Models for Prediction

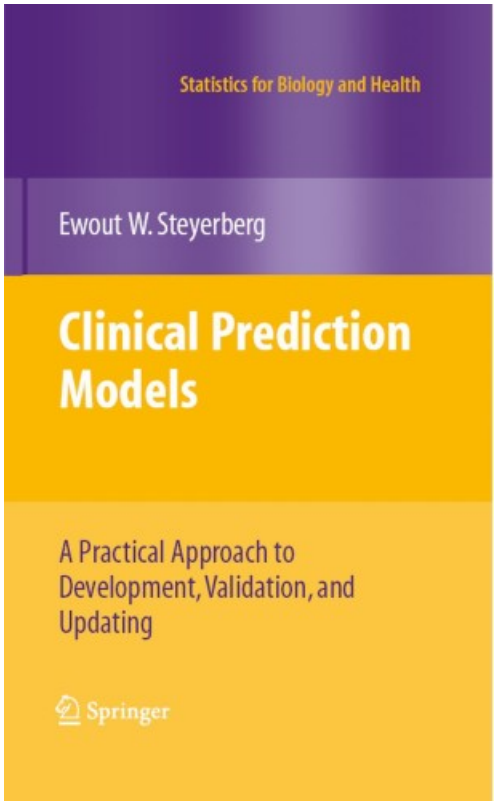
Dave Glidden

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

http://dx.doi.org/10.1007/978-0-387-77244-8



Comprehensive text

Lecture: Ch 5-6, 9-12

Link on the CLE

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

Prediction: Paradigm Shift

- Calculate mean/prob based on Z
- Develop an index predicting outcomes
- Group into strata with similar outcome
- Classification: predict will/won't have event

Objective: Use predictors to minimize prediction error

Machine Learning

- Has been extra active in last two decades
- Powerful approaches from computer science machine learning, deep learning
- Some algorithms are black boxes predictions are not transparent
- Regression is not as flexible as these but can take us very far

Regression

Generalized Linear Model: includes linear and logistic models

- $g(E(Y|Z)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_P$

  link family linear predictor

- Uncertainty
  - which model
  - which variables (lots of choice here)
  - estimate β

Unavoidable Problem

- Apparent Goal: minimize prediction error
- "Once something becomes a metric it ceases to be a good measure"
- Actual goal: minimize prediction error in future samples
- The twin problems: optimism and overfitting

Overfitting

Overfitting is a broad term that refers to entering too many terms into the model -- for the amount of available data.

You fit not only signal, but quirks, of your sample

This produces unstable, variable predictions. The extreme ones *tend* to be both spurious and also *tend* to get the most focus.

Hence, results are frequently unreplicable.

No True Heterogeneity

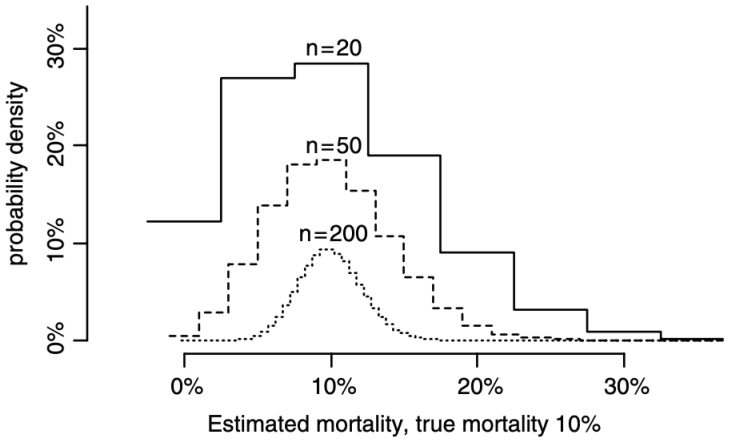


Fig. 5.2 Estimated mortality in relation to sample size. When the true mortality is 10% in samples of n = 20, around 30% of these will contain two deaths (estimated mortality, 10%). With larger sample sizes, observed mortalities are more likely close to 10%

