Predictive Modeling with SAS (for Health)

Lorne Rothman, PhD, P.Stat.
Principal Analytics Services
Lorne.Rothman@sas.com
Overview

- What is Predictive Modeling?
  - Purpose, challenges, and methods
  - Examples
  - The North Carolina Low Birth Weight Data

- Data & Variable Preparation
  - Oversampling, Missing Values, Data splitting & dimension reduction

- Binary Target Modeling
  - Decision Trees with HPSPLIT
  - Logistic Regression with LOGISTIC
  - Comparing ROC curves

- Continuous Target Modeling
  - Model selection with GLMSELECT

- Generalized Linear Predictive Modeling
  - Gamma regression model selection with HPGENSELECT
What is Predictive Modeling?
Purpose of Predictive Modeling

 ✓ To Predict the Future

 X To identify statistically significant attributes or risk factors

 X To publish findings in Science, Nature, or the New England Journal of Medicine

 ✓ To enhance & enable rapid decision making at the level of the individual patient, client, customer, etc.

 X To enable decision making and influence policy through publications and presentations
Data Deluge

OHIP  CIHI  Population Census  Canadian Cancer Registry  Canadian Community Health Survey  Canadian MIS Database  Canadian Organ Joint Replacement Registry  Canadian Organ Replacement Registry  Canadian Tobacco Use Monitoring Survey  Discharge Abstract Database  Disease Surveillance Online General Social Survey – Cycle 6  Health and Activity Limitation Survey  Health Services Access Survey  Hospital Mental Health Survey Hospital Morbidity Database  Maternity Experiences Survey  Natchem/Particulate Matter Database  Natchem/Precipitation Database  National Alcohol and Drug Survey National Pollutant Release Inventory National Population Health Survey (Household)  National Rehabilitation Reporting System  National Trauma Registry Survey on Aging and Independence  Survey on Smoking in Canada  TRAID  Vital Statistics
Challenges: Rare Events
Methodology: Empirical Validation
Predicting the Future with Data Splitting

- Models are fit to Training Data, compared and selected on Validation and tested on a future Test set.
Methodology: Diversity of Algorithms
Jargon...

- **Target** = Dependent Variable.
- **Inputs, Predictors** = Independent Variables.
- **Supervised Classification** = Predicting class membership with algorithms that use a target.
- **Scoring** = The process of generating predictions on new data for decision making. This is not a re-running of models but an application of model results (e.g. equation and parameter estimates) to new data.
- **Scoring Code** = programming code that can be used to prepare and generate predictions on new data including transformations, imputation results, and model parameter estimates and equations.
Examples...

SAS Discharge Disposition and Length of Stay Modeling for Hospitals

- **Length of stay:** Survival modeling to predict ‘target’ discharge date up to 2 days prior to discharge for patients who end up going home with care or without care.

- **Discharge disposition:** Predict discharge disposition 2 days prior to patient discharge for those patients who will go home with home care and those who will go home without homecare.

- Data: Use daily in-hospital data from admissions, OR, DI, pharmacy, lab tests, etc. to score patients daily.

**Predicting End Stage Renal Failure**
Survival modeling to predict probability of developing End Stage Renal Disease given patient attributes and kidney function measures.

**What makes these predictive models?**
- Same algorithms as employed in inferential statistics, but different methodology and modeling **purpose—to score individuals in near real time and use results for rapid and preemptive decision making.**
The North Carolina Birth Records Data


- 7.2% low birth weight births (< 2500 grams) excluding multiple births.

- Data contains information on parents ethnicity, age, education level and marital status.

- Data contains information on mothers health condition and reproductive history.

- 45 potential predictor variables for modeling.
**Scenario: Early Warning System for Birth Weight**

**PREDICTORS**

- **Parent socio-, eco-, demo- graphics, health and behaviour**
  - Age, edu, race, medical conditions, smoking, drinking etc.

- **Prior pregnancy related data**
  - # pregnancies, last outcome, prior pregnancies etc.

