Some thoughts on regression modeling for observational and longitudinal data

W. John Boscardin

October 16, 2018

Outline

- Choosing predictors for a regression model
- Missing data
- Repeated/longitudinal measurements

Recurring Theme

- "Essentially all models are wrong, but some are useful" (George E. P. Box)
- I would add that whichever wrong model you end up using in your analysis, there are many others that you could have used – your model is not particularly special and you should take care not to overinterpret it.

Building a regression model

- Harrell, Lee, Mark (1996)
- Sullivan, Massaro, D'Agostino (2004)
- Steyerberg et al. (2010)

Typical Setting for Model Building

- Long-term survival data on adults age 70+ ($n \approx 1000$, e.g.).
- Have maybe $P = 50$ baseline, admission, discharge characteristics potentially predicting survival
- Goal: build a reasonably parsimonious ($p = 10$ or $p = 15$ predictors), clinically practical and sensible model that has good discrimination and calibration
- Move from statistical model to something simple and clinically useful (make easy to calculate 5 year survival probabilities or median life expectancy for given individual)

Statistical models for risk prediction

- Logistic regression (or other binary regression)
- Cox regression (or other time-to-event models)
- Multinomial regression (for multi-state outcomes)

Predictions

- Key idea: don't just look at (odds/hazard) ratios for the predictors
- Instead focus on predicted probabilities from the fitted models
- For logistic regression get predicted probability of event for given characteristics
- For Cox regression, same but at specific time points

Choosing predictors

- Many possibilities all with pros and cons; combinable Frankenstein's monster-style
  - Theoretical guidance/DAG
  - Existing risk model or index (incremental value of your shiny new predictor)
  - Practicality/Simplicity/Cost of obtaining predictors
  - Bivariate screening
  - Forward/Backward/Stepwise selection
  - "Best" subset methods

Automatic subset selection

- Many sources have criticized stepwise model selection:
  - Standard errors of coefficients artificially small
  - Coefficient estimates biased away from zero
  - R^2 biased upward
  - Performs poorly in presence of multicollinearity
- Best subset selection usually viewed as even worse in all of these senses than stepwise
- Ronan Conroy: "I would no more let an automatic routine select my model than I would let some best-fit procedure pack my suitcase".

A Slightly Different View

- All of these things true (to some extent), but I think there is more important point
- Stepwise selection only shows one model and does not output comparisons to other potential models
- Best subsets regression gives a huge amount of useful information for comparing models, and in practice, a large number of models of reasonable parsimony are statistically nearly indistinguishable
- It is tremendously valuable to clinicians to view a lot of similarly performing prognostic models to choose ones that are most practically applied
- All the other criticisms can be addressed with bootstrapping

Best Subsets Selection

- Computationally infeasible to fit all 2^P possible subset models
- But for each of $p = 1, 2, 3, ..., P - 1$ it is blazingly fast (using both branch and bound and properties of score test) to find the best (or best k) models according to score statistic (similar to log-likelihood)
- This gives a list of $k(P - 1)$ models most of which are good in some sense
- Typical finding (Miao et al., 2014) is that dozens of models will all have same c-statistic and may be quite different in interpretation/simplicity/etc.

Best Subsets 1

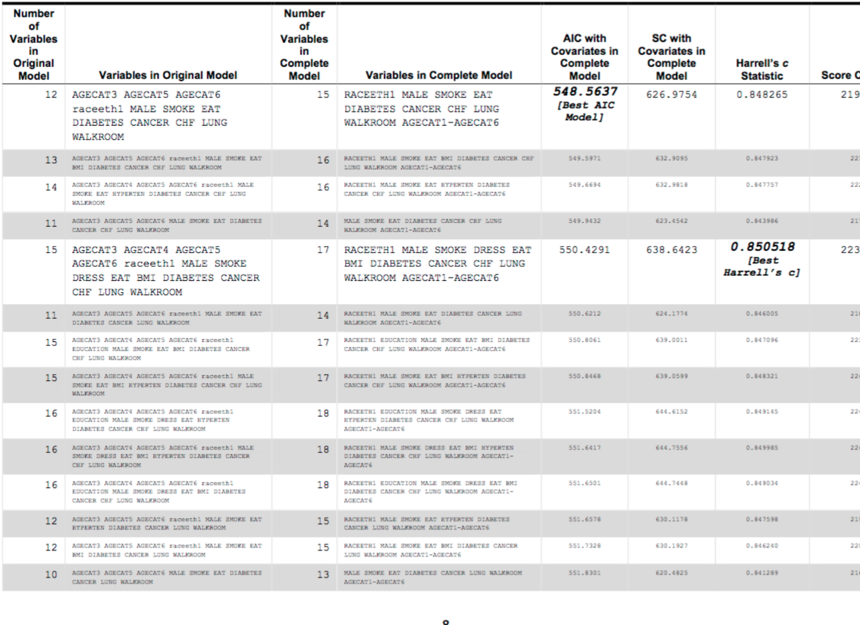
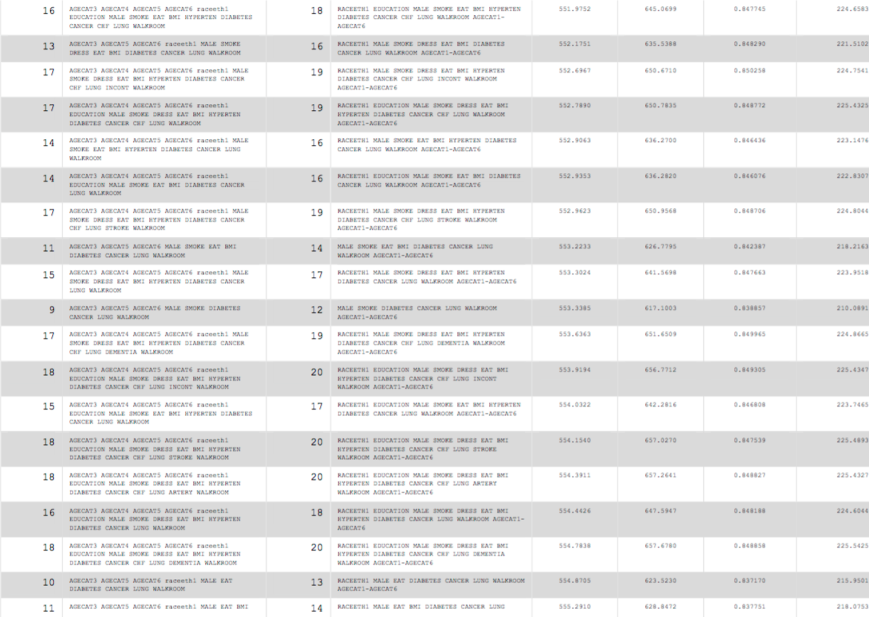


Table 1. Best Models Generated in the Original/Full Data Set by Best Subsets Procedure

