Models for Prediction

Dave Glidden
26 November 2013

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

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

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

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

Building Prediction Models

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

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

Prediction w/ Logistic Regression

Predictor $x_1, \ldots\ldots, x_p$

Score

Linear Pred. $x_1\beta_1 + \ldots\ldots + x_p\beta_p$

Predicted Probability $$\frac{\exp(\beta_0 + x_1\beta_1 + \ldots\ldots + x_p\beta_p)}{1+\exp(\beta_0 + x_1\beta_1 + \ldots\ldots + x_p\beta_p)}$$

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: round($\beta_j/\min(\beta)$)

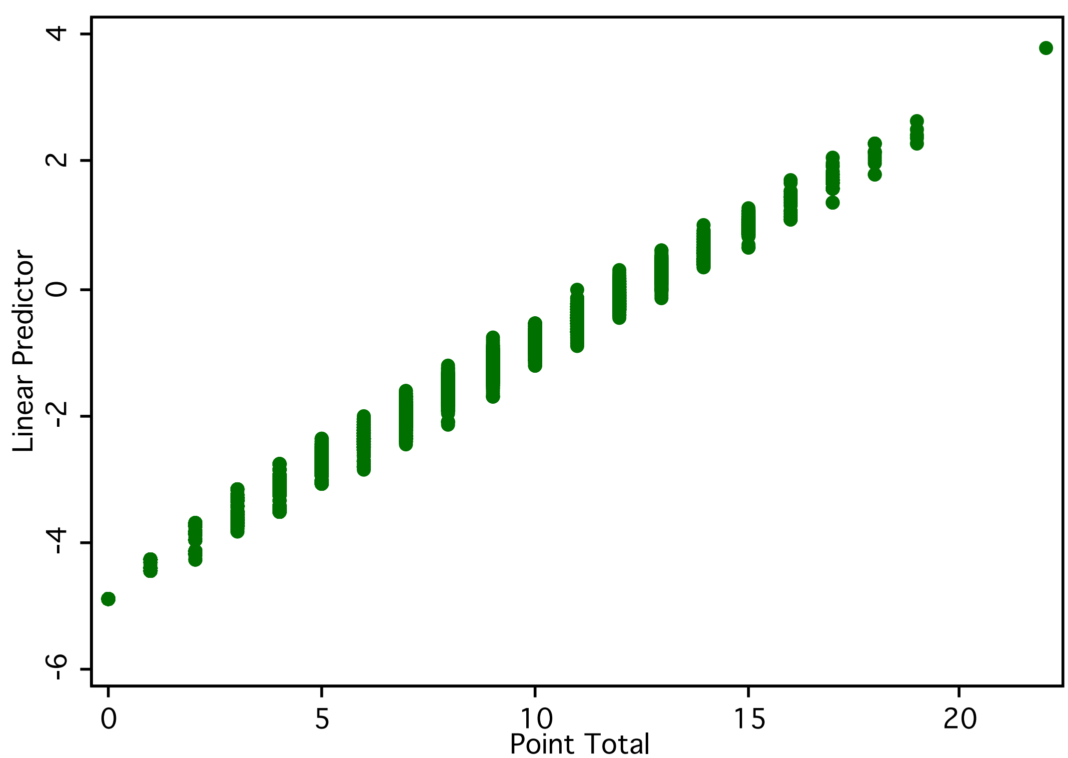
Calculating Points

