Some thoughts on regression modeling for observational and longitudinal data

W. John Boscardin

October 16, 2018
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

- 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
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 $p = 1, 2, 3, \ldots, P - 1$ 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.
<table>
<thead>
<tr>
<th>Number of Variables in Original Model</th>
<th>Variables in Original Model</th>
<th>Number of Variables in Complete Model</th>
<th>Variables in Complete Model</th>
<th>AIC with Covariates in Complete Model</th>
<th>SC with Covariates in Complete Model</th>
<th>Harrell’s c Statistic</th>
<th>Score C</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>AGECAT3 AGECAT6 AGECAT7 RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM</td>
<td>15</td>
<td>RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>548.5637 [Best AIC Model]</td>
<td>626.9754</td>
<td>0.848265</td>
<td>219</td>
</tr>
<tr>
<td>13</td>
<td>AGECAT3 AGECAT6 AGECAT7 RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM</td>
<td>16</td>
<td>RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>549.2971</td>
<td>632.9805</td>
<td>0.847923</td>
<td>221</td>
</tr>
<tr>
<td>14</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT7 RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM</td>
<td>16</td>
<td>RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>549.6494</td>
<td>632.8818</td>
<td>0.847757</td>
<td>222</td>
</tr>
<tr>
<td>11</td>
<td>AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM</td>
<td>14</td>
<td>MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>549.9632</td>
<td>623.4562</td>
<td>0.843826</td>
<td>217</td>
</tr>
<tr>
<td>15</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 RACEETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM</td>
<td>17</td>
<td>RACEETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.4291</td>
<td>638.6423</td>
<td>0.850518 [Best Harrell’s c]</td>
<td>223</td>
</tr>
<tr>
<td>11</td>
<td>AGECAT3 AGECAT7 AGECAT6 RACEETH1 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM</td>
<td>14</td>
<td>RACEETH1 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.68212</td>
<td>624.1977</td>
<td>0.846005</td>
<td>216</td>
</tr>
<tr>
<td>15</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 RACEETH1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM</td>
<td>17</td>
<td>RACEETH1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.8041</td>
<td>639.0011</td>
<td>0.847096</td>
<td>227</td>
</tr>
<tr>
<td>15</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 RACEETH1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM</td>
<td>17</td>
<td>RACEETH1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.8486</td>
<td>639.0199</td>
<td>0.848321</td>
<td>224</td>
</tr>
<tr>
<td>16</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM</td>
<td>18</td>
<td>RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.5204</td>
<td>644.4152</td>
<td>0.849145</td>
<td>226</td>
</tr>
<tr>
<td>16</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM</td>
<td>18</td>
<td>RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.6417</td>
<td>644.7456</td>
<td>0.849985</td>
<td>227</td>
</tr>
<tr>
<td>16</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM</td>
<td>18</td>
<td>RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.6501</td>
<td>644.7448</td>
<td>0.849034</td>
<td>224</td>
</tr>
<tr>
<td>12</td>
<td>AGECAT3 AGECAT5 AGECAT6 RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM</td>
<td>15</td>
<td>RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.6579</td>
<td>639.1170</td>
<td>0.847906</td>
<td>216</td>
</tr>
<tr>
<td>12</td>
<td>AGECAT3 AGECAT5 AGECAT6 RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM</td>
<td>15</td>
<td>RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.7326</td>
<td>639.1927</td>
<td>0.846240</td>
<td>220</td>
</tr>
<tr>
<td>10</td>
<td>AGECAT3 AGECAT5 AGECAT6 MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM</td>
<td>13</td>
<td>MALE SMOKE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.8301</td>
<td>620.3682</td>
<td>0.841289</td>
<td>214</td>
</tr>
<tr>
<td>Column 1</td>
<td>Column 2</td>
<td>Column 3</td>
<td>Column 4</td>
<td>Column 5</td>
