

Models for Prediction

Dave Glidden

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
- $2 \times \dots \times 2$ (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
bmocat	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 cat l=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
- $-2 \times \text{Log-likelihood}$: 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 $-2 \times \log\text{-likelihood}$ is biased
- Assessed on same data use to minimize
- Unbiased version:
- $-2 \times \log\text{-likelihood} - 2 \times K$
- K is # parameters in model
- Call the **Aikake Info. Criterion** (AIC)
- p-value criterion $p < 0.16$

Bayesian Info. Criteria

- A more stringent version
- $-2 \times LL - K \times \log(n)$
- 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

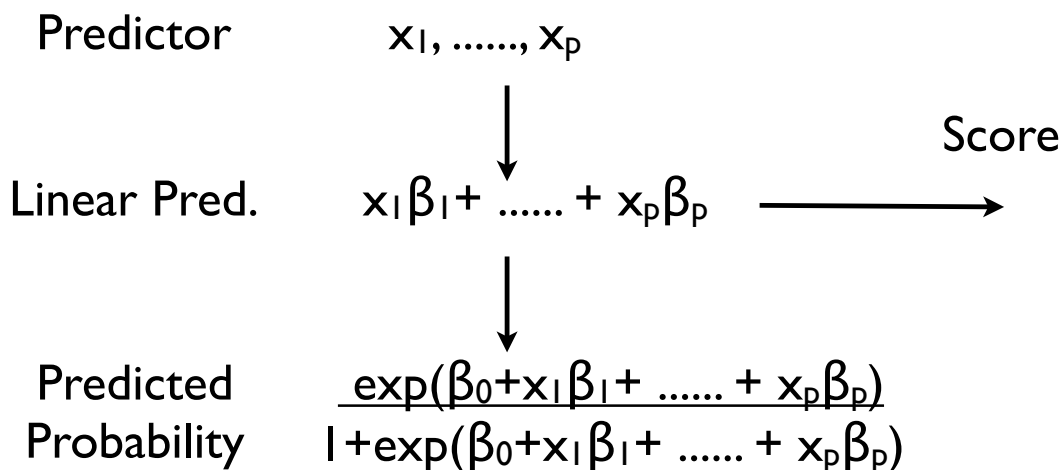
Information Criteria

- 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

Building Prediction Models

- Balance utility, prior knowledge
- Predictors powerful, portable, interpretable, accessible
- Mitigate overfitting: BIC, cross-validation, Construct simple score
- Context dependent cutpoints

