

# Model Validation

Dave Glidden  
17 November 2009

## Mortality in Elderly

- Cohort of community-dwelling elders
- Predictors: age, co-morbidities, functional status
- Follow-up: mortality over 4 years
- Who is at the highest risk of death?

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

# Lee Model

## training set

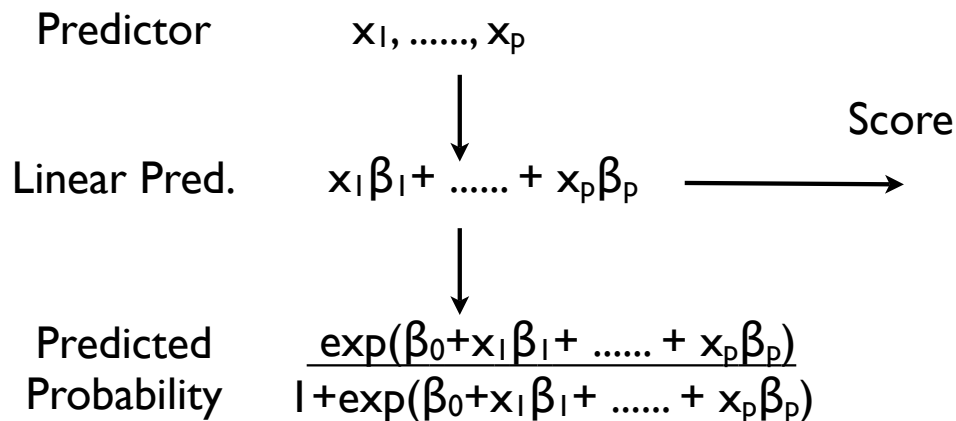
```
. xi: logistic dead i.agecat male hrsdm hrsca hrslung hrschf bmi smoke bath money walkjog push
i.agecat      _Iagecat_0-6      (naturally coded; _Iagecat_0 omitted)
```

```
Logistic regression      Number of obs   =   11652
                          LR chi2(17)           =   2080.12
                          Prob > chi2           =   0.0000
                          Pseudo R2            =   0.2493

Log likelihood = -3132.1774
```