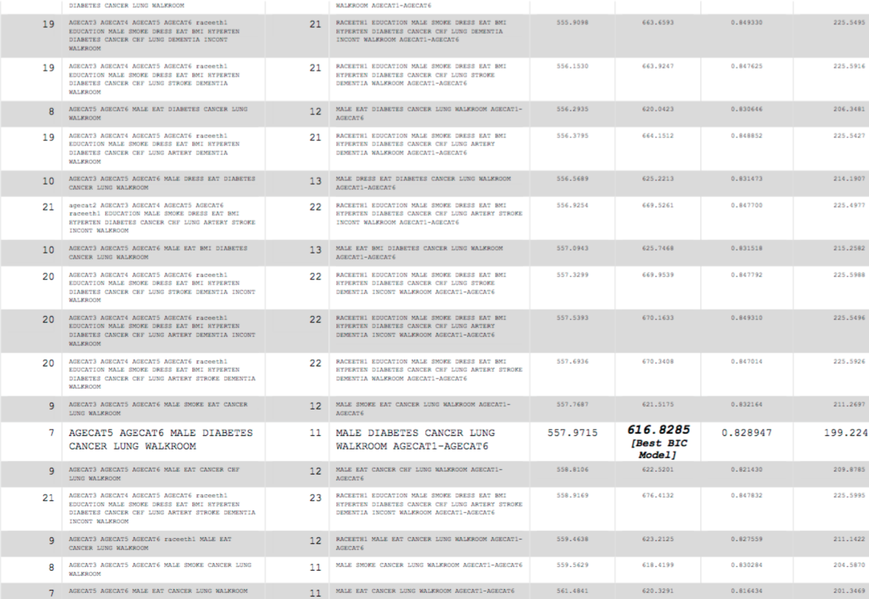
Number of Variables in Original Model	Variables in Original Model	Number of Variables in Complete Model	Variables in Complete Model	AIC with Covariates in Complete Model	SC with Covariates in Complete Model	Harrell's c Statistic	Score C
12	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM	15	RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	***548.5637*** ***[Best AIC Model]***	626.9754	0.848265	219
13	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM	16	RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	549.5971	632.9095	0.847923	22
14	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	16	RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	549.6694	632.9818	0.847757	22
11	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM	14	MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	549.9432	623.4542	0.843986	21
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM	17	RACEETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	550.4291	638.6423	***0.850518*** ***[Best Harrell's c]***	223
11	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM	14	RACEETH1 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	550.6212	624.1774	0.846005	21
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM	17	RACEETH1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	550.8061	639.0011	0.847096	22
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	17	RACEETH1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	550.8468	639.0599	0.848321	22
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 EDUCATION MALE SMOKE DRESS EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.5204	644.6152	0.849145	22
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.6417	644.7556	0.849985	22
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.6501	644.7448	0.849034	22
12	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM	15	RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	551.6578	630.1178	0.847598	21
12	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM	15	RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	551.7328	630.1927	0.846240	22
10	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM	13	MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	551.8301	620.4825	0.841289	21

Best Subsets 2



16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	18	RACEETH1 EDUCATION MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	551.9752	645.0699	0.847745	224.6583
13	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI DIABETES CANCER LUNG WALKROOM	16	RACEETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	552.1751	635.5388	0.848290	221.5102
17	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG INCONT WALKROOM	19	RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG INCONT WALKROOM AGECAT1-AGECAT6	552.6967	650.6710	0.850258	224.7541
17	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM	19	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	552.7890	650.7835	0.848772	225.4325
14	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM	16	RACEETH1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	552.9063	636.2700	0.846436	223.1476
14	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM	16	RACEETH1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	552.9353	636.2820	0.846076	222.8307
17	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE WALKROOM	19	RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE WALKROOM AGECAT1-AGECAT6	552.9623	650.9568	0.848706	224.8044
11	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM	14	MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	553.2233	626.7795	0.842387	218.2163
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM	17	RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	553.3024	641.5698	0.847663	223.9518
9	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE DIABETES CANCER LUNG WALKROOM	12	MALE SMOKE DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	553.3385	617.1003	0.838857	210.0891
17	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG DEMENTIA WALKROOM	19	RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG DEMENTIA WALKROOM AGECAT1-AGECAT6	553.6363	651.6509	0.849965	224.8665
18	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG INCONT WALKROOM	20	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG INCONT WALKROOM AGECAT1-AGECAT6	553.9194	656.7712	0.849305	225.4347
15	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM	17	RACEETH1 EDUCATION MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	554.0322	642.2816	0.846808	223.7465
18	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE WALKROOM	20	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE WALKROOM AGECAT1-AGECAT6	554.1540	657.0270	0.847539	225.4893
18	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY WALKROOM	20	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY WALKROOM AGECAT1-AGECAT6	554.3911	657.2641	0.848827	225.4327
16	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM	18	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	554.4426	647.5947	0.848188	224.6044
18	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG DEMENTIA WALKROOM	20	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG DEMENTIA WALKROOM AGECAT1-AGECAT6	554.7838	657.6780	0.848858	225.5425
10	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE EAT DIABETES CANCER LUNG WALKROOM	13	RACEETH1 MALE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	554.8705	623.5230	0.837170	215.9501
11	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE EAT BMI	14	RACEETH1 MALE EAT BMI DIABETES CANCER LUNG	555.2910	628.8472	0.837751	218.0753