<td>Column 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>18</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>551.9752</td>
<td>645.0499</td>
<td>0.863745</td>
<td>224.6528</td>
</tr>
<tr>
<td>13</td>
<td>Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>16</td>
<td>Racketni Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>552.1751</td>
<td>635.5388</td>
<td>0.866290</td>
<td>221.5102</td>
</tr>
<tr>
<td>17</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>19</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Intox Walkroom</td>
<td>552.6967</td>
<td>650.6710</td>
<td>0.850238</td>
<td>224.7414</td>
</tr>
<tr>
<td>17</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>19</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Intox Walkroom</td>
<td>552.7890</td>
<td>650.7835</td>
<td>0.867172</td>
<td>225.6325</td>
</tr>
<tr>
<td>14</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>16</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>552.9043</td>
<td>636.2700</td>
<td>0.866436</td>
<td>223.1476</td>
</tr>
<tr>
<td>14</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>16</td>
<td>Racketni Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>552.9353</td>
<td>636.2820</td>
<td>0.860706</td>
<td>222.8307</td>
</tr>
<tr>
<td>17</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>19</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Stroke Walkroom</td>
<td>552.9623</td>
<td>650.9560</td>
<td>0.867064</td>
<td>224.8044</td>
</tr>
<tr>
<td>11</td>
<td>Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>14</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>553.2233</td>
<td>626.7795</td>
<td>0.862387</td>
<td>218.2163</td>
</tr>
<tr>
<td>15</td>
<td>Education Male Smoke Eat BMI Diabetes Cancer Lung Walkroom</td>
<td>17</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Stroke Walkroom</td>
<td>553.3024</td>
<td>641.5698</td>
<td>0.867643</td>
<td>223.9518</td>
</tr>
<tr>
<td>9</td>
<td>Education Male Smoke Diabetes Cancer Lung Walkroom</td>
<td>12</td>
<td>Male Smoke Diabetes Cancer Lung Walkroom</td>
<td>553.3385</td>
<td>617.1003</td>
<td>0.836657</td>
<td>210.0091</td>
</tr>
<tr>
<td>17</td>
<td>Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>19</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Dementia Walkroom</td>
<td>553.6363</td>
<td>651.6509</td>
<td>0.869945</td>
<td>224.8665</td>
</tr>
<tr>
<td>18</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>20</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Intox Walkroom</td>
<td>553.9194</td>
<td>656.7712</td>
<td>0.869305</td>
<td>225.6347</td>
</tr>
<tr>
<td>15</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>17</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>554.0322</td>
<td>642.2816</td>
<td>0.868008</td>
<td>223.7465</td>
</tr>
<tr>
<td>18</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>20</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Stroke Walkroom</td>
<td>554.1540</td>
<td>657.0270</td>
<td>0.867039</td>
<td>225.6893</td>
</tr>
<tr>
<td>18</td>
<td>Racketni Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>20</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Stroke Walkroom</td>
<td>554.3911</td>
<td>657.2641</td>
<td>0.868827</td>
<td>225.8327</td>
</tr>
<tr>
<td>16</td>
<td>Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>18</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>554.4426</td>
<td>647.5947</td>
<td>0.868188</td>
<td>224.6044</td>
</tr>
<tr>
<td>18</td>
<td>Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>20</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Dementia Walkroom</td>
<td>554.7838</td>
<td>657.4780</td>
<td>0.868858</td>
<td>225.5425</td>
</tr>
<tr>
<td>18</td>
<td>Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>20</td>
<td>Racketni Education Male Smoke Eat BMI Hypertension Diabetes Cancer Lung Walkroom</td>
<td>554.8705</td>
<td>623.5230</td>
<td>0.837170</td>
<td>215.9501</td>
</tr>
<tr>
<td>10</td>
<td>Education Male Eat Diabetes Cancer Lung Walkroom</td>
<td>13</td>
<td>Racketni Male Eat Diabetes Cancer Lung Walkroom</td>
<td>555.2910</td>
<td>628.8472</td>
<td>0.837751</td>
<td>218.0852</td>
</tr>
<tr>
<td>DIAGNOSIS</td>
<td>CANCER</td>
<td>LUNG</td>
<td>WALKROOM</td>
