

Prediction Models

Biostatistics 210
3 Dec 2019

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

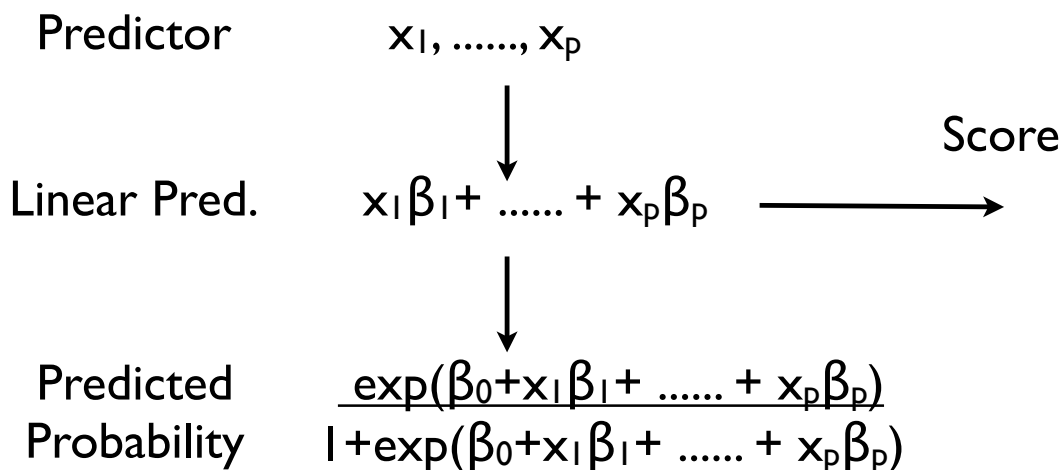
Information Criteria

- 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

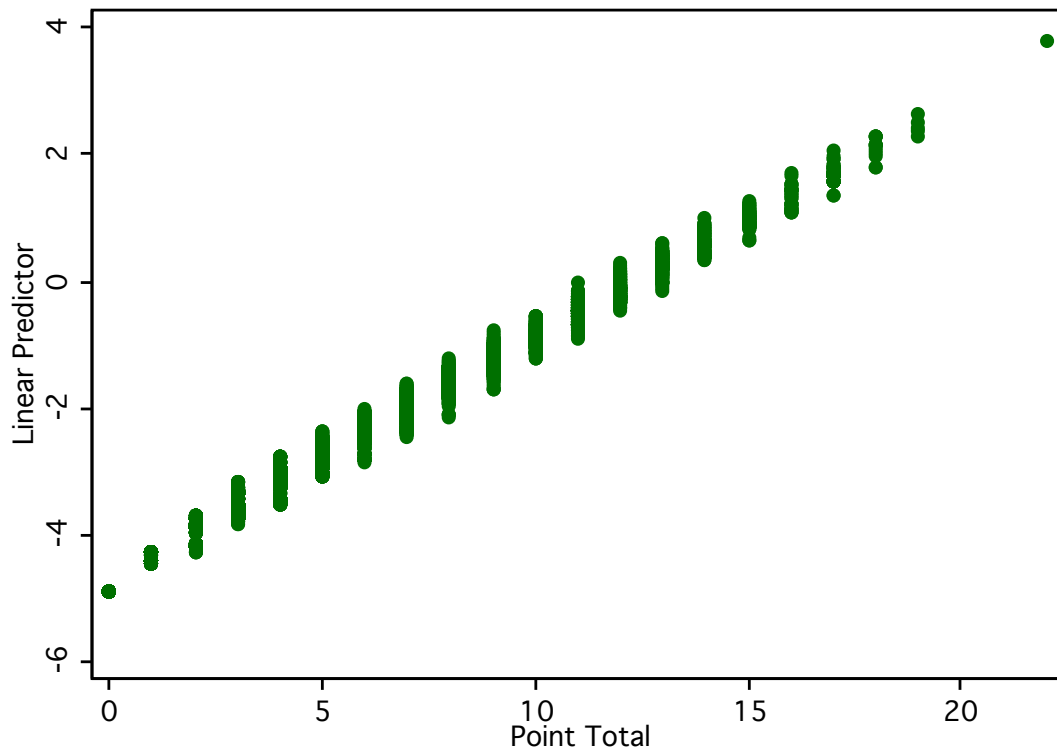
coef = 0.736

points = +2

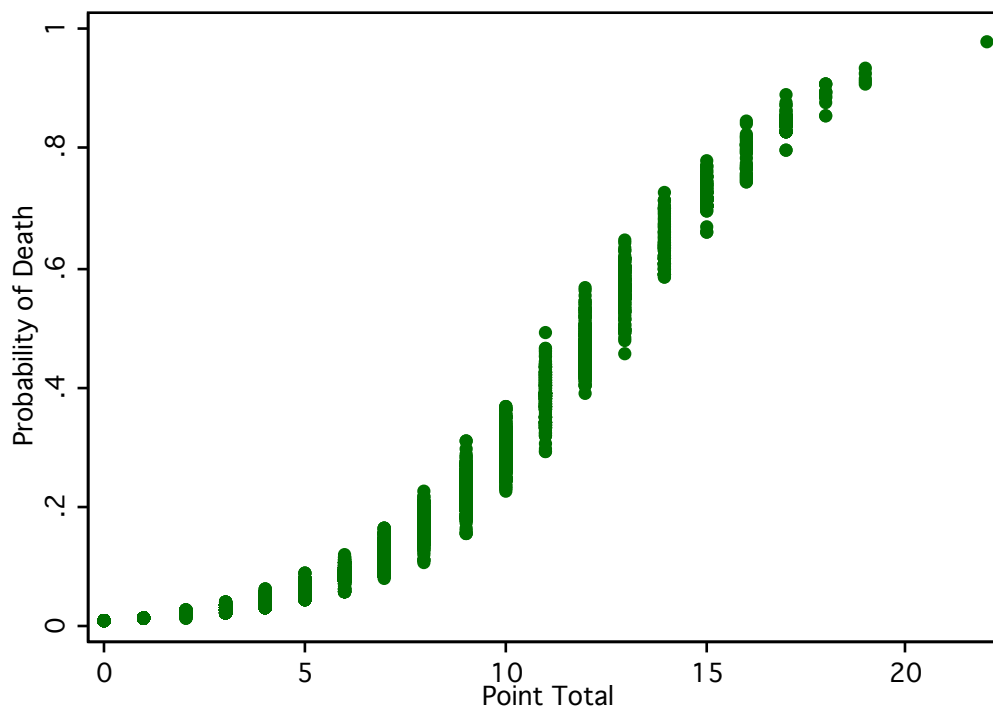
$0.736/0.42 \approx 2$

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



Last Step

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

What groups to form gets to fundamental issues of why you are making a predictive model