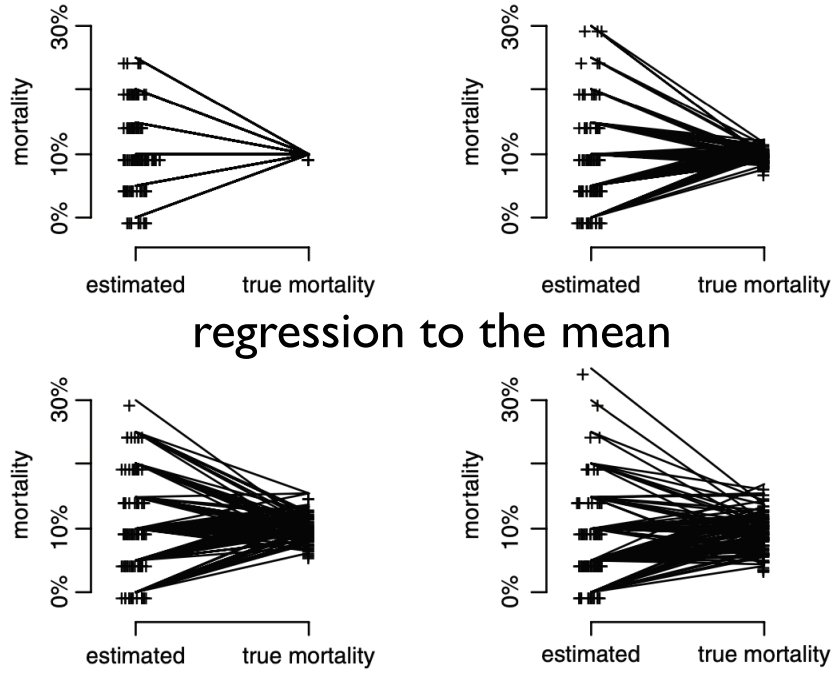


Fig. 5.3 Estimated and true mortality for 100 centres that analyzed 20 subjects each, while the average mortality was 10% for all (*upper left panel*), 10% ± 1% (*upper right panel*), 10% ± 2% (*lower left panel*), 10% ± 3% (*lower right panel*)

Insights: Overfitting

- When estimating groups of things,
  best looks better, worst looks worse
  (regression to the mean)
- Mitigated by actual heterogeneity
- Mitigated by increase n per unit
  df per predictor

Implications

- Need to consider full set of estimations
- The best predictor looks too good
- Best model: bad prognosis too bad, good prognosis too good
- Best model for data will not be best model in replication
- Should calibrate data search with data size

Optimism

- Our estimates look better than they actually are
- Minimizing sample prediction error does not minimize population prediction error
- Once we use the data to decide what to look at, we introduce bias
- "Testimation" bias

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
- Keep variables with $p < 0.05$
- 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

Cost of Data Analysis

- Don't want to misspecify model
  if adapt, we become specific to training data
- More data drive choice => more overfitting
- Bigger n => less overfitting
- Data hungry decision => more overfitting
  regression trees try to split at optimal cutpoint

how to balance curiosity and skepticism

Selecting Variables: Restrict

- Plausibility
  is there literature supporting it?
- Missing values
  prefer vars with complete values
- Narrow distribution
  what if 3% have the trait? restrict. drop.
- Combine within families: activities
  any activity restricted? number of activities restricted?
- Avoid context specific: "foreign-born"
- Avoid experience/hard to measure?

Collinearity

- Correlation ρ between predictor
- Models can handle high corr ($\rho > 0.80$)
- If high ρ, combined vars or choose 1
- Collinearity make coefficient unstable
  but hardly affects predictions
- Overall, it is a minor issue

Universe of Possible Models

- 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

Selecting Variables

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., $p < 0.05$)

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

Steyerberg Advice
Predictor Selection

- Pick variables from among candidates
- Fit model with all of them
- Rationale:
  - probably have few "false" predictors
  - false predictors have low cost
  - selection could lead to under/over fitting