<td>WAHRGUMM</td>
<td>AGECAT1-AGECAT6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>------</td>
<td>----------</td>
<td>----------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>21</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>555.9038</td>
<td>643.6593</td>
<td>0.849330</td>
<td>225.1495</td>
</tr>
<tr>
<td>19</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>21</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>556.1530</td>
<td>663.9247</td>
<td>0.847625</td>
<td>225.5916</td>
</tr>
<tr>
<td>8</td>
<td>AGECAT3 AGECAT6 MALE EAT DIABETES CANCER LUNG WALKROOM</td>
<td>12</td>
<td>MALE EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>556.2935</td>
<td>620.0423</td>
<td>0.830446</td>
<td>206.3481</td>
</tr>
<tr>
<td>19</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>21</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>556.3795</td>
<td>646.1512</td>
<td>0.846852</td>
<td>225.5627</td>
</tr>
<tr>
<td>10</td>
<td>AGECAT3 AGECAT5 AGECAT6 MALE EAT BMI DIABETES CANCER LUNG WALKROOM</td>
<td>13</td>
<td>MALE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>556.2689</td>
<td>625.2213</td>
<td>0.831473</td>
<td>214.1907</td>
</tr>
<tr>
<td>21</td>
<td>agc21 AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG ARTERY DEMENTIA WALKROOM</td>
<td>22</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG ARTERY DEMENTIA WALKROOM AGECAT1-AGECAT6</td>
<td>556.9254</td>
<td>669.5261</td>
<td>0.847700</td>
<td>225.4997</td>
</tr>
<tr>
<td>10</td>
<td>AGECAT3 AGECAT5 AGECAT6 MALE EAT BMI DIABETES CANCER LUNG WALKROOM</td>
<td>13</td>
<td>MALE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>557.0943</td>
<td>625.7468</td>
<td>0.831518</td>
<td>218.2982</td>
</tr>
<tr>
<td>20</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>22</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>557.3299</td>
<td>648.8539</td>
<td>0.847792</td>
<td>225.5988</td>
</tr>
<tr>
<td>20</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>22</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>557.2593</td>
<td>670.1633</td>
<td>0.849310</td>
<td>225.2496</td>
</tr>
<tr>
<td>20</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>22</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>557.6936</td>
<td>670.2408</td>
<td>0.847014</td>
<td>225.5926</td>
</tr>
<tr>
<td>9</td>
<td>AGECAT3 AGECAT6 MALE SMORE EAT CANCER LUNG WALKROOM</td>
<td>12</td>
<td>MALE SMORE CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>557.7487</td>
<td>631.5176</td>
<td>0.832164</td>
<td>211.2697</td>
</tr>
<tr>
<td>7</td>
<td>AGECA7 AGECAT6 MALE DIABETES CANCER LUNG WALKROOM</td>
<td>11</td>
<td>MALE DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>557.9715</td>
<td>616.8285</td>
<td>0.828947</td>
<td>199.224</td>
</tr>
<tr>
<td>9</td>
<td>AGECAT3 AGECAT5 AGECAT6 MALE EAT CANCER CHF LUNG WALKROOM</td>
<td>12</td>
<td>MALE EAT CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>558.8106</td>
<td>622.6201</td>
<td>0.821430</td>
<td>209.8785</td>
</tr>
<tr>
<td>21</td>
<td>AGECAT3 AGECAT4 AGECAT5 AGECAT6 racesth: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM</td>
<td>23</td>
<td>RACETH: EDUCATION MALE SMORE DRESS EAT BMI HYPTERTN DIABETES CANCER CHF LUNG DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6</td>
<td>558.9149</td>
<td>676.4132</td>
<td>0.847832</td>
<td>225.9595</td>
</tr>
<tr>
<td>9</td>
<td>AGECAT3 AGECAT6 AGECAT5 racesth: MALE EAT CANCER LUNG WALKROOM</td>
<td>12</td>
<td>MALE EAT CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>558.4638</td>
<td>623.2125</td>
<td>0.827559</td>
<td>211.1422</td>
</tr>
<tr>
<td>8</td>
<td>AGECAT3 AGECAT6 AGECAT5 MALE SMORE CANCER LUNG WALKROOM</td>
<td>11</td>
<td>MALE SMORE CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>559.5629</td>
<td>619.4199</td>
<td>0.830284</td>
<td>204.6670</td>
</tr>
<tr>
<td>7</td>
<td>AGECAT6 AGECAT5 MALE EAT CANCER LUNG WALKROOM</td>
<td>11</td>
<td>MALE EAT CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>561.6841</td>
<td>620.3291</td>
<td>0.816436</td>
<td>201.3640</td>
</tr>
</tbody>
</table>
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 4. Validation of the Clinical Index: New-Onset Disability at Discharge in Derivation and Validation Cohorts According to Risk Stratum