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

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

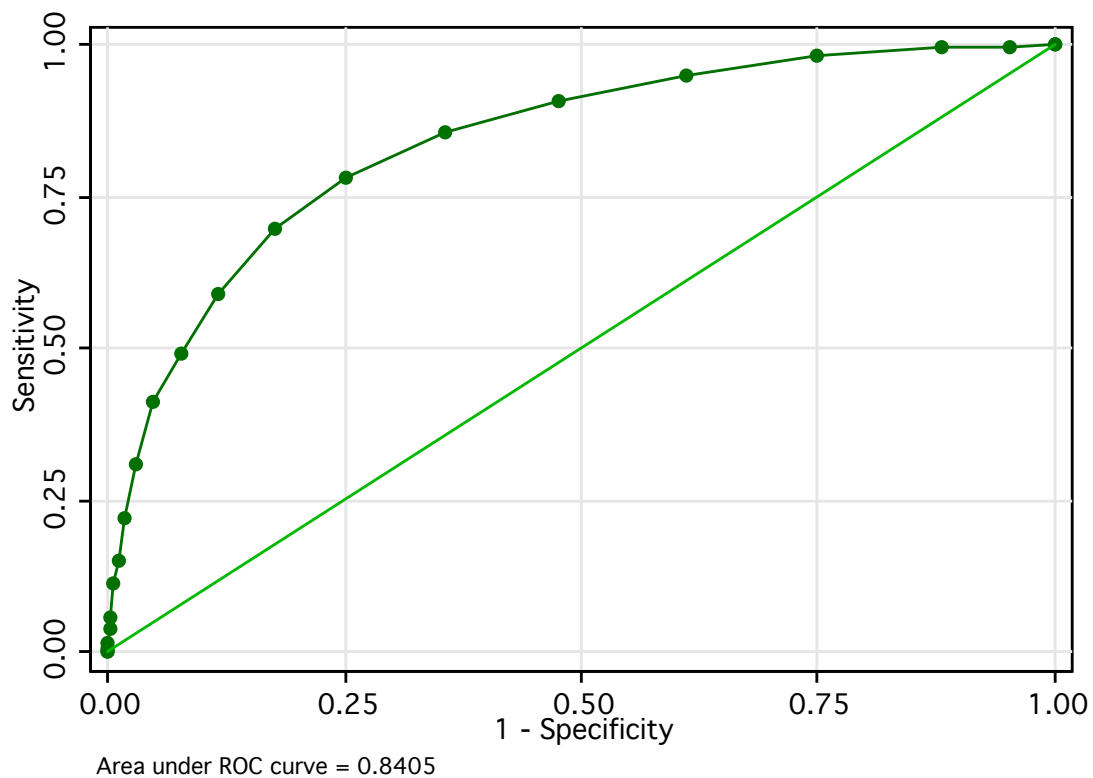
Variable	Training Coef.	Valid Coef.	Ratio
Age 60-64	0.63	0.43	0.68
Age 65-69	1.04	0.81	0.78
Age 70-74	1.31	0.95	0.72
Age 75-79	1.69	1.41	0.84
Age 80-84	2.12	2.03	0.96
Age > 85	2.79	2.68	0.96
Male	0.71	0.77	1.09
Diabetes	0.58	0.57	0.98
Cancer	0.72	0.51	0.71
Lung Disease	0.82	0.46	0.56
Heart Failure	0.85	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

