Models for Prediction

Dave Glidden
18 November 2014

Distinct Objectives

for a regression model

1. Developing a Model for Prediction/ Stratification
2. Adjusted Effect of Single Variable
3. Identifying Multiple Predictors of Outcome
4. Adjustment to Improve Efficiency in RCT

Example: Prediction

- Cohort: 11,701 community-dwelling elders
- Predictors: age, co-morbidities, functional status
- 42 candidate predictors
- Follow-up: mortality over 4 years
  1,361 deaths
- Who is at the highest risk of death?

Lee, SJ et al. JAMA. 2006;295:801-808.

Study Context

- Objective: inform patients/risk stratification
- Predictors obtained from patient report
- Wanted index "as simple as possible"
- Combine functional measure + co-morbidities
- Doesn't state if high or low risk more important: important for grouping

Statistical Objective

- Develop a regression model
- Discriminates between high and low risk
  how to quantify discriminatory ability?
- Pure prediction is not enough
  want some parsimony/interpretation
  external validity
- Implications for predictor selection
  all 42 significantly associated with death

Getting Best Model

- Need a measure of model predictiveness
- Sift through large # models
- Handle optimism/overfitting
  "how many predictors can the data handle?"
- Consider what will validate externally

Finding Our Model

Look at every possible model and choose the best one in terms of prediction

1. Infeasible in many cases
2. Susceptible to "overfitting"
3. Places no value on simplicity or plausibility

Measures of Prediction

- Log-likelihood
  measure how close model is to the data
  how max. likelihood methods to select model
  analogous to a sum of squares in linear model
- Binary data: C-statistic
  area under ROC curve
  "proportion of person with event has higher risk than someone without event"

A good model optimizes these

Best Subsets

- Consider all possible models
- Each variable may be in or out
  2 possibilities
- Repeated for each predictor
  total number of predictors p
- 2 x ..... x 2 (p times) = 2^p
- Increases very fast 2^{42} = 4.4 trillion models
  trillion seconds > 31,000 years
  cut predictors in half: 2.1 million models

Backward Selection

1. Start all variables in the model
2. Fit model
3. Drop term with highest p-value
4. Repeat 2 and 3 until all p-value less than some threshold (e.g., $p < 0.05$)

what is the right threshold? Avoid overfitting

Stepwise Model Selection

- Reduces the model selection immensely
- 42 predictors => 42 possible models
  43 predictors => 43 possible models
- Gives a logical sequence for fitting
  probably don't need to fit all 42
  fit until some criterion is reached
- But will not select the best model

Overfitting

Overfitting is a broad term that refers to entering too many terms into the model -- for the amount of available data.

This produces unstable, variable estimates. The extreme ones *tend* to be both spurious and also *tend* to get the most focus.

Hence, results are frequently unreplicable.

Testimation Bias

The combination of testing a hypothesis and then estimating a quantity, resulting in estimation of an effect which is stronger than is true. One example the process of estimating regression coefficients only among significant predictors.

Example

- Take 5% sample of Lee data
- Fit all variables to data subset
- 0.05 criterion with backward selection
- Would results be reliable?

Model Replication

mortality model

	Odds Ratio	
Variable	test data	full data
hrsdm	2.9	1.8
bmicat	3.3	2.2
hrsmemory	45.7	2.7
hrsarth	7	1.3
hrscad	4.2	1.4
hrslung	4.8	2.3
meals	5.4	2.7
walkjog	9.5	3.8

yellow: test result lies outside CI for full data

Poor Validation

- Odds Ratios consistently overestimated
  consistent overestimation -> bias
- Many estimates outside CI
- More extreme OR -> more overestimated
- Artifact of model selection procedure
  retained only if $p < 0.05$

Cross-Validation

- Correcting for model-based optimism
- Split data into a series of random subsets
  usually 10 of them
- one-at-time, leave subset out
- Build model on 90%
  use other 10% for validation

Stata Code

- gen u = uniform()
  makes random # and puts in "u"
- xtile cat=u, nq(10)
  creates "cat" 10 groups based on "u"
- logistic dead *predictors* if cat!=k
  xtile cat1=u if dead==1, nq(10)
- predict *prediction* if cat==k
  save prediction for kth group

How it works

- Uses data to build and validate model
- 90% builds model, 10% validates it
- Every observation gets used in 1 test set
- Provides an internal validation

Model Fit and Complexity

- Log-likelihood (LL) measures closeness of data to model
- -2*Log-likelihood: like a sum of squares
  smaller means more variation explained
- Always goes down with more predictors
- LR test: must go down by 3.84 for 1 predictor to be significant

Aikake Info. Criterion

- Turns out -2*log-likelihood is biased
- Assessed on same data use to minimize
- Unbiased version:
- -2*log-likelihood - 2*K
- K is # parameters in model
- Call the Aikake Info. Criterion (AIC)
- p-value criterion $p < 0.16$

Bayesian Info. Criteria

- A more stringent version
- -2*LL - K*log(n)
- n is number of predictors
- bigger penalty for adding predictor
  increases with sample size