<table>
<thead>
<tr>
<th>Risk Stratum</th>
<th>Derivation Cohort</th>
<th>Validation Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients Disabled at Discharge/All Patients</td>
<td>% (95% CI)</td>
</tr>
<tr>
<td>Quintile of risk* (logistic regression model)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13/167</td>
<td>8 (4–13)</td>
</tr>
<tr>
<td>2</td>
<td>23/172</td>
<td>13 (9–19)</td>
</tr>
<tr>
<td>3</td>
<td>36/182</td>
<td>20 (14–26)</td>
</tr>
<tr>
<td>4</td>
<td>59/182</td>
<td>32 (26–40)</td>
</tr>
<tr>
<td>5</td>
<td>121/182</td>
<td>66 (59–73)</td>
</tr>
<tr>
<td>AUC</td>
<td>0.787</td>
<td></td>
</tr>
<tr>
<td>Risk group points (risk scoring system)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>7/125</td>
<td>6 (2–11)</td>
</tr>
<tr>
<td>1</td>
<td>27/207</td>
<td>13 (9–18)</td>
</tr>
<tr>
<td>2</td>
<td>30/167</td>
<td>18 (12–24)</td>
</tr>
<tr>
<td>3</td>
<td>42/125</td>
<td>34 (25–43)</td>
</tr>
<tr>
<td>4</td>
<td>30/85</td>
<td>35 (25–46)</td>
</tr>
<tr>
<td>5</td>
<td>21/47</td>
<td>45 (30–60)</td>
</tr>
<tr>
<td>6</td>
<td>23/46</td>
<td>50 (35–65)</td>
</tr>
<tr>
<td>7</td>
<td>21/28</td>
<td>75 (55–89)</td>
</tr>
<tr>
<td>8</td>
<td>16/18</td>
<td>89 (65–99)</td>
</tr>
<tr>
<td>9</td>
<td>13/15</td>
<td>87 (60–98)</td>
</tr>
<tr>
<td>≥10</td>
<td>22/22</td>
<td>100 (85–100)</td>
</tr>
<tr>
<td>AUC</td>
<td>0.784</td>
<td></td>
</tr>
<tr>
<td>Hosmer-Lemeshow test P-value</td>
<td>.40</td>
<td></td>
</tr>
</tbody>
</table>
Discrimination Plot (Steyerberg, 2010)