Area Under ROC

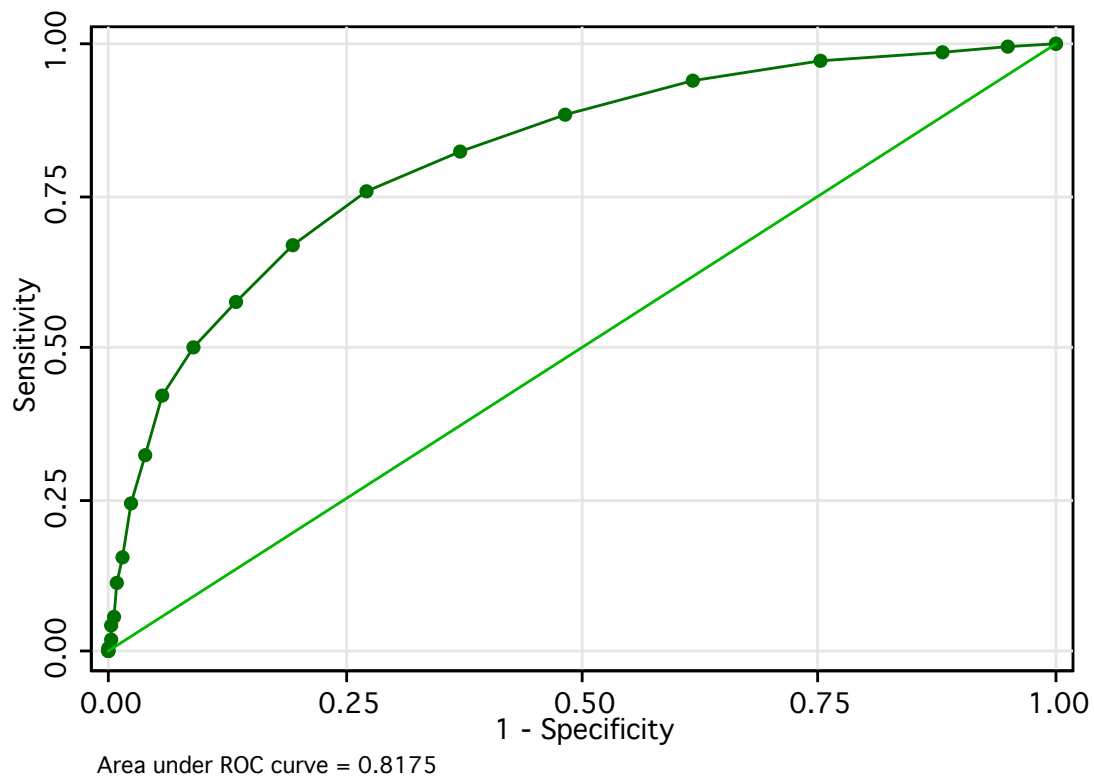
- Also known as C-statistic
- Summary measure of AUC curve *between 0 and 1*
- Probability that dead person has more points than a living person
- Measures purity of groups
- Clearly related to discrimination

