends on how many models are searched
    - Another reason criteria like AIC that relate optimism to d.f. don't always work well
    - Big difference between
      - Filtering 7 variables down to a 4 variable model by hand
      - And sifting through 500 derived variables with stepwise to arrive at a 4 variable model!
  - Humans don't search many by hand
  - Machines search quite a few

Accuracy and Objectivity

Cross-Validation

- Intensity scale of model search:



Cross-Validation

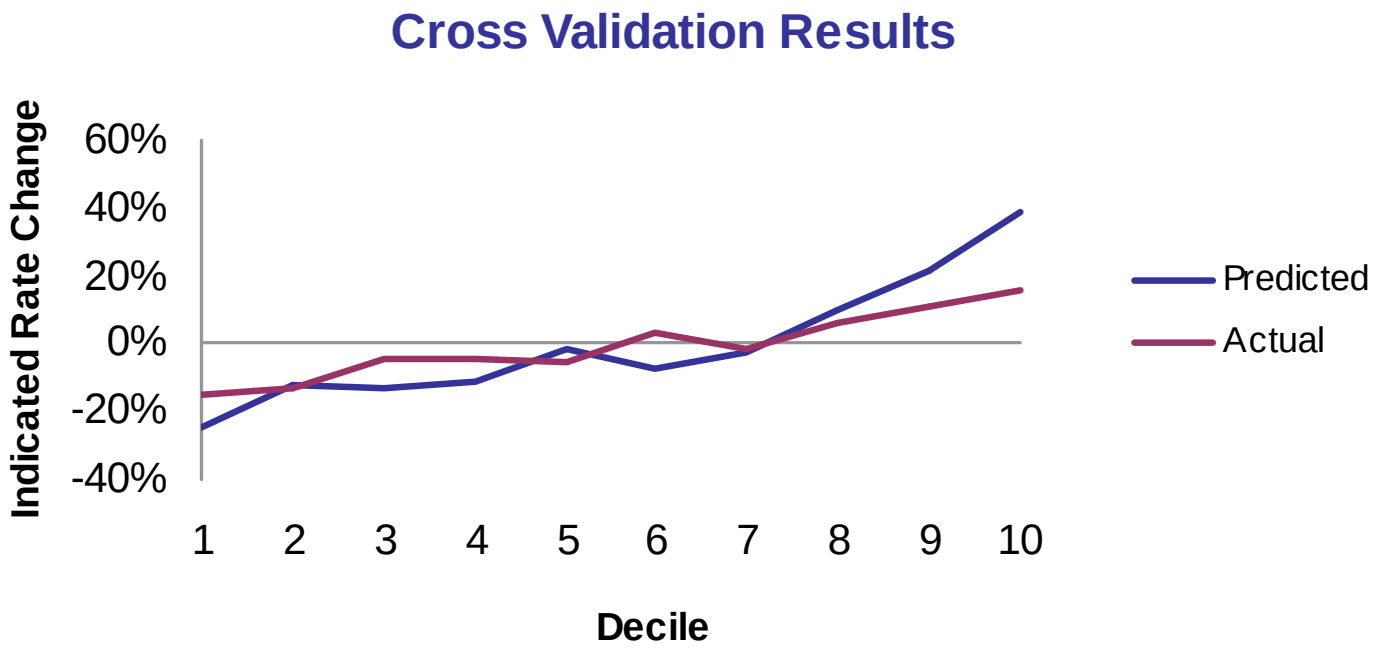
- Another item to be aware of:
  - All else being equal, cross-validation tends to be slightly pessimistic about model performance
    - For example, if K=10, it's based on the performance of models that only look at 90% of the data
    - For some measurements (e.g., prediction error) there are corrections available

So my model is overfit….what now?

- It depends
  - If you have lots of data, or a complicated model, fit a simpler model
  - But what if you have limited data and the model is not particularly complicated?
    - In response modeling, you use an overfit model, but state performance metrics based on performance on out-of-sample data
    - For ratemaking, set **rates** based on performance on out-of-sample data

Performance

So my model is overfit….what now?