A Development data, slope=0.3
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
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
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 normality; 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)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

------------+------------------
17,751 | 1 1 1 1 1 1

| 954 | 1 1 1 0 0 0

618 | 1 1 0 1 1 1

464 | 1 0 1 1 1 1

62 | 1 0 1 0 0 0

57 | 1 1 0 0 0 0

42 | 0 1 0 1 1 1

29 | 1 0 0 1 1 1

22 | 0 1 0 0 0 0

Variables are (1) smoke (2) maxwt (3) age (4) dbp (5) sbp
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
------------------------------------------------------------------------------
                               Coef.     Std. Err.     t     P>|t|     [95\% Conf. Interval]
------------------------------------------------------------------------------
    kstones | Coef.     Std. Err.     t     P>|t|     [95\% Conf. Interval]
------------------------------------------------------------------------------
sbp     |  0.0013643  0.0020714   0.66   0.510     -0.0027044   0.005433
     |  0.0002225  0.0000516   0.10   0.066     0.0002038   0.0002842
    dbp     |  0.0059833  0.0030078   1.99   0.047     0.0000822   0.0118843
     |  0.0002665  0.0000477   0.08   0.009     0.0002537   0.0003112
    male    |  0.4154024  0.0730054   5.69   0.000     0.2723141   0.5584907
     |  0.0012741  0.0000642   0.02   0.000     0.0012587   0.0013017
     age     |  0.0270061  0.0022828  11.83   0.000     0.0225226   0.0314895
     |  0.0002419  0.0000342   0.20   0.000     0.0002485   0.0002558
    smoke   | -0.1498978  0.0771103  -1.94   0.052    -0.301032    0.0012364
     |  0.0018016  0.0001164   0.02   0.003     0.0017323   0.0018961
    maxwt   |  0.0050521  0.0008346   6.05   0.000     0.0034163   0.0066878
     |  0.0000302  2.72e-06   0.04   0.000     0.0000302   0.0000311
   _cons   | -6.207185   0.2778718  -22.34   0.000    -6.752047   -5.662324
     |  0.0200949  0.0046794   0.43   0.000     0.0157049   0.0271742
------------------------------------------------------------------------------
Note: values displayed beneath estimates are Monte Carlo error estimates.
```
. mi estimate, vartable nocitable

Multiple-imputation estimates Imputations = 10
Logistic regression

Variance information

<table>
<thead>
<tr>
<th></th>
<th>Imputation variance</th>
<th>Relative efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within</td>
<td>Between</td>
</tr>
<tr>
<td>sbp</td>
<td>3.7e-06</td>
<td>5.0e-07</td>
</tr>
<tr>
<td>dbp</td>
<td>8.3e-06</td>
<td>7.1e-07</td>
</tr>
<tr>
<td>male</td>
<td>.005312</td>
<td>.000016</td>
</tr>
<tr>
<td>age</td>
<td>4.6e-06</td>
<td>5.9e-07</td>
</tr>
<tr>
<td>smoke</td>
<td>.00591</td>
<td>.000032</td>
</tr>
<tr>
<td>maxwt</td>
<td>6.9e-07</td>
<td>9.1e-09</td>
</tr>
<tr>
<td>_cons</td>
<td>.072771</td>
<td>.004038</td>
</tr>
</tbody>
</table>
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
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

Cubic Spline Trajectory Example

Logarithmic Recovery Example


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

![Graph showing discrete timepoints of study with a horizontal axis labeled "Years on Study" and a vertical axis labeled "Functional Score". The graph displays multiple lines representing different subjects or conditions, with the years on study ranging from 0 to 8 and the functional score ranging from 0 to 120.]
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. $r_{ij} \equiv \log(h_{ij}t_{ij} + 1)$
- $\beta_1 =$ Intercept (average score at $t = 0$)
- $\beta_2 =$ Pre-hospitalization slope
- $\beta_3 =$ Amount score drops at time of hospitalization
- $\beta_4 =$ Recovery slope on logarithmic time scale
Logarithmic recovery model

Years Post–Hospitalization

Slope $\beta_2$

$\beta_1$

$\beta_3$

Slope (log) $\beta_4$
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).
Figure 1. Trajectories of Disability in the Last Year of Life among 383 Decedents.
Latent Trajectory Discussion

- Visually can be a 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 a 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


References (2)


References (3)