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

# Prediction w/ Logistic Regression



# 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  $j$ th predictor:  $\text{round}(\beta_j/\min(\beta))$

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

# 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

Altman and Royston

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

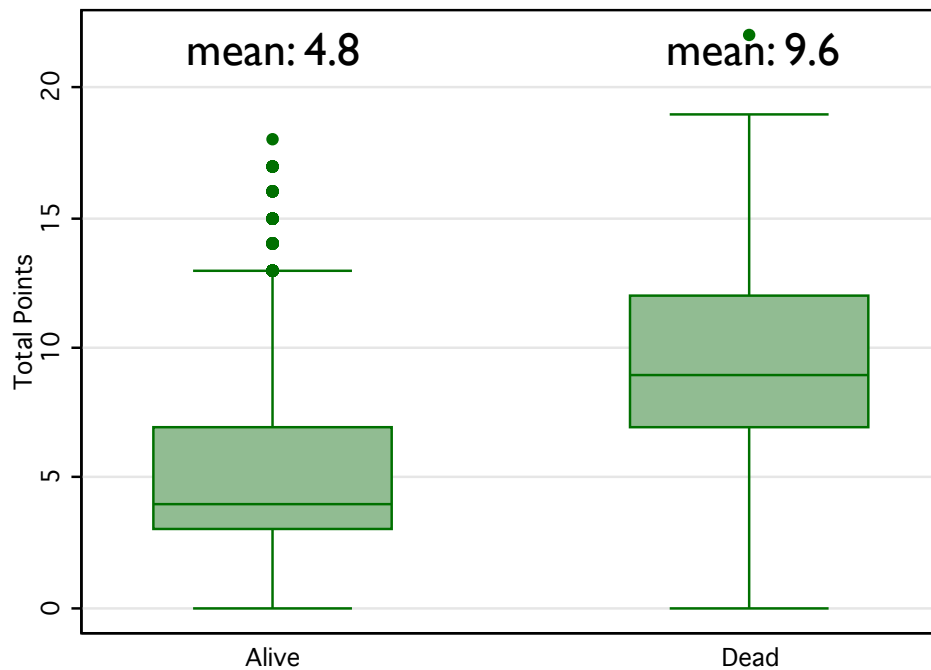
# Statistical Fit

- 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

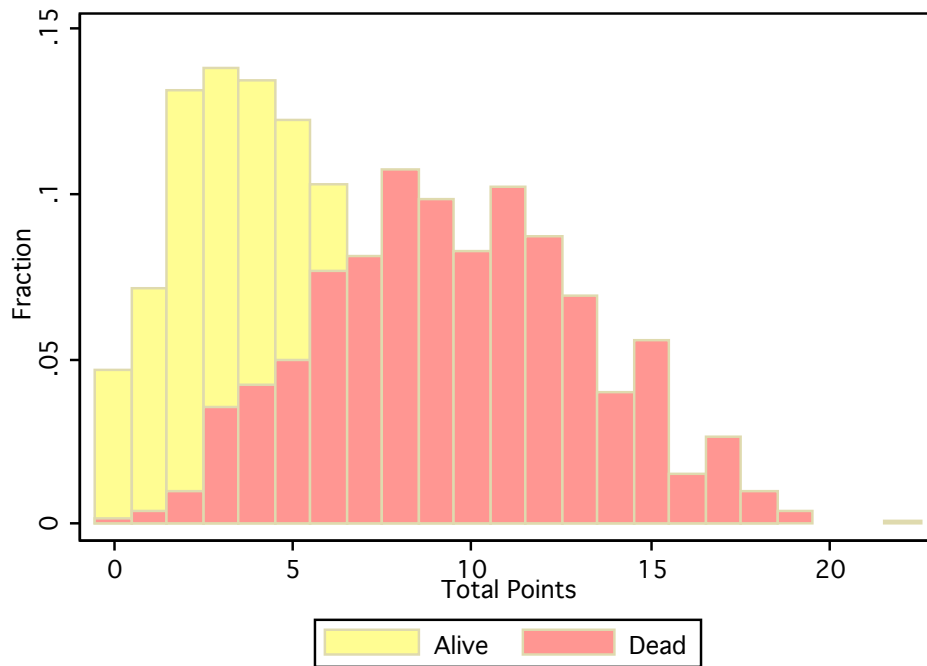
# Score by Group



## 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 point than a living person
- Measures purity of groups
- Clearly related to discrimination

# Histogram by Outcome



## Statistical Fit

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

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

Risk for Each  
Levels of Points

points2	N(prob)	mean(prob)
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

## Assessing Overfitting

- In training, selected most significant variables
- Estimate of OR are biased  
*biased away from the null*  
*bias depends on n, # var, and strictness*
- Validation set allows assessment of avg. bias
- How much to discount HRs

# Assessing Overfitting

## training set

```
. logistic dead lp if valid==1, coef
```

```
Logistic regression          Number of obs   =       7973
                             LR chi2(1)            =    1375.05
                             Prob > chi2           =       0.0000
Log likelihood = -2443.1414   Pseudo R2       =       0.2196
```

dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lp	.8777149	.0270914	32.40	0.000	.8246168	.930813
_cons	-.0964505	.0565834	-1.70	0.088	-.2073518	.0144509