Best Subsets 3



	DIABETES CANCER LUNG WALKROOM		WALKROOM AGECAT1-AGECAT6				
19	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM	21	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	555.9098	663.6593	0.849330	225.5495
19	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE DEMENTIA WALKROOM	21	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE DEMENTIA WALKROOM AGECAT1-AGECAT6	556.1530	663.9247	0.847625	225.5916
8	AGECAT5 AGECAT6 MALE EAT DIABETES CANCER LUNG WALKROOM	12	MALE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	556.2935	620.0423	0.830646	206.3481
19	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY DEMENTIA WALKROOM	21	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY DEMENTIA WALKROOM AGECAT1-AGECAT6	556.3795	664.1512	0.848852	225.5427
10	AGECAT3 AGECAT5 AGECAT6 MALE DRESS EAT DIABETES CANCER LUNG WALKROOM	13	MALE DRESS EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	556.5689	625.2213	0.831473	214.1907
21	agecat2 AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY STROKE INCONT WALKROOM	22	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY STROKE INCONT WALKROOM AGECAT1-AGECAT6	556.9254	669.5261	0.847700	225.4977
10	AGECAT3 AGECAT5 AGECAT6 MALE EAT BMI DIABETES CANCER LUNG WALKROOM	13	MALE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	557.0943	625.7468	0.831518	215.2582
20	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE DEMENTIA INCONT WALKROOM	22	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG STROKE DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	557.3299	669.9539	0.847792	225.5988
20	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY DEMENTIA INCONT WALKROOM	22	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	557.5393	670.1633	0.849310	225.5496
20	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA WALKROOM	22	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA WALKROOM AGECAT1-AGECAT6	557.6936	670.3408	0.847014	225.5926
9	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT CANCER LUNG WALKROOM	12	MALE SMOKE EAT CANCER LUNG WALKROOM AGECAT1-AGECAT6	557.7687	621.5175	0.832164	211.2697
7	AGECAT5 AGECAT6 MALE DIABETES CANCER LUNG WALKROOM	11	MALE DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6	557.9715	***616.8285 [Best BIC Model]***	0.828947	199.224
9	AGECAT3 AGECAT5 AGECAT6 MALE EAT CANCER CHF LUNG WALKROOM	12	MALE EAT CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	558.8106	622.5201	0.821430	209.8785
21	AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA INCONT WALKROOM	23	RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	558.9169	676.4132	0.847832	225.5995
9	AGECAT3 AGECAT5 AGECAT6 raceeth1 MALE EAT CANCER LUNG WALKROOM	12	RACEETH1 MALE EAT CANCER LUNG WALKROOM AGECAT1-AGECAT6	559.4638	623.2125	0.827559	211.1422
8	AGECAT3 AGECAT5 AGECAT6 MALE SMOKE CANCER LUNG WALKROOM	11	MALE SMOKE CANCER LUNG WALKROOM AGECAT1-AGECAT6	559.5629	618.4199	0.830284	204.5870
7	AGECAT5 AGECAT6 MALE EAT CANCER LUNG WALKROOM	11	MALE EAT CANCER LUNG WALKROOM AGECAT1-AGECAT6	561.4841	620.3291	0.816434	201.3469

Assessing fit: discrimination

- Discrimination demonstrations:
  - C-statistic for binary regression
  - Harrell's c-statistic for time-to-event models
  - Ad hoc variants on c-statistic for multi-state outcome models (e.g. collapse outcomes into dichotomous versions and use binary regression methods)
  - Graphically/Tabularly want to show large differences in predicted outcomes

Harrell's c statistic

- Generalization of logistic regression c-statistic to Cox model
- Denominator is all pairs of "evaluable" subjects where one known to have had the event before the other (so two cases, either both had event or one with event and other censored after that time)
- Numerator is number of denominator pairs where earlier event time has shorter predicted survival ("concordant")
- Harrell's c is proportion of evaluable pairs which are concordant in predicted and actual survival
- Does not work well with heavy censoring; other alternatives (e.g. Gönen and Heller, 2005)

Incremental value

- Typically improvement in c-statistic very small with your shiny new predictor (e.g. c=0.83 new model vs. c=0.82 old model)
- Sometimes reclassification indices provide better insight into incremental value (NRI and IDI)

Assessing fit: calibration

- Calibration demonstrations:
  - Binary regression: show that predicted event rates and observed event rates match up (graphically use calibration plot, numerically use Hosmer-Lemeshow test maybe)
  - Time-to-event models: look at fixed time points (e.g. 5 year survival) and use binary regression methods
  - Multi-state models: show that predicted event rates and observed event rates match up

Calibration plot (Steyerberg, 2010)

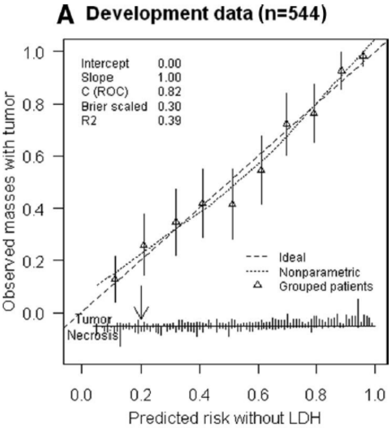


Table Format (Mehta, 2011)

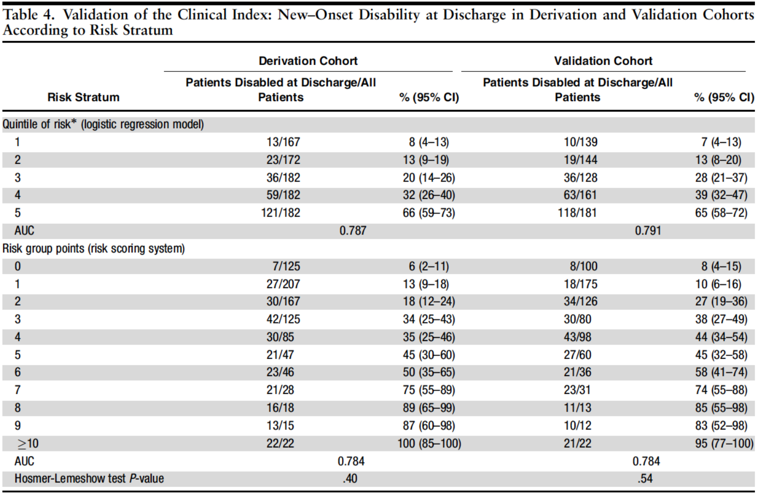
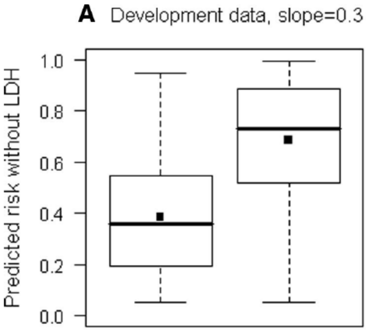


Table 4. Validation of the Clinical Index: New–Onset Disability at Discharge in Derivation and Validation Cohorts According to Risk Stratum

	Derivation Cohort		Validation Cohort	
Risk Stratum	**Patients Disabled at Discharge/All Patients**	**% (95% CI)**	**Patients Disabled at Discharge/All Patients**	**% (95% CI)**
Quintile of risk* (logistic regression model)				
1	13/167	8 (4–13)	10/139	7 (4–13)
2	23/172	13 (9–19)	19/144	13 (8–20)
3	36/182	20 (14–26)	36/128	28 (21–37)
4	59/182	32 (26–40)	63/161	39 (32–47)
5	121/182	66 (59–73)	118/181	65 (58–72)
AUC	0.787		0.791	
Risk group points (risk scoring system)				
0	7/125	6 (2–11)	8/100	8 (4–15)
1	27/207	13 (9–18)	18/175	10 (6–16)
2	30/167	18 (12–24)	34/126	27 (19–36)
3	42/125	34 (25–43)	30/80	38 (27–49)
4	30/85	35 (25–46)	43/98	44 (34–54)
5	21/47	45 (30–60)	27/60	45 (32–58)
6	23/46	50 (35–65)	21/36	58 (41–74)
7	21/28	75 (55–89)	23/31	74 (55–88)
8	16/18	89 (65–99)	11/13	85 (55–98)
9	13/15	87 (60–98)	10/12	83 (52–98)
$\geq$10	22/22	100 (85–100)	21/22	95 (77–100)
AUC	0.784		0.784	
Hosmer-Lemeshow test *P*-value	.40		.54	

Discrimination Plot (Steyerberg, 2010)



Simplex discrimination (Barnes, 2013)

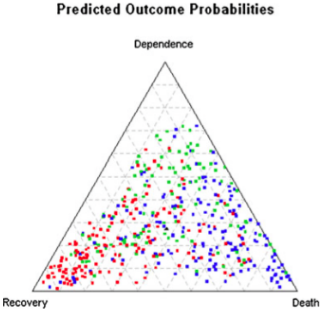


Figure 2. Predicted probabilities of recovery, dependence and death in all subjects combined. Figure 2 is created by stacking Figure 1a–c on top of each other, with colors used to reflect the actual outcomes for each subject (*red*=recovery, *green*=dependence, *blue*=death). The *black triangle* in the center reflects the marginal predicted probabilities for the three outcomes (36 % recovery, 27 % dependence, 37 % death).