So my model is overfit….what now?

Decile	Predicted	Actual
1	-24%	-15%
2	-13%	-13%
3	-13%	-4%
4	-12%	-5%
5	-2%	-5%
6	-8%	3%
7	-2%	-2%
8	10%	6%
9	22%	11%
10	39%	16%

- If the deciles are scoring tiers, use the predicteds rather than the actuals for the rating factors

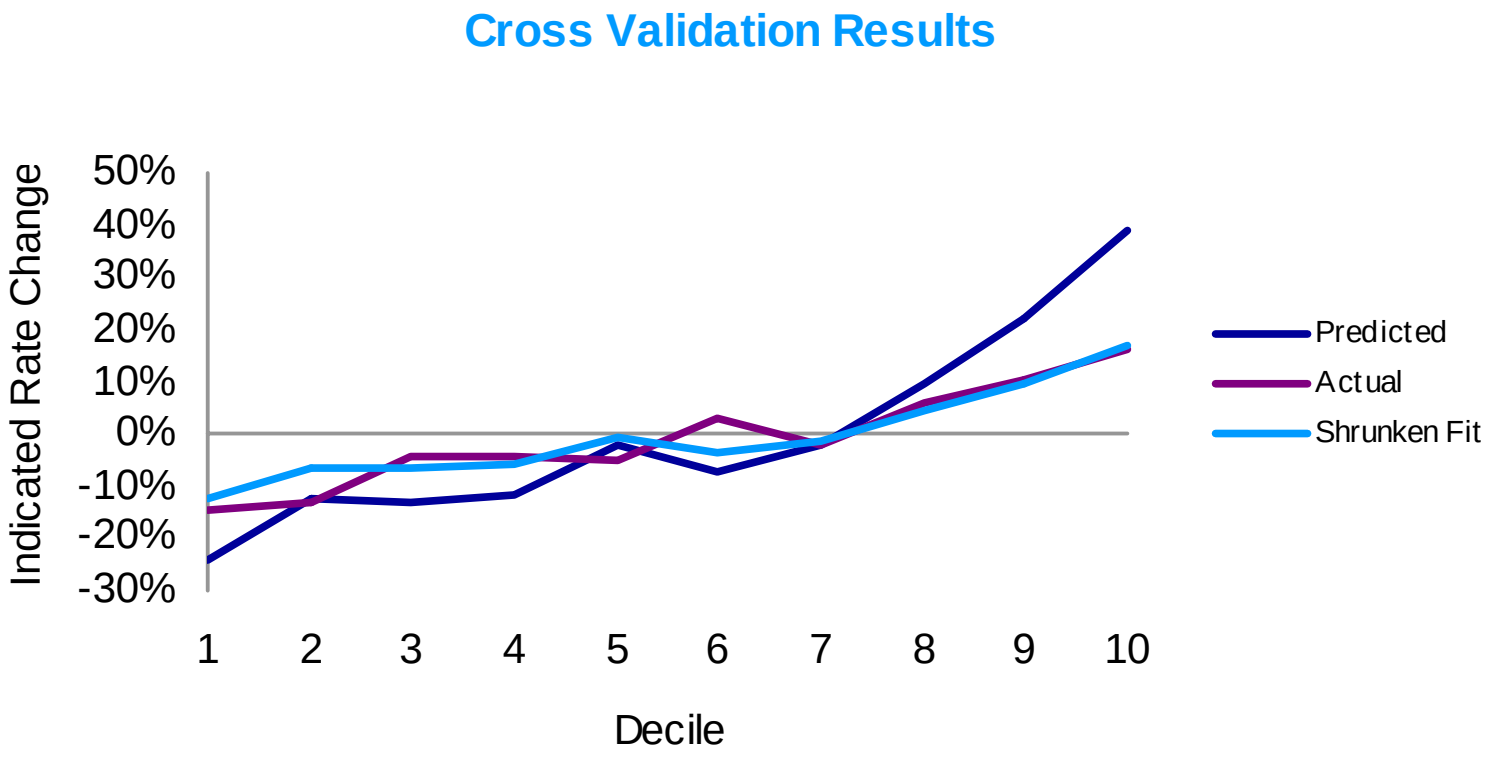
Shrunken Fit

So my model is overfit….what now?