## 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

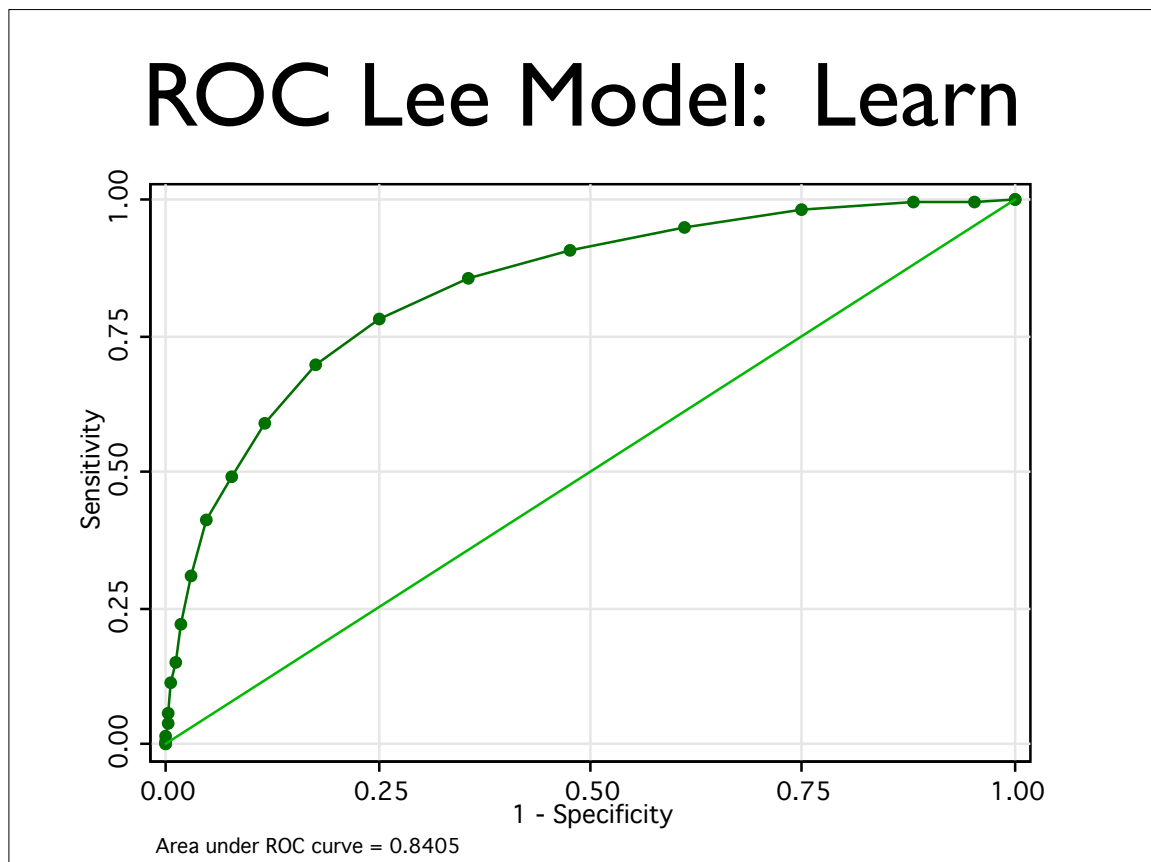
## training set

```
. 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)
```

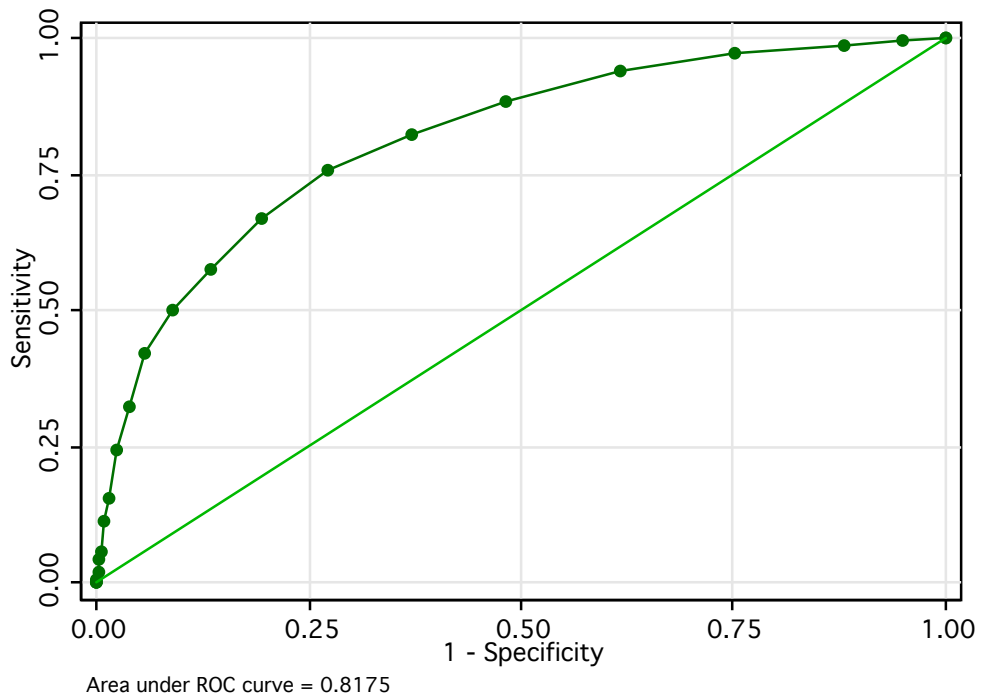
```
Logistic regression      Number of obs   =     11652
                        LR chi2(17)      =     2080.12
                        Prob > chi2      =      0.0000
Log likelihood = -3132.1774      Pseudo R2      =      0.2493
```