C-statistic (test):
0.8424 (linear predictor), 0.8405 (points)

ROC Lee Model: Learn



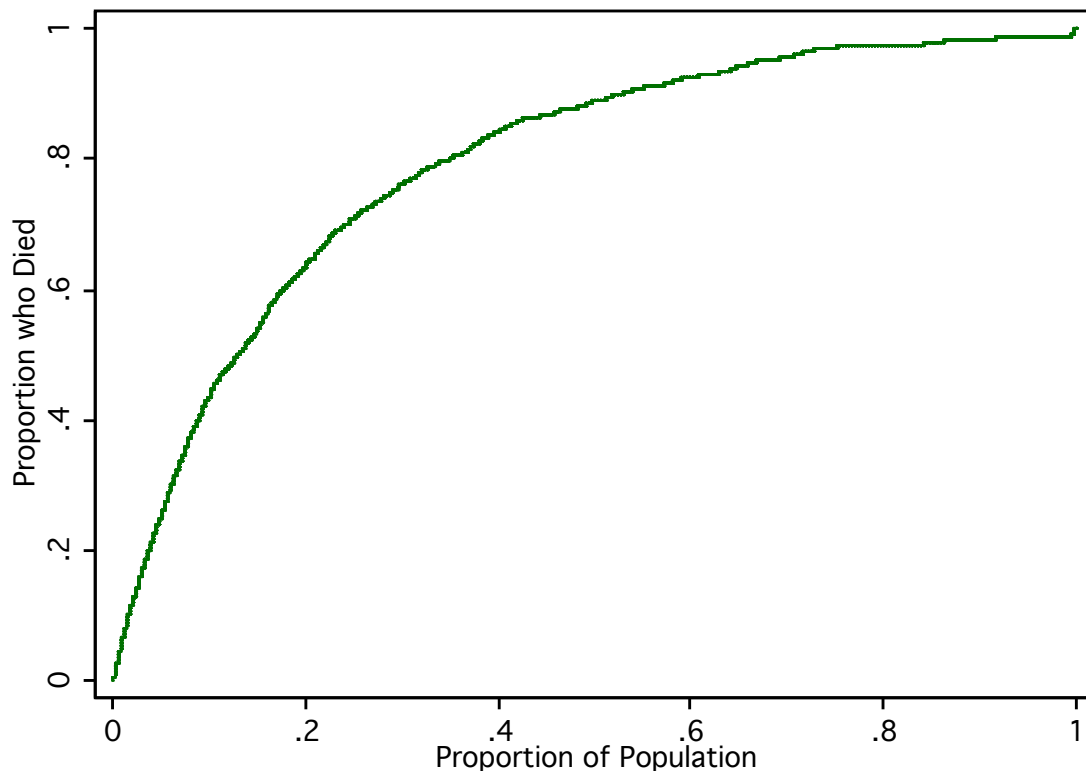
ROC: Test



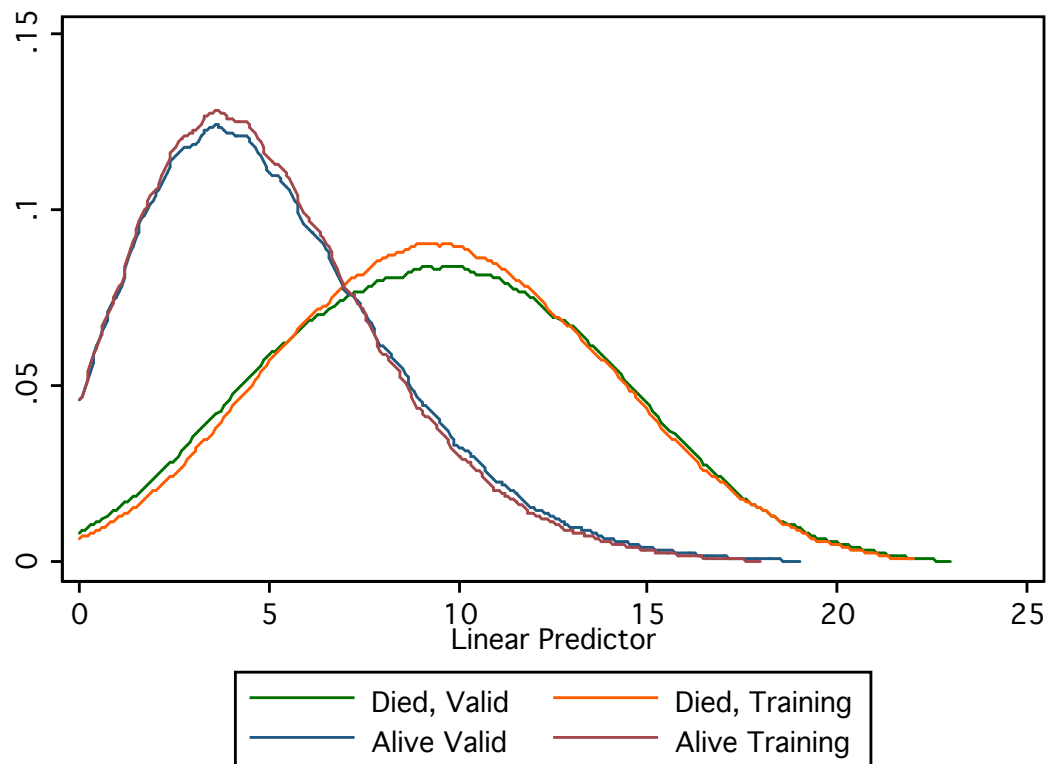
Lorenz Curve

- 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 popⁿ has 40% of mortality risk

Lorenz



Density



Calibration

Calibration

- Model: predicted probability
- Does predicted probability match truth
- Important for individual decision making
- Classic model-checking problem
