Propensity Scores continued

15 November 2016
Dave Glidden
Biostatistics 210

Example

- UNOS data
- Survival of pediatric kidney rx
- Effect of living v. cadaveric organ on survival
- 351 deaths
- Model on level of outcome (survival) or
- Model on level of propensity (txtype)

The steps

- Predictor Selection
- Build model for exposure
  calculate the prob. of exposure for each person
- Assess overlap/positivity
- Compare exposure by level of propensity
  quintiles, splines, matching

Stata Code

- logistic txtype *predictors*
- predict prop, p
- xtile group=prop, nq(5)
- cc death txtype, by(group)
- logistic death txtype i.group

Quintile Overlap

5 quantiles of prop	txtype Living	Cadaveric	Total
1	1,375	95	1,470
2	1,252	217	1,469
3	666	804	1,470
4	197	1,272	1,469
5	14	1,455	1,469
Total	3,504	3,843	7,347

Quintile

```
. logistic death i.txtype i.group

Logistic regression                                     Number of obs   =       7347
                                                        LR chi2(5)      =      23.85
                                                        Prob > chi2     =     0.0002
Log likelihood = -1398.0372                             Pseudo R2       =     0.0085

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.txtype |   1.460813   .2433996     2.27   0.023     1.053823    2.024982
             |
       group |
          2  |   1.301893   .2544977     1.35   0.177     .8875235    1.909723
          3  |   1.274882   .2676887     1.16   0.247     .8447766    1.923968
          4  |    1.07496   .2522625     0.31   0.758     .6786397    1.702727
          5  |   1.480443   .3480316     1.67   0.095     .9338694    2.346915
             |
       _cons |   .0327919   .0048379   -23.16   0.000     .0245576    .0437872
------------------------------------------------------------------------------
```

Causal Effects

- p_0 = *Probability of death if everyone got living*
- p_1 = *Probability of death " " got cadaveric*
- Can be estimated by "margins" based on a particular model
- $p_1 - p_0$ = *ACE*
- p_1/p_0 = *Causal RR*
- $p_1 (1-p_0)/(1-p_1) p_0$ *Causal OR*

Marginal Effects

Margined Over Quintile

```
. margins txtype

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      txtype |
          0  |   .0386335   .0041469     9.32   0.000     .0305057    .0467613
          1  |   .0554319   .0046479    11.93   0.000     .0463223    .0645416
------------------------------------------------------------------------------

. margins, dydx(txtype)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.txtype |   .0167985   .0072948     2.30   0.021     .0025008    .0310961
------------------------------------------------------------------------------
```

Causal RR -- Quintile Fit

```
logistic death i.txtype i.group
margins txtype, post

 nlcom (risk_ratio: _b[1.txtype]/_b[0.txtype] )

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  risk_ratio |   1.434816    .228147     6.29   0.000     .9876558    1.881976
------------------------------------------------------------------------------


logistic death i.txtype i.group
margins txtype, post
. nlcom (logRR: log(_b[1.txtype]) - log(_b[0.txtype]))

       logRR:  log(_b[1.txtype]) - log(_b[0.txtype])

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       logRR |   .3610364   .1590079     2.27   0.023     .0493867    .6726861
------------------------------------------------------------------------------
```

marginal RR = 1.43 (1.05, 1.96)

Regression	Propensity Score
assess confounding	assess confounding
draw DAG	draw DAG
assess predictor size	assess predictor size
select confounders	select confounders
model for outcome	model exposure
check overlap	check overlap
	check for interaction
	compare outcome exposure adjusted for propensity

Major Differences

- Propensity scores (PS) only useful for binary exposure
- Model size can be very different
  PS great for rare outcome/common exposure
- PS involves less modeling of the outcome
- More choices for controlling for propensity
- PS involves many ad-hoc choices

Final Recommendations

- Useful for binary variable
  or possibly ordinal variable
- Switches modeling to exposure
  rather than outcome
  - Advantage if exposure is common,
    outcome is rare
- Makes overlap transparent
- Analysis for like "randomized trial"
  less modeling of the outcome
- Often similar results

Models for Prediction

Dave Glidden
17 November 2015

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