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

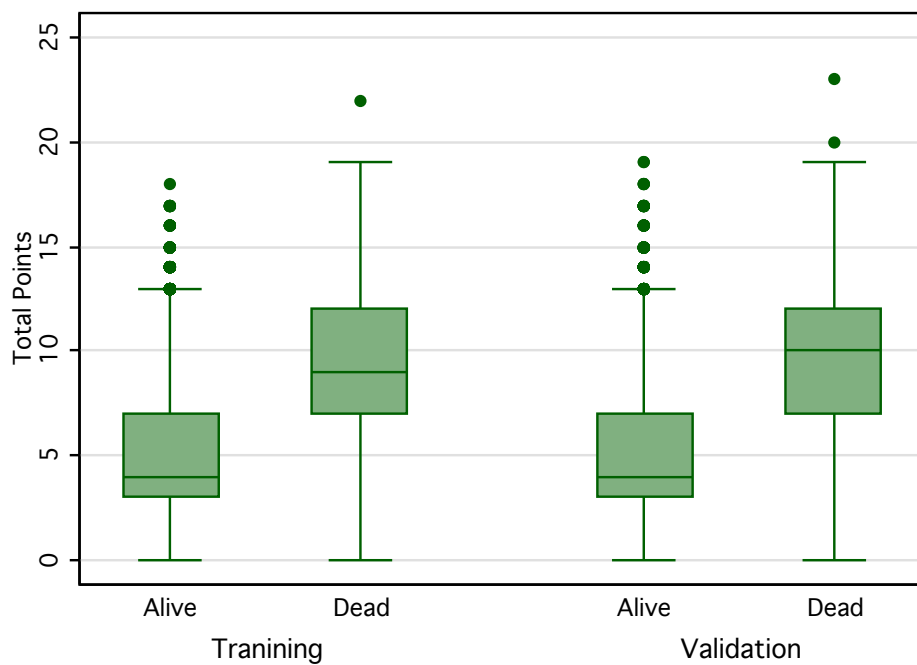
# Discrimination



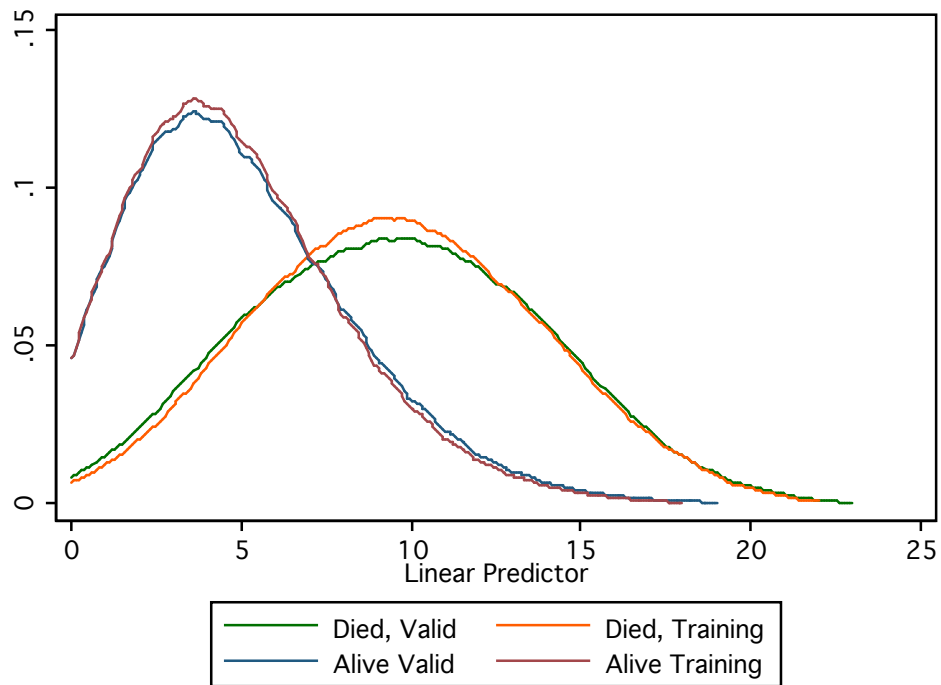
# ROC: Test



# Boxplot of Risk Scores



# 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

# Degree of Overfitting

- Measured by  $R = \beta_{\text{valid}}/\beta_{\text{train}}$
- Overfitting results in  $R < 1$
- If  $R$  is the same of each variable
- $LP_{\text{train}} = X_1 \beta_{\text{train}, 1} + \dots + X_p \beta_{\text{train}, p}$
- Fit validation set with  $LP_{\text{train}}$  as predictor
- Coefficient estimates  $R$