- Incorporate statistical tests and descriptions of magnitude
- Find model misspecified in some way

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
-----+-----						

Aspects of Model Checking

- Involves
 - wrong functional form
 - undetected interaction

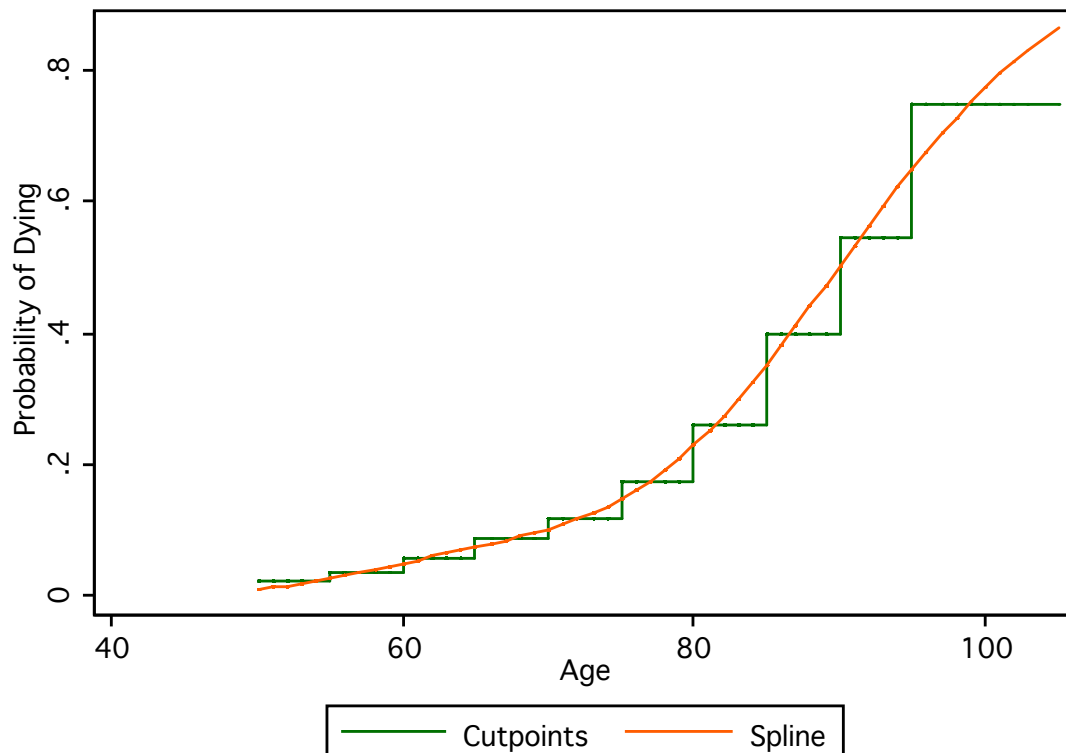
Functional Form

- Involves continuous covariate: e.g., age
- Maybe assumed linear when it is not
- Can be too few cutpoints
 - choice of two
- Result: some ages systematically predicted wrong