Prediction w/ Logistic Regression



17 df predictor model

coefficients

```
. xi: logistic dead i.agecat male hrsdm hrsca hrslung hrschf bmi smoke bath money walkjog push, coef
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	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Iagecat_1	.6345948	.1454731	4.36	0.000	.3494727	.9197169
_Iagecat_2	1.039148	.1413782	7.35	0.000	.7620513	1.316244
_Iagecat_3	1.314136	.1370846	9.59	0.000	1.045455	1.582817
_Iagecat_4	1.689447	.1371779	12.32	0.000	1.420583	1.958311
_Iagecat_5	2.11972	.1429917	14.82	0.000	1.839461	2.399979
_Iagecat_6	2.789075	.1451113	19.22	0.000	2.504662	3.073487
male	.712165	.0704824	10.10	0.000	.5740221	.8503079
hrsdm	.5781861	.0848568	6.81	0.000	.4118699	.7445023
hrsca	.7208027	.0851805	8.46	0.000	.553852	.8877534
hrslung	.8225759	.1244037	6.61	0.000	.5787491	1.066403
hrschf	.851144	.145293	5.86	0.000	.5663751	1.135913
bmicat	.5017212	.0698006	7.19	0.000	.3649147	.6385278
smoke	.7204122	.0927079	7.77	0.000	.5387081	.9021163
bath	.6721684	.1055242	6.37	0.000	.4653448	.8789919
money	.643781	.0961402	6.70	0.000	.4553498	.8322122
walkjog	.7358055	.0786537	9.36	0.000	.5816471	.8899639
push	.4231833	.0791656	5.35	0.000	.2680215	.578345
_cons	-4.903092	.1321857	-37.09	0.000	-5.162171	-4.644013

Creating a Score

- Mimics the linear predictor (lp)
if model is right, lp summarizes risk
- Simple score requires categorical predictor
- Recode them so 1=high risk, 0 = low
- Score is sum of points for each predictor
- Score for jth predictor: $\text{round}(\beta_j / \min(\beta))$

Calculating Points

Push: HR = 1.53

coef = 0.42

points = +1

(smallest coef)

$0.42/0.42$

Walk/Jog HR = 2.1

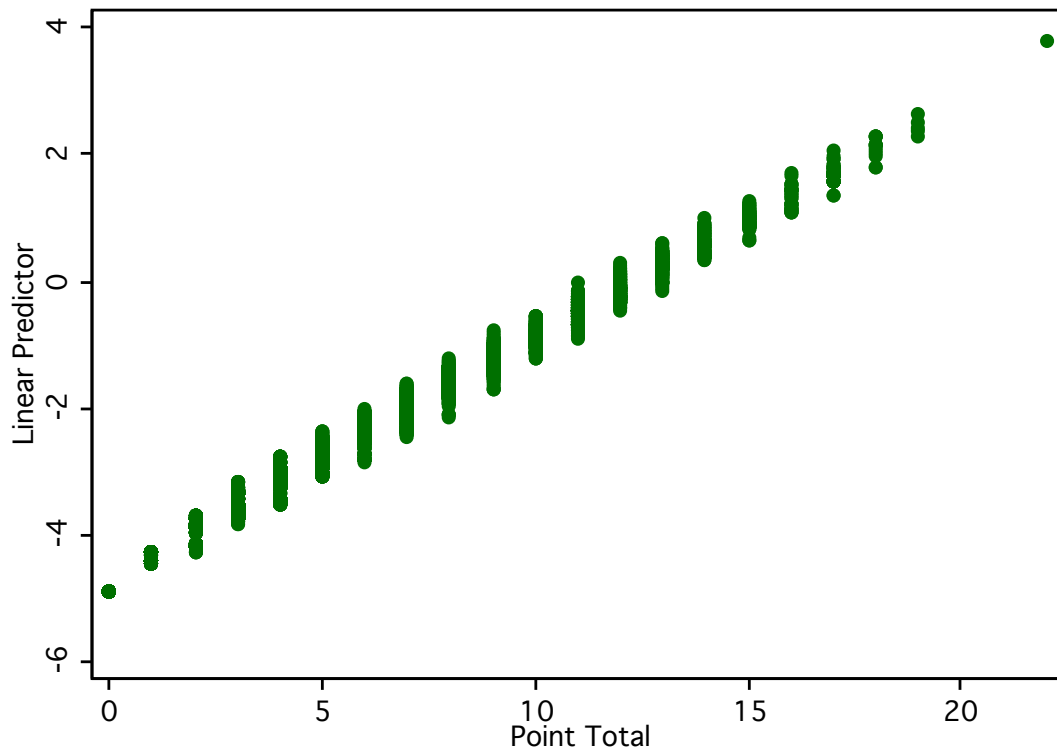
coef = 0.736

points = +2

$0.736/0.42 \approx 2$

Variable	Lee Points
Age 60-64	+1
Age 65-69	+2
Age 70-74	+3
Age 75-79	+4
Age 80-84	+5
Age > 85	+7
Male	+2
Diabetes	+1
Cancer	+2
Lung Disease	+2
Heart Failure	+2
BMI < 25	+1
Smoking	+2
Bathing	+2
Money	+2
Walking	+2
Pushing	+1

Points v. Linear Predictor



Probability of Death