# Assessing OR Inflation

- $\beta_{\text{valid}}/\beta_{\text{train}}$  can measure OR discounting
- $\beta_{\text{valid, age 1}} = 0.87 * \beta_{\text{train, age 2}}$
- $OR_{\text{valid, age 1}} = [OR_{\text{train, age 2}}]^{0.87}$
- $[1.88]^{0.87} = 1.73$
- Observed value was 1.54

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

# Checking the Score validation set

```

Logistic regression
Log likelihood = -2423.8528
Number of obs = 7973
LR chi2(17) = 1413.63
Prob > chi2 = 0.0000
Pseudo R2 = 0.2258
  
```

dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_lagecat_1	.0470738	.1371089	0.34	0.731	-.2216546 .3158022
_lagecat_2	.0435945	.1288616	0.34	0.735	-.2089696 .2961587
_lagecat_3	-.1995713	.1249111	-1.60	0.110	-.4443927 .04525
_lagecat_4	-.1198981	.1151314	-1.04	0.298	-.3455516 .1057553
_lagecat_5	.1127879	.1264054	0.89	0.372	-.134962 .3605379
male	.0067407	.0881151	0.08	0.939	-.1659617 .1794431
hrsdm	.1826408	.0942081	1.94	0.053	-.0020036 .3672852
hrsca	-.2572313	.1113103	-2.31	0.021	-.4753954 -.0390672
hrslung	-.3042641	.1419048	-2.14	0.032	-.5823923 -.0261359
hrsCHF	-.0074083	.166359	-0.04	0.964	-.333466 .3186493
bmecat	.1683109	.084531	1.99	0.046	-.0026332 .3339887
smoke	-.3002105	.0994421	-3.02	0.003	-.4951134 -.1053075
bath	-.3782041	.125638	-3.01	0.003	-.6244501 -.1319581
money	-.4159105	.1185917	-3.51	0.000	-.648346 -.183475
walkjog	-.0756177	.1058295	-0.71	0.475	-.2830397 .1318043
push	.0159256	.0928418	0.17	0.864	-.166041 .1978922
points2	.3832932	.0218129	17.57	0.000	.3405407 .4260456
_cons	-4.411714	.1329939	-33.17	0.000	-4.672377 -4.15105