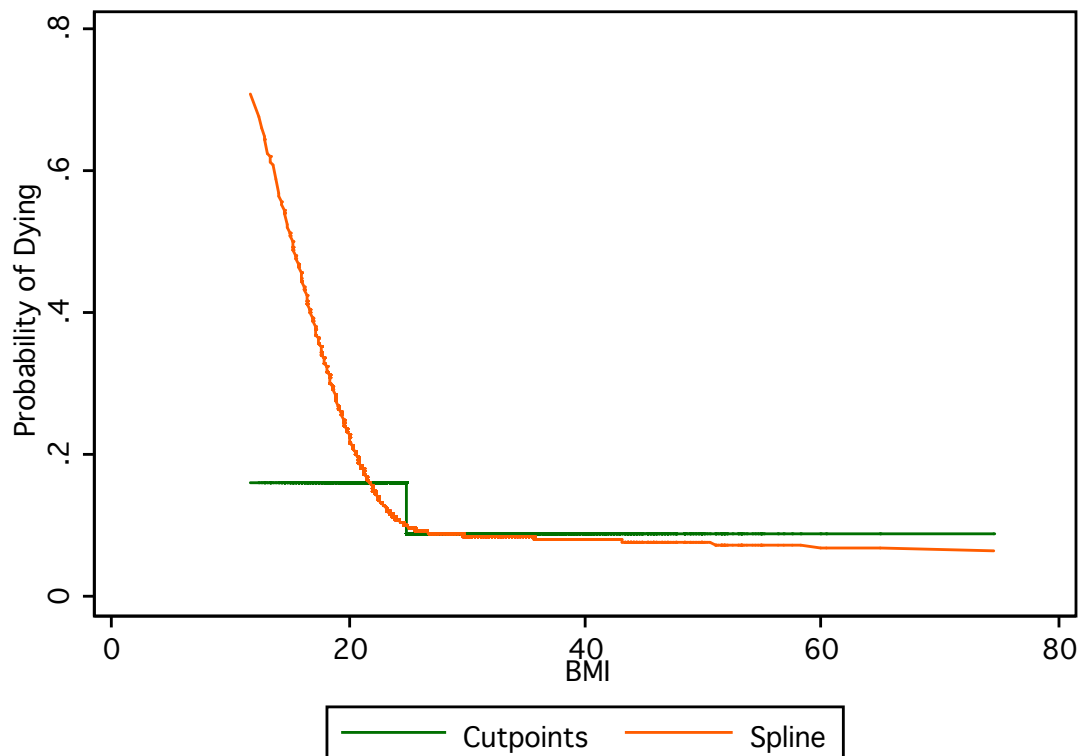
Functional Form

- Fit the spline model to continuous predictor(s)
- Fit the other model
- Plot the difference in fit
- Tests not useful with large or small n

Age



BMI



Functional Form

- Look basically fine
- Age has enough cutpoints
 - and we can afford them
 - underestimate mortality for 90yo
- BMI: only 5% of values less than 20
 - apparent diff => minimal for prediction

Undetected Interaction

- Effect of one variable depends on another
- For someone with cancer, is diabetes important predictor?
- If modeled, points depend on two variables
- Just 17 predictors => 66 possible interactions
- That's just two-way interactions

Inspecting Calibration

- Form series of groups based on risk
- Compare expected mortality with observed
- Expected: average(probs)
- Observed: average(death)
= *proportion who died*
- Within each group
our point total furnishes such a group