Validation

- Internal validation (in same data set): can either do random or purposeful split sample.
- I am not a fan of single random split sample (nor is Harrell, 1996) which attempts to measure overfitting but confounds it with statistical variability
- Multiple random split sample preferable (e.g. cross-validation or bootstrapping)
- Internal purposeful split sample gets at validity of model in subgroups (demographic, geographical, temporal, etc.)
- External validation carries more weight

Overfitting

- “Over-optimism” has two components
- 1. whatever procedure was used to select a good model was almost certainly driven by data at hand
- 2. the coefficients for that model are optimized to provide the best fit to the data at hand
- When assessing the model performance in a new data set, we will almost always have degradation in the model performance measure
- With a single split sample, you can’t separate random variability from systematic overfitting
- Can address with repeated split sampling (e.g. cross-validation or bootstrapping)

Implementation

- Crucial aspect of prognostic modeling: turn statistical model into something simple, clinically useful
- Can do point scoring or keep predictions from actual model
- Point scoring super useful pre-computer age (add up points; look up score)
- Original model no big deal now with tablet apps (tap on risk factors; get predictions instantly)
- That said, point score models still very popular

Interesting Directions (with Sei Lee and Alex Smith)

- Many, many models have nearly identical c-statistics
- Have interface asking for risk factors to be input
- Fill in as many as have – give prediction using appropriate model for the non-missing ones (if acceptable level of discrimination)
- Also looking into model fit indices taking into account time-collection cost

Handling Missing Data

- White, Royston, Wood (2011)
- Royston (2004)
- Little and Rubin (2002)

Common Setting for Missing Data

- Have a large number of potential predictors in a regression analysis
- Regression software will drop cases that have any missing data
- Many of the predictors are missing in the data set
- Even if small percentage of missing for any particular predictor might have only a handful of subjects with no missing data for any predictors

Imputation of Missing Data

- Various procedures to fill in the missing data so that these subjects are not dropped from the analysis have been used for past 40 years
- Suppose want to regress blood pressure on weight, height, gender, etc, but that some weights are missing in the data set
- One idea is to fill in the mean weight for all missing weights
- Slightly better idea is to fill in the mean weight for all those with same height and gender

Multiple Imputation

- Problem is that the fill-in is uncertain.
- Key idea: fill in a random draw from the set of all weights of those with same height and gender and do so a number (M) of times. This is called Multiple Imputation (MI)
- Can then do the regression in each of these M complete data sets
- Combine the M sets of regression coefficients using Rubin's rules (Little and Rubin, 2002; Carlin et al., 2008)

Rubin's Rules

- Estimate of a parameter is the average of the parameter estimates from each imputed data set
- Standard error of a parameter combines the within imputation SE and between imputation SD

MI in software and in practice

- Twenty years ago, MI was a nice idea in theory
- Now MI is easily available in SAS, Stata, and R, for example
- First widely available algorithm was NORM (Schafer, 1997). Assumes multivariate normal distribution for all quantities of interest. Variants allow some relaxation of this. Backbone of Proc MI in SAS
- Specially modified version of NORM used to do the official multiple imputations for NHANES III (Schafer et al., 1996); other government data have official MIs (e.g. Schenker et al., 2006)

NHANES III imputation

(from the official documentation of NHANES III-MI "NH3MI.DOC")

```
One key feature of the imputation models is that they are based
upon an assumption of multivariate nomality; that is, they assume
that the variables to be imputed are (individually and jointly)
normally distributed within demographic subgroups defined by age,
sex, and race/ethnicity.  Some variables that consist of discrete
categories (e.g.  self-reported health status, which takes values
from 1 = excellent to 5 = poor) were modeled and imputed as if they
were normally distributed, and the continuous imputed values were
rounded off to the nearest category.  Other variables whose
distributions were skewed were transformed by standard power
functions such as the logarithm, square root, or reciprocal square
root; modeling and imputation were carried out on the transformed
data, and after imputation they were transformed back to the
original scale....
```

Simple example in NHANES

```
. mi set mlong
. mi register imputed kstones sbp dbp male age smoke maxwt
. mi misstable patterns, frequency
      Missing-value patterns
        (1 means complete)
              |   Pattern
    Frequency |  1  2  3  4    5
  ------------+------------------
       17,751 |  1  1  1  1    1
              |
          954 |  1  1  1  0    0
          618 |  1  1  0  1    1
          464 |  1  0  1  1    1
           62 |  1  0  1  0    0
           57 |  1  1  0  0    0
           42 |  0  1  0  1    1
           29 |  1  0  0  1    1
           22 |  0  1  0  0    0
```

Complete cases regression

Lose almost 12% of the data set.

```
logistic kstones sbp dbp male age smoke maxwt

Logistic regression                               Number of obs   =      17751
                                                  LR chi2(6)      =     346.65
                                                  Prob > chi2     =     0.0000
Log likelihood = -3170.0128                       Pseudo R2       =     0.0518

------------------------------------------------------------------------------
     kstones | Odds Ratio   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |   .9998493   .0020883    -0.07   0.942     .9957646    1.003951
         dbp |   1.008443   .0031585     2.68   0.007     1.002272    1.014653
        male |   1.474334   .1137626     5.03   0.000     1.267405    1.715048
         age |   1.028355   .0023434    12.27   0.000     1.023773    1.032959
       smoke |   .8614465    .069716    -1.84   0.065     .7350916    1.009521
       maxwt |   1.005423   .0008747     6.22   0.000      1.00371    1.007139
------------------------------------------------------------------------------
```