## Aspects of Model Checking

- Involves
  - wrong functional form
  - undetected interaction

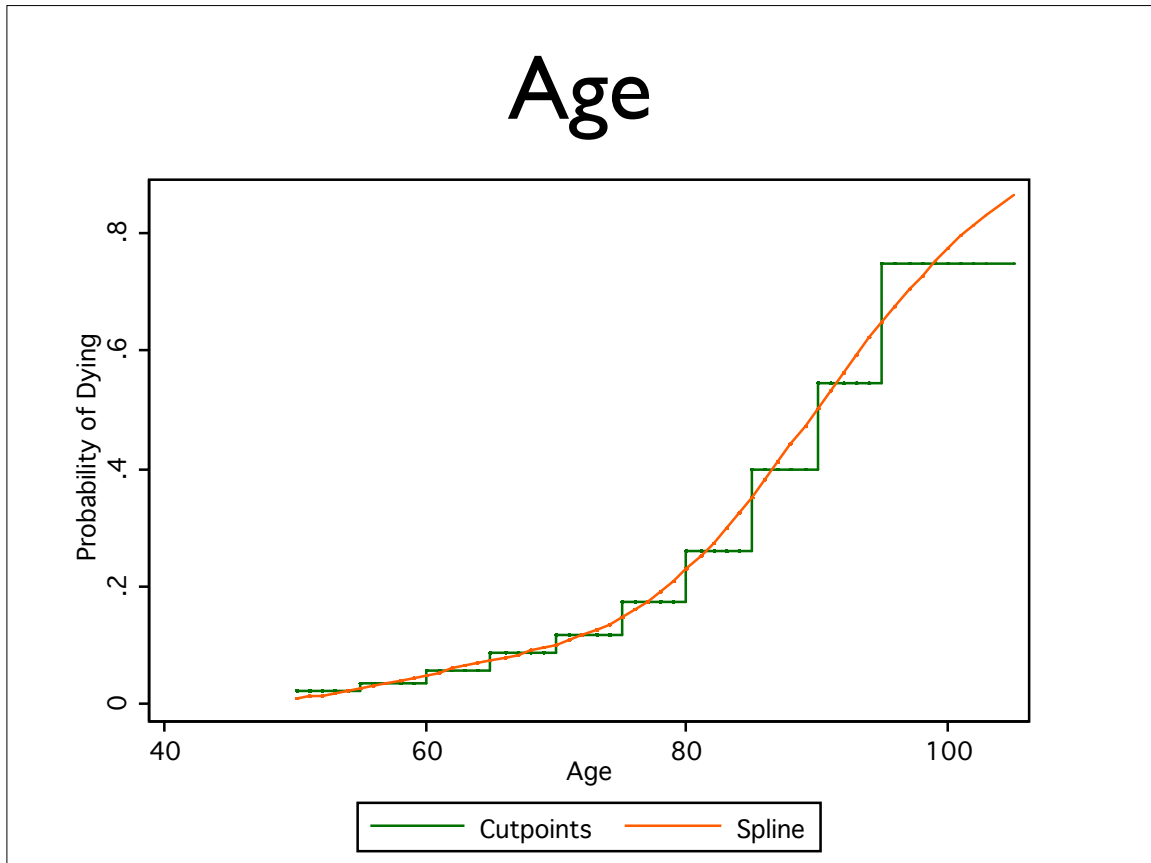
# Function Form

- Lecture #4
- 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

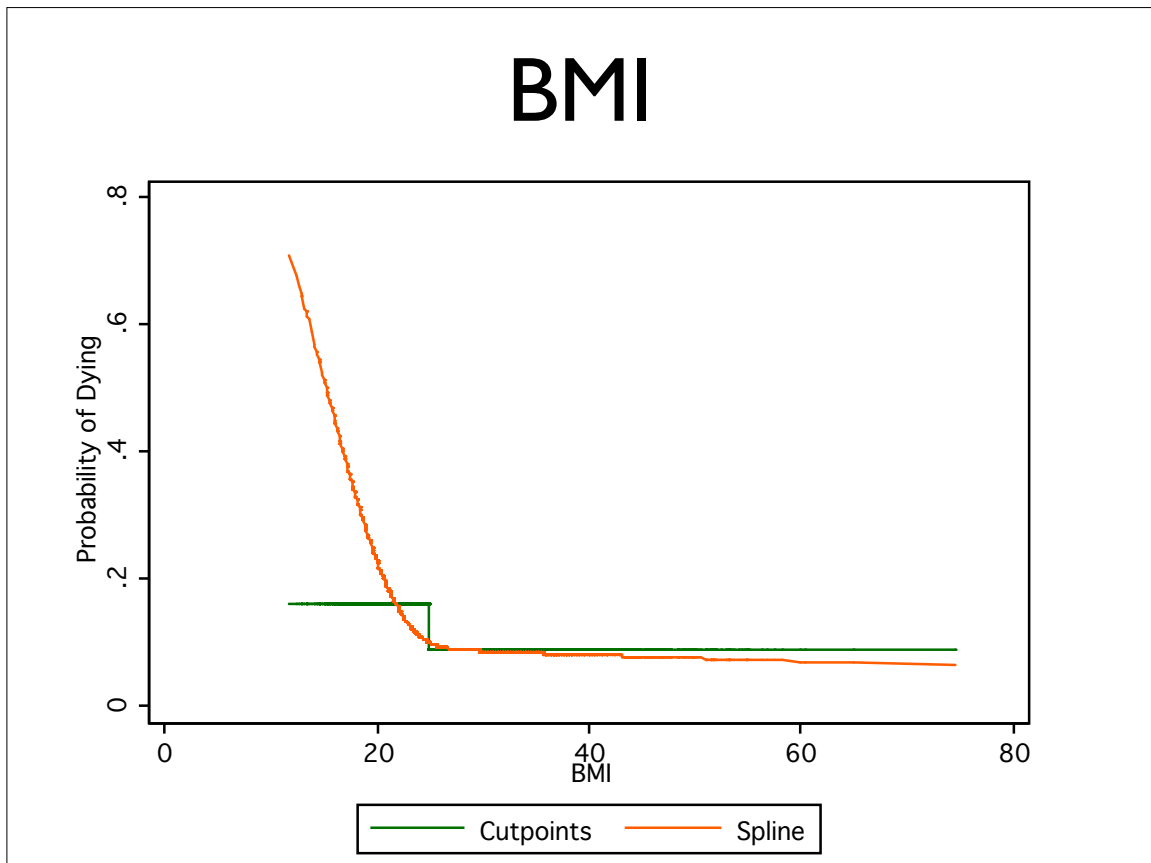
# Functional Form

- Best tool is splines
- Fit the spline model to continuous
- 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 an 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

# Interactions with Age

Variable	Age < 60	Age 60-80	Age > 80
Male	2.69	2.15	1.81
Diabetes	2.03	1.94	1.22
Cancer	5.29	2.40	1.34
Lung Disease	3.96	2.08	1.72
Heart Failure	5.26	2.11	2.10