Test Set

```
. table pts2 if valid==0, c(n dead mean dead mean prob)
```

Points	N	mean(dead)	mean(prob)
0	486	.0041152	.0073689
1	739	.0067659	.0126751
2	1,366	.0095168	.017859
3	1,474	.0325645	.0277967
4	1,445	.0394464	.0407945
5	1,330	.0503759	.0589664
6	1,162	.0886403	.0849856
7	886	.1230248	.1186558
8	758	.1912929	.1653885
9	551	.2413793	.2242661
10	407	.2727273	.2922267
11	320	.43125	.384265
12	244	.4836065	.4604374
13	174	.5344828	.5641574
14	107	.5046729	.6368158
15+	203	.7438424	.7780437

Validation Set

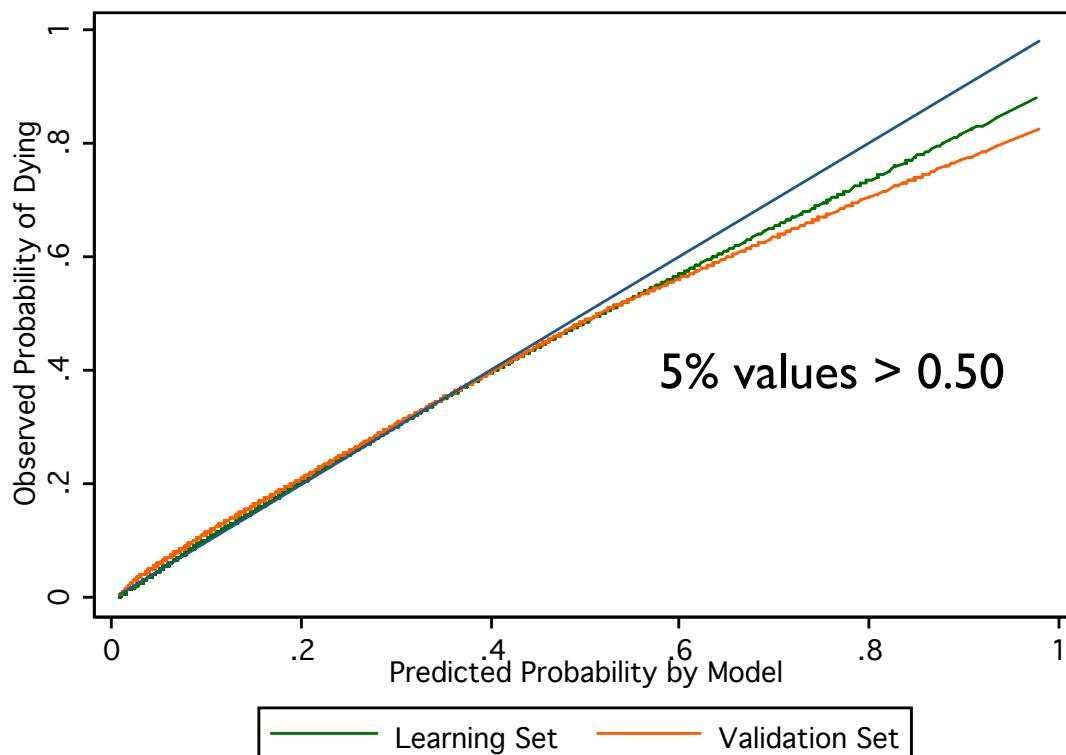
```
. table pts2 if valid==1, c(n dead mean dead mean prob)
```

Points	N	mean(dead)	mean(prob)
0	354	.0056497	.0073689
1	489	.0204499	.0127029
2	889	.0168729	.0174934
3	971	.0360453	.0275617
4	986	.0598377	.0408728
5	842	.0771971	.0585236
6	758	.0936675	.0851236
7	637	.1507064	.1188091
8	498	.1967871	.1658374
9	401	.19202	.2229141
10	308	.275974	.2988213
11	232	.4568965	.3782594
12	192	.4427083	.4634854
13	159	.591195	.5602076
14	77	.5584416	.6408411
15+	180	.6777778	.789714

Smoothing

- Alternative graphical approach
- Draw smooth curve of mortality by predicted probability
- Should look like straight line
- Assessed in test and validation set

Smoothed Calibration



Summary

- See classic overfitting pattern
 - mortality lower than pred at high end
 - mortality higher than pred at low end
- Large calibrations at high end
 - represent 5% of the prediction
 - fixing would complicate model

Validation

- Calibration: examination of model spec
 - largely internal
- Discrimination: measure separation
 - largely external
- Calibration can be fixed
- Discrimination dictates utility of model