Push: HR = 1.53
coef = 0.42 (smallest coef)
points = +1 0.42/0.42

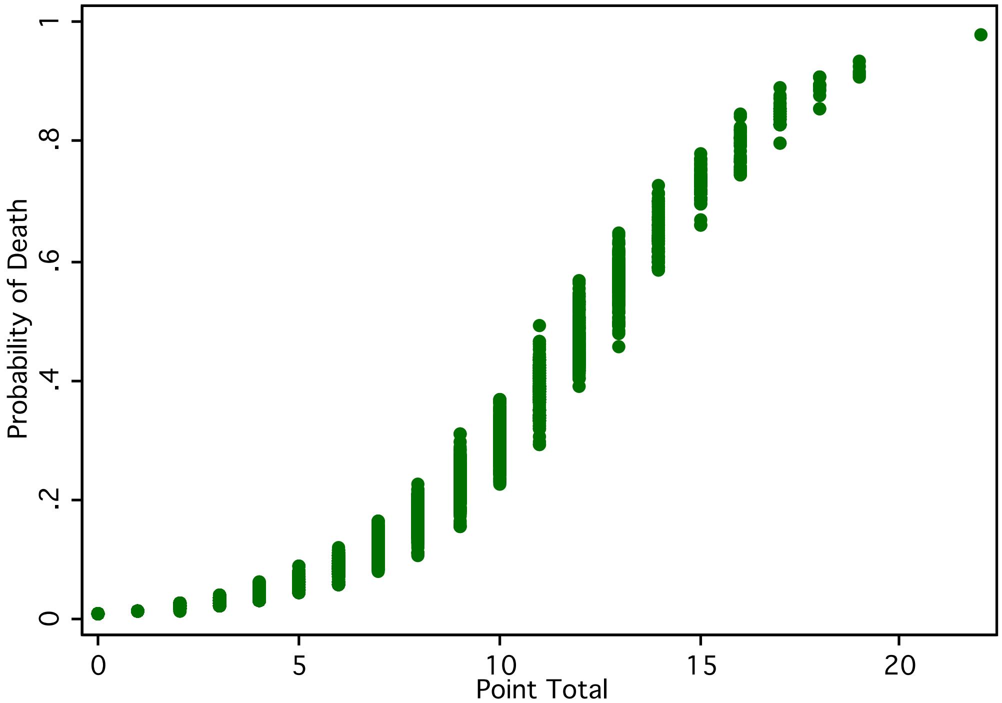
Walk/Jog HR = 2.1
coef = 0.736
points = +2 $0.736/0.42 \approx 2$

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



Last Step

- The score be used directly
- $p_{death} \approx \exp(-4.9+0.42*\text{score})$
- Can also be used predict/stratify
- Would need a cutpoint
  Lee uses quantiles 0-5, 6-9, 10-13, ≥14

End Result

- Score from 0-23
- Lower score -> lower risk of death
- 90% of people have score of 10 or less
- Is it a good summary of risk?
- Can it predict who will die?
- What does a "7" on the score mean?

Model Validation

Model Validation

Altman and Royston

- Statistical and clinical considerations
  - best model for predictors?
  - accurate enough for purposes?
- Interplay between statistic and context