BMI < 25	1.38	1.78	1.95
Smoking	1.71	1.81	1.50
Bathing	1.04	2.06	2.33
Money	1.38	1.87	2.18
Walking	2.41	2.48	1.79
Pushing	1.24	1.60	1.44

# 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)
```

```
-----
```

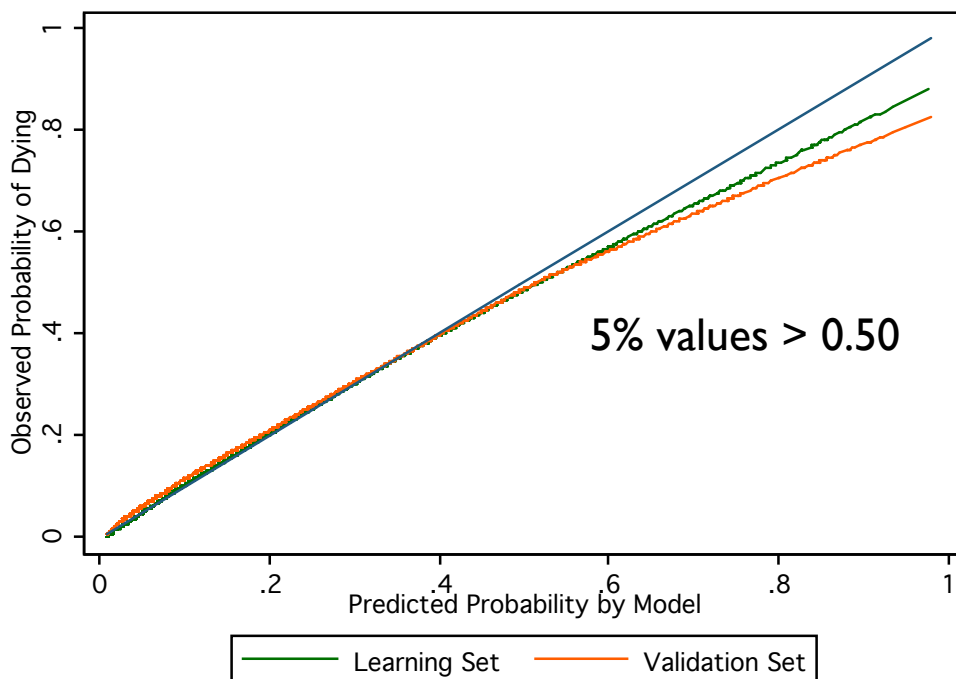
Points	N	mean(dead)	mean(prob)
0	354	.0056497	.0073689
1	489	.0204499	.0127029
2	889	.0168729	.0174934
3	971	.0360453	.0275617
4	986	.0598377	.0408728
5	842	.0771971	.0585236
6	758	.0936675	.0851236
7	637	.1507064	.1188091
8	498	.1967871	.1658374
9	401	.19202	.2229141
10	308	.275974	.2988213
11	232	.4568965	.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