Imputing data

```
mi impute chained (regress) sbp (regress) dbp (logit) male (regress)
age (logit) smoke (regress) maxwt, add(10)

note: variable male contains no soft missing (.) values; imputing nothing

Conditional models:
             smoke: logit smoke i.male maxwt age sbp dbp
             maxwt: regress maxwt i.male i.smoke age sbp dbp
               age: regress age i.male i.smoke maxwt sbp dbp
               sbp: regress sbp i.male i.smoke maxwt age dbp
               dbp: regress dbp i.male i.smoke maxwt age sbp

Performing chained iterations ...

Multivariate imputation                     Imputations =       10
Chained equations                                 added =       10
Imputed: m=1 through m=10                       updated =        0

Initialization: monotone                     Iterations =      100
                                                burn-in =       10
```

MI Logistic Regression (MC error)

```
 . mi estimate, mcerror: logistic kstones sbp dbp male age smoke maxwt
Logistic regression                                 Number of obs    =      20029
                                                    Largest FMI      =     0.1300
------------------------------------------------------------------------------
     kstones |      Coef.   Std. Err.      t    P>|t|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |   .0013643   .0020714     0.66   0.510    -.0027044     .005433
             |   .0002225   .0000516     0.10   0.066     .0002038    .0002842
             |
         dbp |   .0059833   .0030078     1.99   0.047     .0000822    .0118843
             |   .0002665   .0000477     0.08   0.009     .0002537    .0003112
             |
        male |   .4154024   .0730054     5.69   0.000     .2723141    .5584907
             |   .0012741   .0000642     0.02   0.000     .0012587    .0013017
             |
         age |   .0270061   .0022828    11.83   0.000     .0225226    .0314895
             |   .0002419   .0000342     0.20   0.000     .0002485    .0002558
             |
       smoke |  -.1498978   .0771103    -1.94   0.052     -.301032    .0012364
             |   .0018016   .0001164     0.02   0.003     .0017323    .0018961
             |
       maxwt |   .0050521   .0008346     6.05   0.000     .0034163    .0066878
             |   .0000302   2.72e-06     0.04   0.000     .0000302    .0000311
             |
       _cons |  -6.207185   .2778718   -22.34   0.000    -6.752047   -5.662324
             |   .0200949   .0046794     0.43   0.000     .0157049    .0271742
------------------------------------------------------------------------------
Note: values displayed beneath estimates are Monte Carlo error estimates.
```

MI Logistic Regression (look at FMIs)

```
. mi estimate, vartable nocitable

Multiple-imputation estimates                     Imputations     =         10
Logistic regression

Variance information
------------------------------------------------------------------------------
             |        Imputation variance                             Relative
             |    Within   Between     Total       RVI       FMI    efficiency
-------------+----------------------------------------------------------------
         sbp |   3.7e-06   5.0e-07   4.3e-06   .145351   .130013       .987166
         dbp |   8.3e-06   7.1e-07   9.0e-06   .094518   .087866        .99129
        male |   .005312   .000016    .00533   .003362   .003353       .999665
         age |   4.6e-06   5.9e-07   5.2e-06   .140889   .126446       .987513
       smoke |    .00591   .000032   .005946   .006041   .006012       .999399
       maxwt |   6.9e-07   9.1e-09   7.0e-07   .014592   .014428       .998559
       _cons |   .072771   .004038   .077213   .061039    .05822       .994212
------------------------------------------------------------------------------
```

Alphabet Soup

- Missing completely at random (MCAR): probability of missing does not depend on observed or on missing data (e.g. recording instrument fails 10% of the time)
- Missing at random (MAR): probability of missing depends only on observed data (e.g. men who smoke more likely to be missing blood pressure)
- Missing not at random (MNAR): missingness probability depends on missing values (e.g. when maximum weight was very high more likely not to report it)

Assumptions and Methods

- Dropping subjects with any missing data (listwise deletion) may lead to biased estimates of parameters (unless MCAR) and always leads to inefficient estimates
- Multiple imputation routines can give unbiased estimates of parameters of interest assuming data are MAR and will almost always be more efficient
- Practical advice: if include everything remotely relevant in chained equations then MAR is much more plausible (similar to propensity score idea)
- If MNAR, need to run more advanced models as sensitivity check (Daniels and Hogan, 2008)

And Now for Something Completely Different

- Recent problem reflecting real-world complexity!
- VA data on providers (approximately 10K providers)
- About 15% missing provider gender
- Several potential strategies:
  - Gender from name algorithms (e.g. `genderize.io`) with external databases
  - Same algorithms but in internal database
  - Impute using multiple imputation from provider specialty, patient mix demographics, etc
  - Assign some by expert judgement!
- Optimal combination of strategies is not trivial

Summary

- Switching regressions (SR) is incredibly intuitive and flexible method for generating multiple imputations
- SR is still on somewhat shaky theoretical ground statistically, but a number of recent papers (e.g. Lee and Carlin, 2010) have shown it works quite well
- SR is now seamlessly integrated into Stata (`mi impute chained`) as of Version 12
- Stata multiple imputation works with nearly every regression routine and also handles survey weights!

Repeated Measures and Longitudinal Data

- Repeated or longitudinal measurements on subjects very common in studies
- Can be very useful to address confounding but also adds complexity to modeling
- Will discuss some commonly used models

Mixed Effects Regression Models

- Make model for average trajectory in time
- Assume each patient has their own patient-specific trajectory centered around the average trajectory
- Many parametric choices for shape of trajectory (e.g. linear, polynomial, spline)
- Can also use non-parametric shapes via penalized or smoothing splines

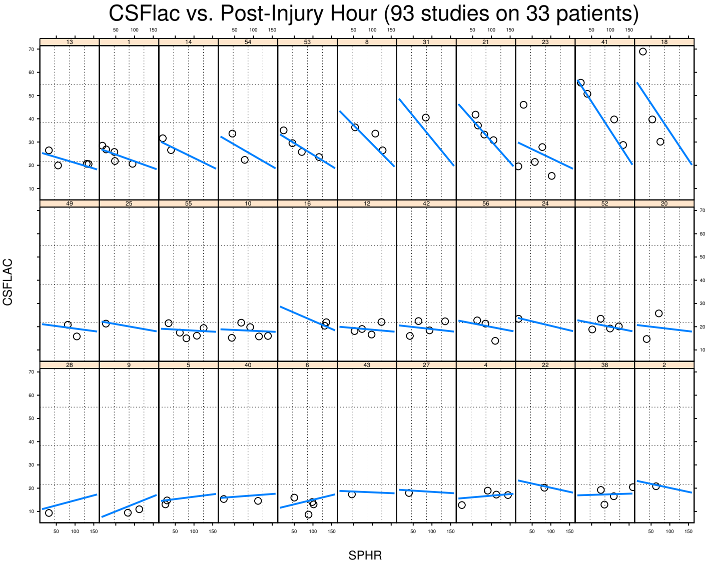
Average Shape Remarks

- Might focus on average trajectory. The mixed model used to properly account for intra-patient correlation of longitudinal data
- Buries under rug the likely inter-patient heterogeneity around these average trajectories
- Two components of heterogeneity: (i) noisiness of individual data; (ii) variation of shape in individual trajectories