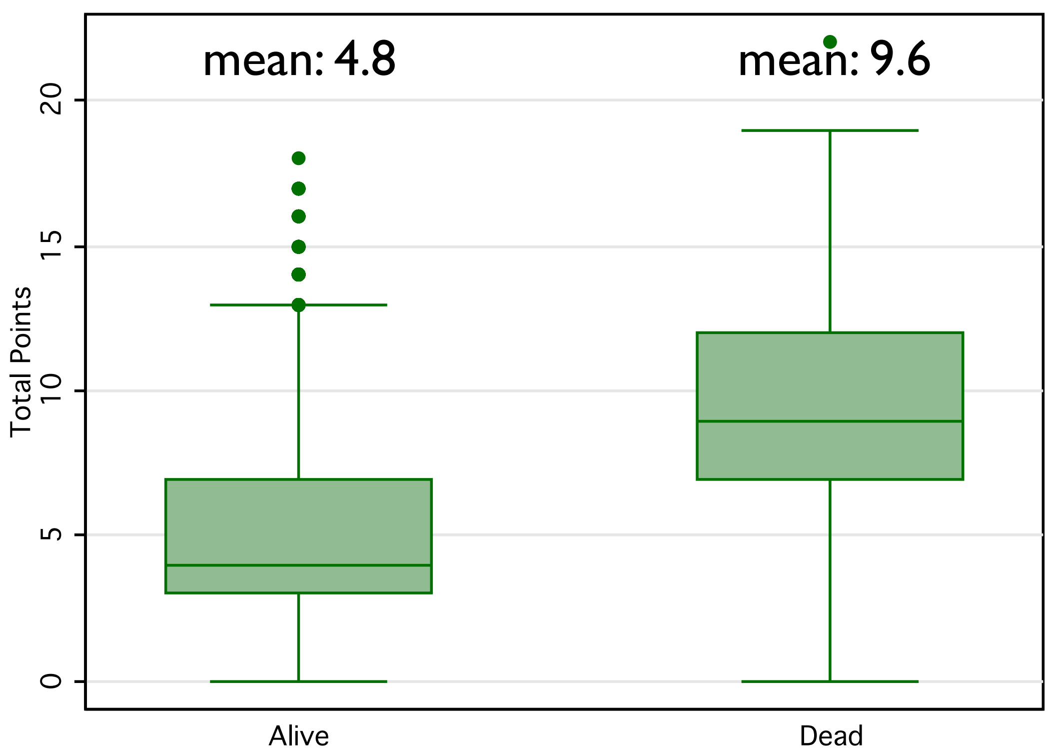
Statistical Fit

- Discrimination
  - Is it a strong model?
  - Can we effectively predict?
- Calibration
  - Are predictions accurate?

Discrimination

- Difference between groups
- Various measures
  - mean difference in score
  - c statistic