Downside of Selection

Uncertainty: Model Form

- Our models assume some form of additivity and linearity
- Is every year of age have same effect
  inspect with splines
- Are there interactions?
  need interaction terms
- Steyerberg: inspect only if effects are plausible or if effective sample size can tolerate

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

Mitigating Overfitting

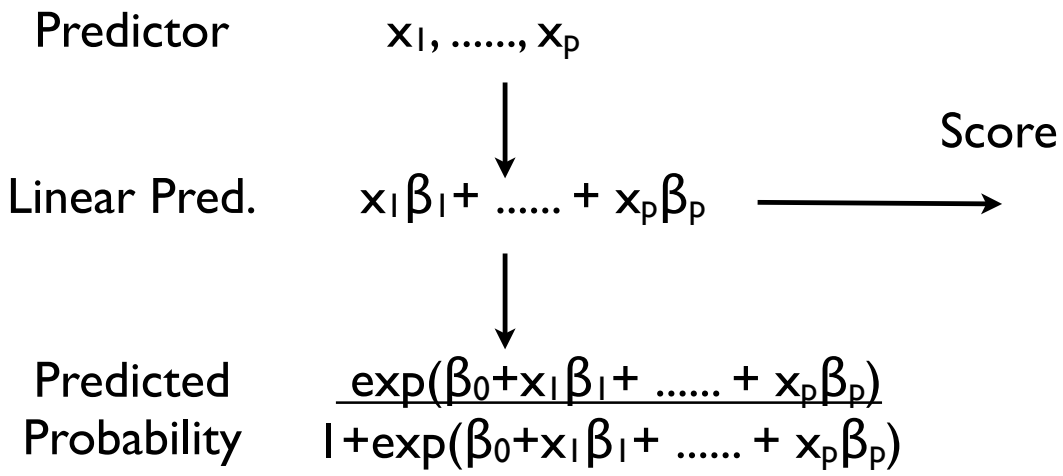
Statistically

- Reduces overfitting in pre-specified model
  the overfitting due to betas
- "Shrinks" coefficients toward 0
- Smooths the data
- Particularly interesting example is LASSO
  can shrink some coefficients to 0
- The "0" functions as variable selection

Building Prediction Models

- Beware overfitting/underfitting degrades prediction
- External data to select subset of predictors
- Compare AIC/BIC of subset v. full model
- If full model is superior in AIC/BIC use all
- If subset is superior in AIC/BIC, use it
- LASSO is also helpful
- Interactions only if external knowledge

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

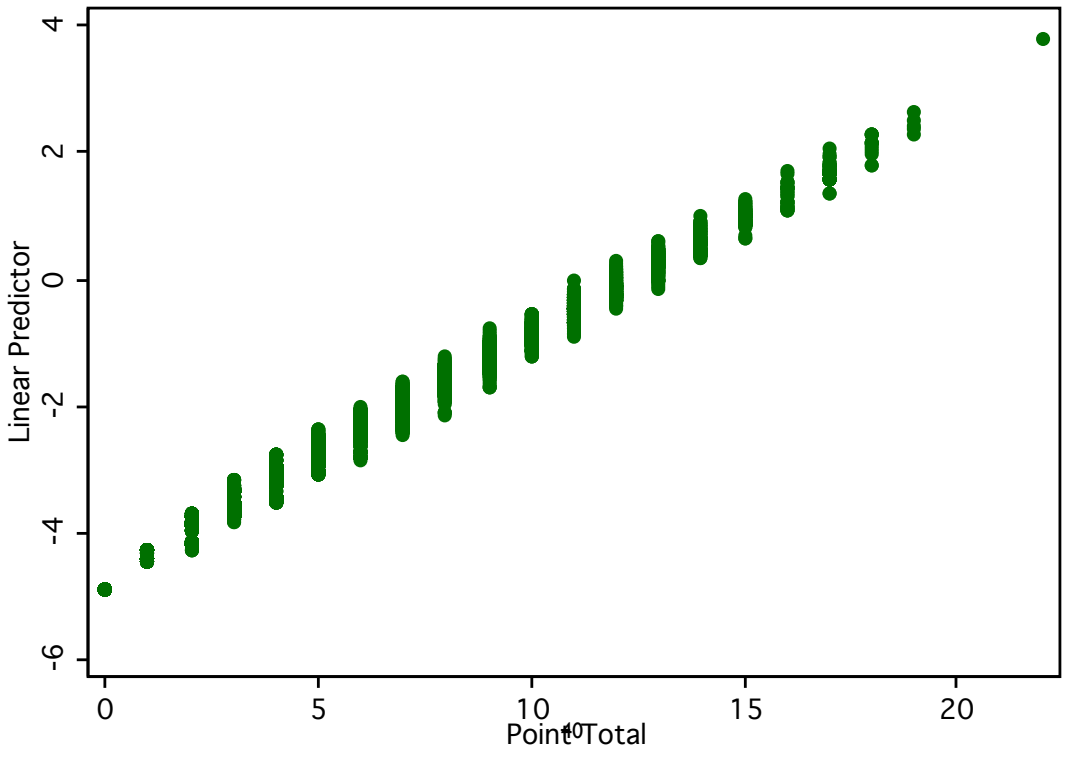
- Mimics the linear predictor (lp)
  if model is right, lp summarizes risk
- Simple score requires categorical predictor
- Recode them so 1=high risk, 0 = low
- Score is sum of points for each predictor
- Score for jth predictor: round($\beta_j/\min(\beta)$)