Subject-Specific Trajectories

- Goal: estimate subject-specific linear trajectories.
- Can use these estimated trajectories for variety of purposes:
  - Direct statistically valid inference on subject-specific time trajectories
  - Classify subjects according to characteristics of subject-specific time trajectories (e.g. increasers vs. decreasers, slow vs. quick increasers)
  - Inference on time to threshold crossing
- More reliant on model being correct

Random Slopes and Intercepts Example

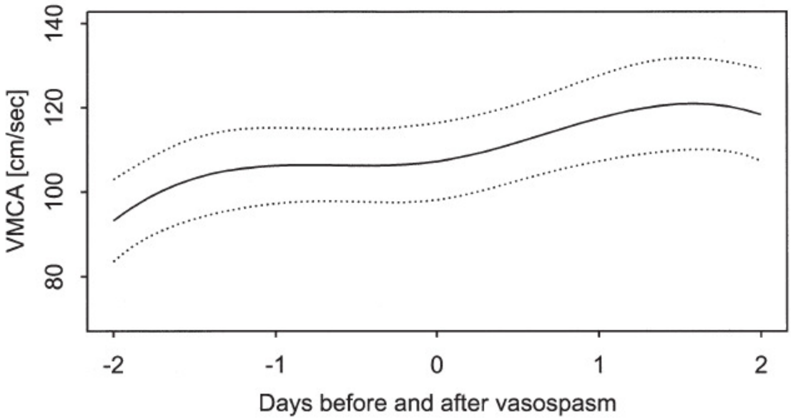


CSFlac vs. Post-Injury Hour (93 studies on 33 patients)

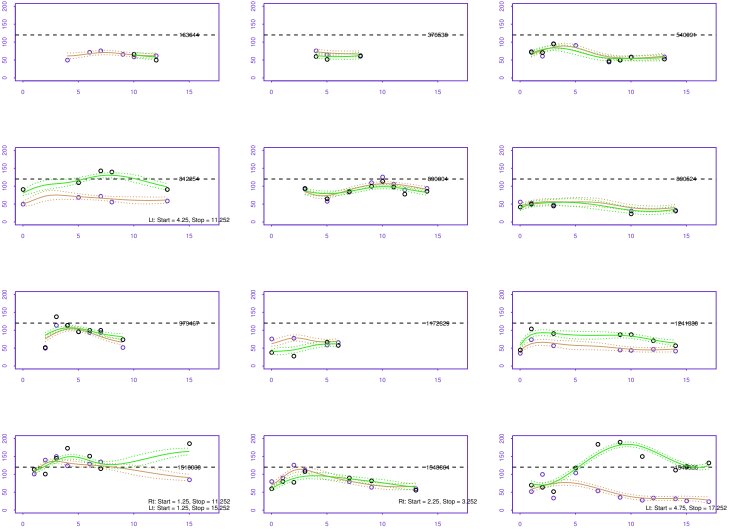
Cubic Splines Average Trajectory

5/25/2011 9:55:24 AM

Oertel et al. (2005). J Neurosurg 103:812–824.



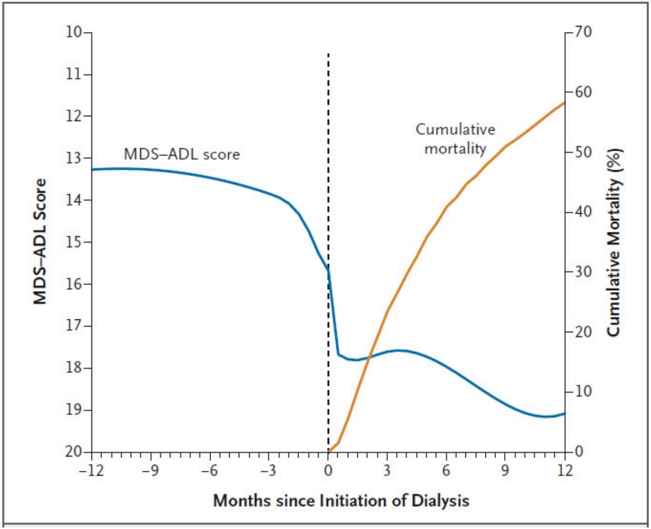
Cubic Splines Individual Trajectories



Cubic Spline Trajectory Example

5/24/2011 3:20:49 PM

Tamura et al. (2009). N Engl J Med;361:1539-47.

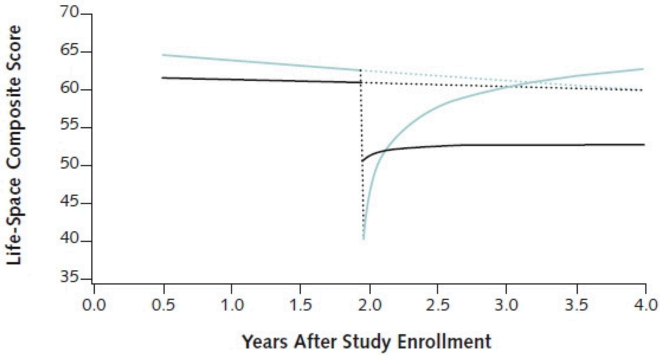


Logarithmic Recovery Example

5/24/2011 3:23:23 PM

Brown et al. (2009). Ann Intern Med;150:372-378.

Figure 2. Trajectory of adjusted life-space mobility decrease and recovery.



Discrete Timepoints

- Most longitudinal cohort studies have only a small number of timepoints that are common to all subjects (e.g. at baseline and at every 2 years after baseline)
- Rather than looking at baseline as time 0 can look at post-enrollment event-of-interest times (which are often obtained continuously) as time 0
- Example: physical functioning before and after hospitalization
- For an event that happens at roughly constant rate, timepoints will be roughly uniformly distributed

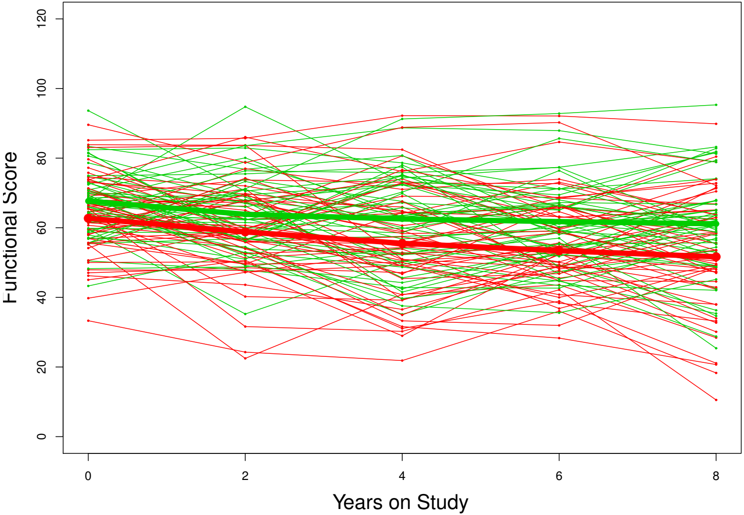
HRS Data Setting

- Nationally representative study of older Americans.
- Bigger sample size but fewer timepoints (less frequent) than Brown
- Total of 7000 subjects, 5000 hospitalized during study (and smaller subgroups are of interest)
- Have 5 measurement occasions per subject (every 2 years vs. every 6 months in Brown)
- How good of a job can we do estimating Brown model with 2-3 before and 2-3 after?
- How can we argue that we have enough power to look at questions of interest!