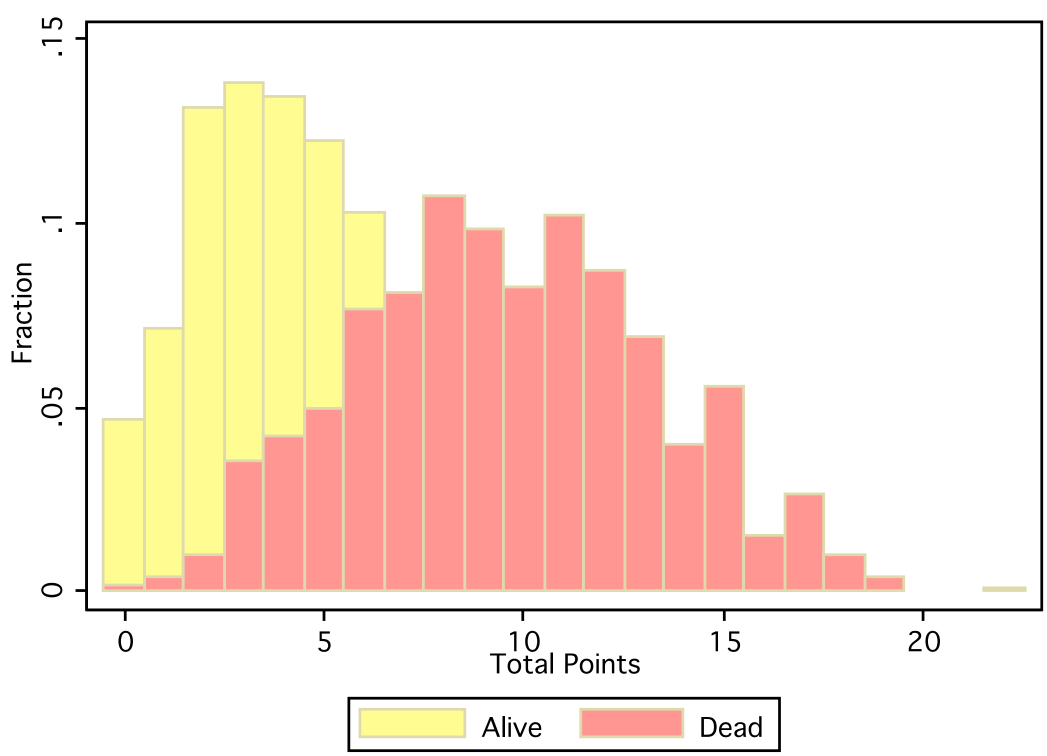
Score by Group



Area Under ROC

- Also known as C-statistic
- Summary measure of AUC curve
- Must be between 0 and 1
- Probability that dead person has more points than a living person
- Measures purity of groups
- Clearly related to discrimination

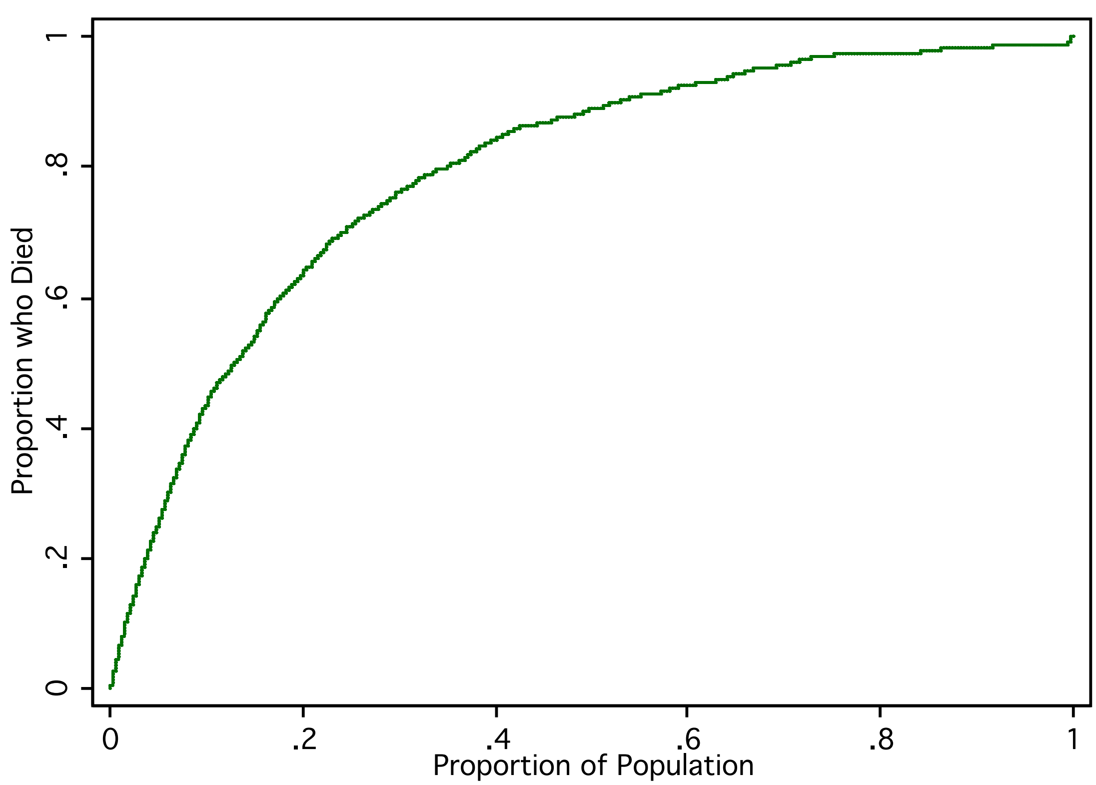
Histogram by Outcome



Lorenz Curve

- Cumulative proportion of outcome by cumulative positive classifications
- Works in prevalence information
- Gives a sense of screening performance
- Also used for income distributions
  10% of pop^n has 40% of mortality risk