Decile	Predicted	Actual
1	-24%	-15%
2	-13%	-13%
3	-13%	-4%
4	-12%	-5%
5	-2%	-5%
6	-8%	3%
7	-2%	-2%
8	10%	6%
9	22%	11%
10	39%	16%

- Note on the left, sample variance of the Predicted column is 19%, while 9% for the actual column.
- Thus, if supports a rating plan, could shrink all the parameters by a factor of roughly 9/19.
- Thus, if the model has a factor of 2.00 for class A, the new model would have a factor of 1.39=exp(9/19*log(2))
- May need to adjust the 9/19 ratio to make the variance match precisely

Shrunken Fit



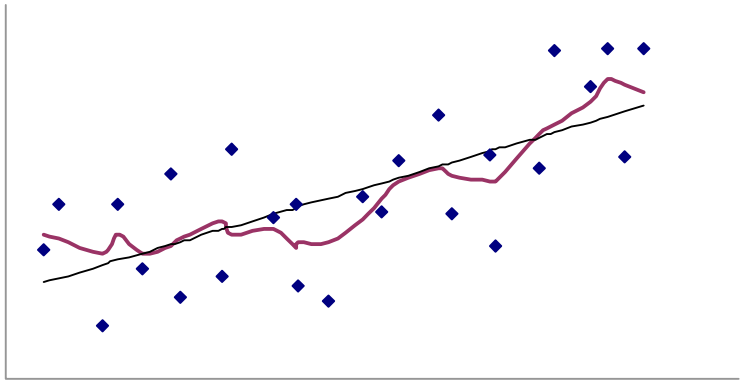
Brief aside on
Cross-Validation for Tuning Parameters

Using Cross-Validation to Tune a Parameter

- Data miners will tell you all about cost-complexity parameters and ridge regression and lasso parameters
  - E.g., ridge regression: $y=\Sigma x_i\beta_i$ subject to $\Sigma\beta_i^2 <= \Lambda$
    - Use cross-validation to optimize Λ
- Other examples of parameters that are routinely cross-validated:
  - Mixing proportion for how much weight to give to a pooled variance estimate and how much to give to a separate variance estimate for each class
    - For example, gives rise to compromise between linear and quadratic discriminant analysis

Using Cross-Validation to Tune a Parameter

- Other examples of parameters that are routinely cross-validated:
  - Window width for local regression



  - Roughness penalty for a smoothing spline
    - Typical formulation:
      - Minimize squared error plus $\lambda \int f''$

Using Cross-Validation to Tune a Parameter

- We actuaries make extensive use of a parameter you can also cross-validate
  - K in Bayesian credibility
  - So can apply cross-validation even to a simple one-way class ratemaking problem!
- Nice thing about cross-validation here is it automatically takes into account how accurate the complement of credibility is

Food for Thought

- Other practices around model validation and quantifying model performance? Best practices?
- Should there be a standard of practice? If so, what would it say?
- Motivating references.
  - Breiman, Friedman, Olshen, and Stone, *Classification and Regression Trees*
  - Davison and Hinckley, *Bootstrapping Methods and Their Applications*
  - Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*
  - Domingos, “MetaCost: A General Method for Making Classifiers Cost-Sensitive”, 5th Int’l KDD Conference, 155-164
  - Domingos, “The Role of Occam's Razor in Knowledge Discovery”, Data Mining and Knowledge Discovery, 3, 409-425