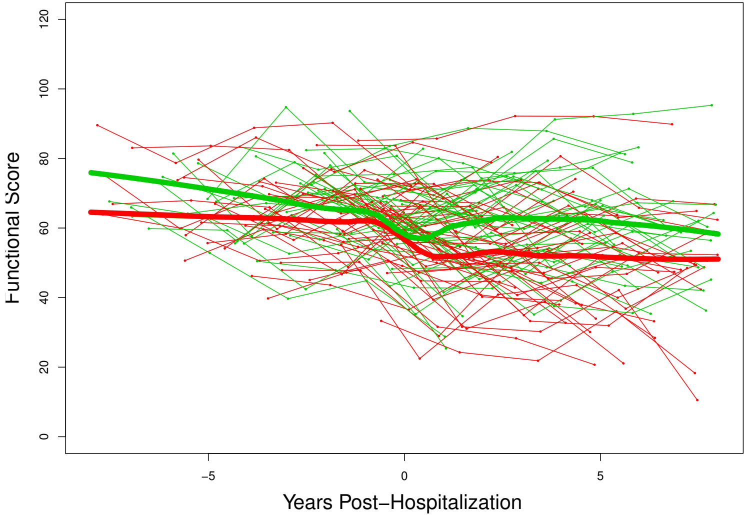
Simulated Data

- Use fitted model from Brown to simulate data
- Use sample size and time point frequency from my data source of interest

Discrete Timepoints of Study



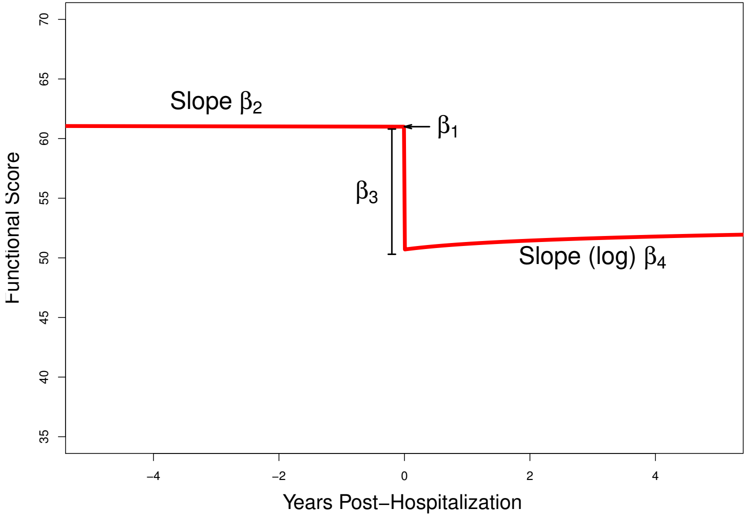
Hospitalization-Centered Timepoints



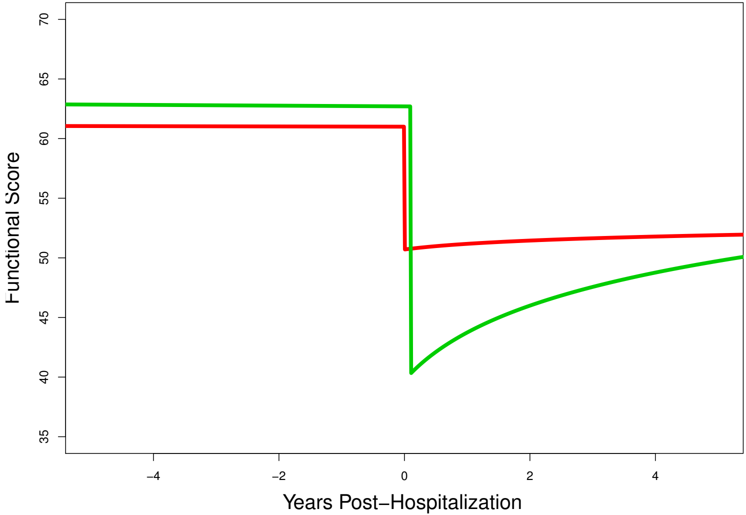
Four parameter model (Single group)

- t_{ij} = Time since hospitalization for subject i, occasion $j = 1, 2, 3, 4, 5$
- h_{ij} = Indicator of post-hospitalization time, i.e. $h_{ij} \equiv 1_{t_{ij}>0}$
- r_{ij} = Logarithmic time post-hospitalization, i.e. $r_{ij} \equiv \log(h_{ij} t_{ij} + 1)$
- β_1 = Intercept (average score at $t = 0$)
- β_2 = Pre-hospitalization slope
- β_3 = Amount score drops at time of hospitalization
- β_4 = Recovery slope on logarithmic time scale

Logarithmic recovery model



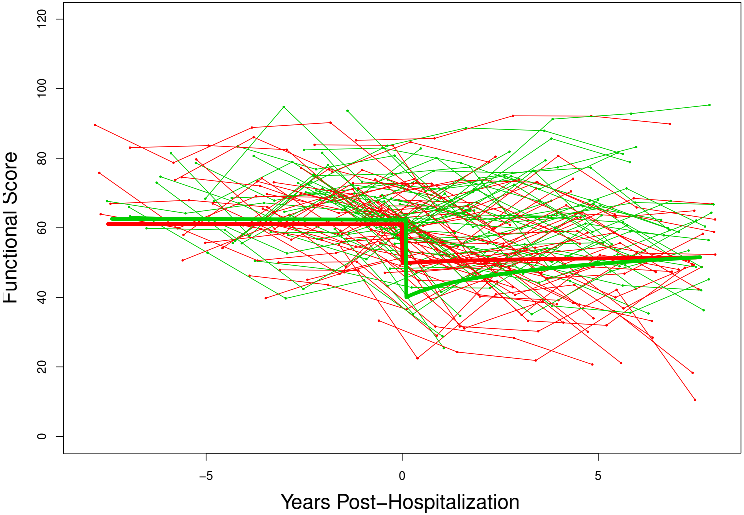
Two group logarithmic recovery model



A few more modeling details

- Random effects for intercept and drop at time of hospitalization in what I will discuss here
- Have binary covariate indicating type of hospitalization (surgical vs. non-surgical) distributed roughly 50/50
- Add interactions of intercept, slope, drop, recovery with the covariate