Lorenz



```
table points2, c(n prob mean prob)
```

Risk for Each Levels of Points

points2	N(prob)	Pr Death (Model)
0	840	.0073689
1	1,228	.0126862
2	2,255	.0177149
3	2,445	.0277034
4	2,431	.0408263
5	2,172	.0587947
6	1,920	.0850401
7	1,523	.1187199
8	1,256	.1655665
9	952	.2236966
10	715	.2950674
11	552	.3817409
12	436	.4617796
13	333	.5622715
14	184	.6385003
15	194	.7240268
16	55	.7923723
17	86	.8453672
18	31	.8894324
19	14	.9191387
20	1	.9487265
22	1	.9775991
23	1	.981464

Validation Set

- Total dataset 19,710 people available
- 11,701: east, west, central region US
  learning set
- 8009: south
  test set
- Deliberately non-comparable
- Rationale: *assess portability*

Lee Model

validation set

```
. 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   =       7973
                                                  LR chi2(17)     =    1413.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -2423.8528                       Pseudo R2       =     0.2258

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagecat_1 |   1.537822   .2259016     2.93   0.003       1.1531    2.050903
  _Iagecat_2 |   2.248315   .3291624     5.53   0.000     1.687476    2.995549
  _Iagecat_3 |   2.586506    .381196     6.45   0.000     1.937602     3.45273
  _Iagecat_4 |   4.109389   .5813912     9.99   0.000     3.114228    5.422558
  _Iagecat_5 |   7.608406    1.14937    13.43   0.000     5.658564    10.23013
  _Iagecat_6 |   14.62968   2.233807    17.57   0.000     10.84588    19.73353
        male |   2.166964   .1712289     9.79   0.000     1.856058    2.529949
       hrsdm |   1.761092   .1639891     6.08   0.000     1.467303    2.113704
       hrsca |   1.664218   .1689592     5.02   0.000      1.36393    2.030617
     hrslung |   1.587757   .2179521     3.37   0.001     1.213219     2.07792
      hrschf |   2.136519   .3437415     4.72   0.000     1.558685    2.928569
      bmicat |   1.736036   .1361793     7.03   0.000     1.488635    2.024553
       smoke |   1.594206   .1606953     4.63   0.000     1.308409    1.942429
        bath |   1.474593   .1711166     3.35   0.001     1.174616    1.851179
       money |   1.420027   .1534369     3.25   0.001     1.149006    1.754974
     walkjog |   1.995648   .1817115     7.59   0.000      1.66947    2.385553
        push |    1.49066   .1350732     4.41   0.000     1.248098    1.780363
------------------------------------------------------------------------------
```

Lee Model

training set

```
. 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
------------------------------------------------------------------------------
```

Variable	Training Coef.	Valid Coef.	Ratio
Age 60-64	0.63	0.43	0.68
Age 65-69	1.04	0.81	0.78
Age 70-74	1.31	0.95	0.72
Age 75-79	1.69	1.41	0.84
Age 80-84	2.12	2.03	0.96
Age > 85	2.79	2.68	0.96
Male	0.71	0.77	1.09
Diabetes	0.58	0.57	0.98
Cancer	0.72	0.51	0.71
Lung Disease	0.82	0.46	0.56
Heart Failure	0.85	0.76	0.89
BMI < 25	0.50	0.55	1.10
Smoking	0.72	0.47	0.65
Bathing	0.67	0.39	0.58
Money	0.64	0.35	0.54
Walking	0.74	0.69	0.94
Pushing	0.42	0.40	0.94

Variable	Lee Points	Valid Points
Age 60-64	1	1
Age 65-69	2	2
Age 70-74	3	3
Age 75-79	4	4
Age 80-84	5	6
Age > 85	7	8
Male	2	2
Diabetes	1	2
Cancer	2	1
Lung Disease	2	1
Heart Failure	2	2
BMI < 25	1	2
Smoking	2	1
Bathing	2	1
Money	2	1
Walking	2	2
Pushing	1	1