BIC Table

N	Chi2 >	p-value
20	3	0.083
100	4.6	0.032
200	5.3	0.021
500	6.2	0.013
2000	7.6	0.006

Information Criteria

- Choose model with lowest AIC/BIC
- Very big difference
- AIC will allow some non-sig predictors
- BIC will exclude some non-sig predictors
  BIC can lead to underfitting

Other Strategies

Model Selection

- Considered families of variables
  - demographic (age/sex)
  - behavior (alcohol / BMI)
  - co-morbidities (hypertension/falls)
  - functional (walking, managing $)

Families of Variables

- Likely correlated
- May measure similar things conceptually
- May make sense to exclude as a group
- Or include some measures, not others
- Facilities interpretability

Family & Overfitting

- 8 measure of difficult walking
  all slightly different
- Examine their "bv" associations
- Understand differences in what they measure
- Pick 1 for "multivariate analysis"
- Greatly reduces the # models considered
  mitigates overfitting, increases interpretation

Exclusion: # Surgeries

- Varies by region not by disease severity
- Doesn't measure disease clearly
- Not a portable predictor
- Especially outside US
- Use of external information in selection

Stepwise Regression

- Lee et all uses backwards
- $p < 0.05$
- Use all 42 variables
- 19 significant
- Can produce unstable models
    - Small changes > very different models

Lee Model Reduction

- Took original model with BIC
  19 predictors
- Deleted least significant predictors
- Until BIC no longer went down
- p-cut off about 0.0022
- Strong protection again false positives
- Resulted in 12 (17 df) predictor model

17 df predictor model

```
. xi: logistic dead i.agecat male hrsdm hrsca hrslung hrschf bmi smoke bath money walkjog push
i.agecat          _Iagecat_0-6        (naturally coded; _Iagecat_0 omitted)

Logistic regression                               Number of obs   =      11652
                                                  LR chi2(17)     =    2080.12
                                                  Prob > chi2     =     0.0000
Log likelihood = -3132.1774                       Pseudo R2       =     0.2493

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagecat_1 |   1.886258   .2743998     4.36   0.000     1.418319     2.50858
  _Iagecat_2 |   2.826806   .3996489     7.35   0.000     2.142667    3.729387
  _Iagecat_3 |   3.721534   .5101651     9.59   0.000     2.844693    4.868651
  _Iagecat_4 |   5.416485   .7430223    12.32   0.000     4.139534    7.087346
  _Iagecat_5 |   8.328805    1.19095    14.82   0.000     6.293148    11.02294
  _Iagecat_6 |   16.26596   2.360374    19.22   0.000     12.23942    21.61716
        male |     2.0384   .1436712    10.10   0.000     1.775394    2.340367
       hrsdm |   1.782802   .1512828     6.81   0.000     1.509638    2.105393
       hrsca |   2.056083   .1751382     8.46   0.000     1.739942    2.429665
     hrslung |   2.276356   .2831871     6.61   0.000     1.783806    2.904911
      hrschf |   2.342325   .3403233     5.86   0.000     1.761869    3.114015
      bmicat |   1.651562   .1152799     7.19   0.000     1.440391    1.893691
       smoke |    2.05528   .1905406     7.77   0.000     1.713791    2.464814
        bath |   1.958479   .2066669     6.37   0.000     1.592563    2.408471
       money |   1.903665   .1830186     6.70   0.000     1.576725    2.298398
     walkjog |   2.087163    .164163     9.36   0.000     1.788983    2.435042
        push |   1.526814   .1208712     5.35   0.000     1.307375    1.783085
------------------------------------------------------------------------------
```

AIC/BIC in Stata

Age as categorical

```
. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |  11652    -4172.24   -3132.177     18     6300.355    6432.893
-----------------------------------------------------------------------------
```

Age as continuous

```
estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |  11652    -4172.24   -3116.505     13      6259.01    6354.732
-----------------------------------------------------------------------------
```

BIC lower with age as continuous

Model Reduction

- 19 (26df) variables to 12 (17df) variables
- Modest reduction in C-statistic
  0.85 v. 0.84
  negligible prediction lost
- Can help to reduce false positives
- Makes model more simple

Sensitivity Analyses

- All variables based on BIC
- Model reduction in predictor families
- Identified redundant/collinear variables
  tried replacing them
- Tried total co-morbidities
- All corroborated final model

Nice example of sensitivity analysis

Model Choice

- Substantive/statistical considerations
  thoughtful exclusions
- Emphasis on parsimony
  appropriate for prediction
- Extensive sensitivity analyses

Building Prediction Models

- Balance utility, prior knowledge
- Predictors powerful, portable, interpretable
- Mitigate overfitting: BIC, cross-validation, make prediction selection less data hungry
- Construct simple score
- Context dependent cutpoints

Summary

- Example of building for pure prediction
- Regression models => additive point scores
- Keep prediction models parsimonious
- Consider portability, interpretability, expense and difficult
- AIC/BIC can be a useful way to choose models

Prediction Issues

- How big should model be?
- Categorize predictors?
- How regression leads to "score"
- Want model good for new data
  not just recreating our data