Fitted Model



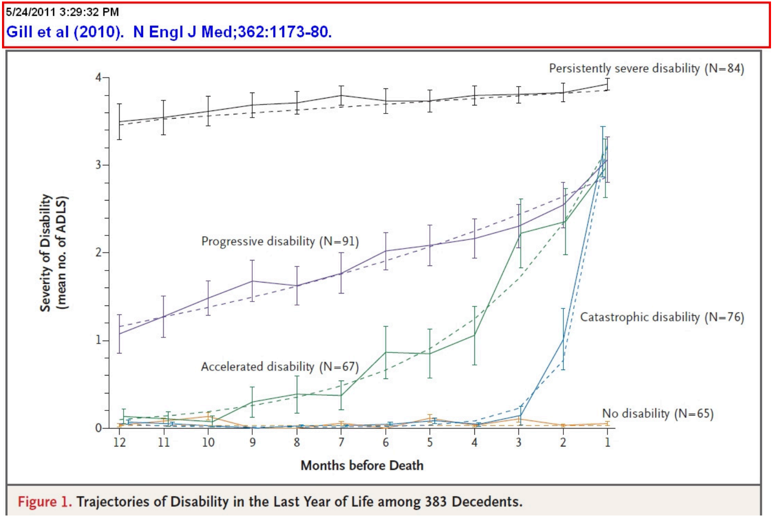
General remarks

- A sample size of 5000 subjects with a 50/50 split of the hospitalization-type covariate is enough to get extremely precise estimates of the parameters in the Brown et al. model
- Acceptable precision (i.e. enough power to discern clinically meaningful differences)

Latent Trajectory Models

- Assume there are a small number (k) of discrete categories of patients
- Category is unknown (latent) for each patient
- Simultaneously estimate the latent class memberships (get a probability for each patient belonging to each class) and the k trajectories
- Has become very popular method in past several years largely due to SAS Traj procedure (Jones et al., 2001) and recent NEJM article (Gill et al., 2010)

Latent Trajectory Example 1



5/24/2011 3:29:32 PM

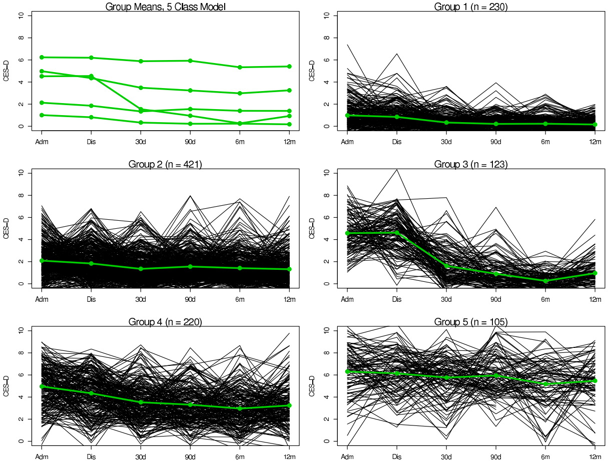
Gill et al (2010). N Engl J Med;362:1173-80.

Figure 1. Trajectories of Disability in the Last Year of Life among 383 Decedents.

Latent Trajectory Discussion

- Visually can be powerful way to display heterogeneity in data
- However, estimated latent trajectories may not be clinically distinct
- Statistically “optimal” solution masks near-optimality of many quite different solutions
- Look at Bandeen-Roche plots (next slide) for visual check on model fit

Latent Trajectory Example 1



Latent Trajectory vs. Mixed Effects

- As mentioned can also use mixed effects to categorize patients into small number of groups based on the subject-specific trajectories
- This is fairly common application in practice
- Ignores uncertainty in group membership, however
- Latent class makes this explicit by reporting probabilities of group membership
- Can get group membership probabilities for mixed effects approach using Bayesian inference (now available in procs mixed/mcmc)

References

- Barnes DE, Mehta KM, Boscardin WJ, Fortinsky RH, Palmer RM, Kirby KA, Landefeld CS (2013). A prognostic index to predict recovery, dependence, or death in elders who become disabled during hospitalization. *J Gen Intern Med*, **28**:261-268.
- Harrell FE, Lee KL, Mark DB (1996). Tutorial in Biostatistics: Multivariable prognostic models. *Stat Med*, **15**, 361–387.
- King (2003). Running a best-subsets logistic regression: an alternative to stepwise methods. *Educ Psych Meas*, **63**, 392–403.
- Mehta KM, Pierluissi E, Boscardin WJ, Kirby K, Walter L, Chren M, Palmer R, Counsell S, Landefeld CS (2011). A clinical index to stratify hospitalized older patients according to risk for new-onset disability. *J Am Geriatr Soc*, **59**:1206-1216.
- Miao Y, Cenzer I, Kirby K, Boscardin WJ. (2013) Estimating Harrell?s Optimism on Predictive Indices Using Bootstrap Samples *Proc SAS Global Forum*, **2013:504**.
- Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidem*, **21**, 128–138.
- Sullivan LM, Massaro JM, D'Agostino RB (2004), Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med*, **23**,1631?1660.

References (2)

- Carlin JB, Galati JC, Royston P (2008). A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* **8**, 49–67.
- Daniels MJ, Hogan JW (2008). *Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis*. New York: CRC.
- Jones B, Nagin D, Roeder K (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol Method Res*, **29**, 374–393.
- Lee KJ and Carlin JB (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am J of Epid*, **171**, 624–632.
- Li, K.-H. 1988. Imputation using Markov chains. *J Stat Comp Simulation*.
- Little RJA and Rubin DB (2002). *Statistical analysis with missing data, 2nd ed.*. New York: Wiley.
- Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P (2001). A sequential regression imputation for survey data. *Survey Methodology*, **27**, 85–96.
- Royston P (2004). Multiple imputation of missing values: update. *Stata Journal*, **5**, 188–201.

References (3)

- Rubin, DB (1976). Inference and Missing Data. *Biometrika*, **63**, 581–592.
- Schafer, JL (1997). *Analysis of incomplete multivariate data*. New York: CRC.
- Schafer JL, Ezzatti-Rice TM, Johnson W, Khare M, Little RJA, Rubin, DB (1996). The NHANES III multiple imputation project. *ASA Proc of Survey Research Methods Section*, 28–37.
- Schenker N, Raghunathan TE, Chiu PL, Makuc DM, Zhang GY, Cohen AJ (2006). Multiple imputation of missing income data in the National Health Interview Survey. *J Am Stat Assoc*, **101**, 924–933.
- van Buuren S, Boshuizen HC, Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*, **18**, 681–694.
- van Buuren S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB (2006). Fully conditional specification in multivariate imputation. *J Stat Comp Sim*, **76**, 1049–1064.
- van Buuren S (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Meth Med Res*, **16**, 219–242.
- White IR, Royston P, Wood AM (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, **30**, 377-99.