Calculating Points

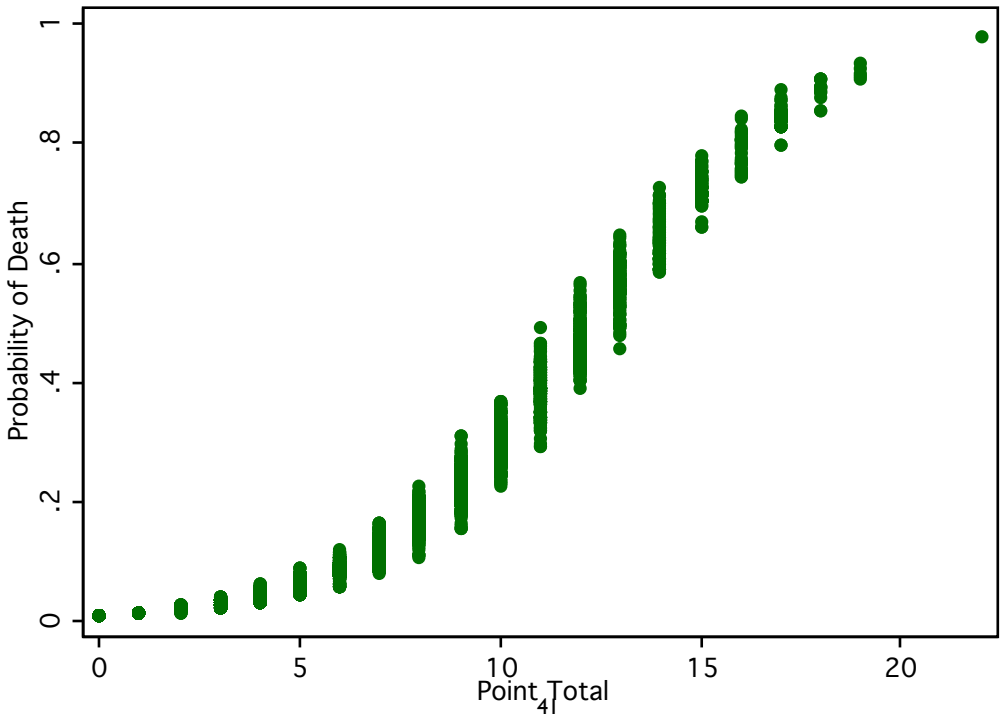
Push: HR = 1.53
coef = 0.42 (smallest coef)
points = +1 0.42/0.42

Walk/Jog HR = 2.1
coef = 0.736
points = +2 0.736/0.42 ≈ 2

Points v. Linear Predictor



Probability of Death



Next Week

- Discuss methods for assess a model
- Discuss model validation