- **Medical History for pregnancy**
  - Hypertension during pregnancy, eclampsia, incompetent cervix, etc.

- **Obstetric procedures**
  - Amniocentesis, ultrasound, etc.

- **Events of Labor**
  - Breech, fetal distress etc.

- **Method of delivery**
  - Vaginal, c-section etc.

- **New born characteristics**
  - Congenital anomalies (spinabifida, heart), APGAR score, anemia
Beware of Temporal Infidelity…..

- Parent socio-, eco-, demo- graphics and behaviour
- Prior health and pregnancy related data
- Medical history for this pregnancy
- Obstetric procedures
- Events of Labor
- Method of delivery
- New born characteristics

Time
Data & Variable Preparation
Preparing the Modeling Data

- time range
- attributes
- latency period
- target window
- start date
- $T_{\text{max}}$
- relative time
- end date
- 0
- target date

Attributes include:
- latency period
- target window

Target date is represented by the 0 mark on the relative time axis.
Oversample Rare Events

```
proc sort data=bwt00;
   by lbwt;
run;

proc surveyselect data=bwt00
   samprate=(.075,1) out=Osbwt00 seed=5;
   strata lbwt;
run;

proc freq data=Osbwt00;
   tables lbwt;
run;
```

- SURVEYSELECT is used to sample 7.5% of non-events and 100% of events.
- Data must be sorted by the target prior to oversampling.
Create Missing Indicators

- Create missing indicators to capture associations between missingness and the target in development data.
- The process is repeated for Test data.
- This step is unnecessary for Decision Trees as they accommodate missing values directly.

```sas
data bwt00;
set bwt00;

array vars{*} fage mage feduc meduc totalp bdead terms
   loutcome prenatal marital children cignum drinknum anemia
   cardiac aclung diabetes herpes hydram hemoglob hyperch
   hyperpr eclamp cervix pinfant preterm renal rhsen uterine
   amnio ultra YrsLastLiveBirth YrsLastFetalDeath drinker smoker marital;

array mvars{*} M_fage M_mage M_feduc M_meduc M_totalp M_bdead M_terms
   M_loutcome M_prenatal M_marital M_children M_cignum M_drinknum M_anemia
   M_cardiac M_aclung M_diabetes M_herpes M_hyadram M_hemoglob M_hyperch
   M_hyperpr M_eclamp M_cervix M_pinfant M_preterm M_renal M_rhsen M_uterine
   M_amnio M_ultra M_YrsLastLiveBirth M_YrsLastFetalDeath M_drinker M_smoker M_marital;
   do i=1 to dim(vars);
      mvars{i}=(vars{i}=.);
   end;
run;
```
Partition Data for Empirical Validation

- SURVEYSELECT is used to partition data into Training (67%) and Validation (33%) sets.
- The OUTALL option provides one dataset with a variable, SELECTED that indicates dataset membership.
- For class targets, stratification on the target, LBWT ensures equal representation of low birth weight cases in training and validation sets.
- Since HPSPLIT and HPGENSELECT procedures do not accept separate train and validate sets, the dataset output from SURVEYSELECT will be used before physically splitting into train and validation data.

```sas
proc surveyselect
  data=OSbwt00
  samprate=.6667
  out=dev00
  seed=44444
  outall;
  strata lbwt;
run;

data train valid;
  set dev00;
  if selected then output train;
  else output valid;
run;
```
Impute Missing Values

- STDIZE will replace missing values (REPONLY) and is applied to the Training data.
- The OUTSTAT option saves a dataset to be used to insert results (score) into Validation and Test sets.
- The METHOD=IN (MED) uses the imputation information from the training data to score the Validation and Test data.
- Imputation is unnecessary for Decision Trees as they accommodate missing values directly.

```sas
proc stdize data=train reponly method=median out=trainI outstat=med;
   var &numvars;
run;

proc stdize data=valid out=validI reponly method=in(med);
   var &numvars;
run;

proc stdize data=test01 out=testI reponly method=in(med);
   var &numvars;
run;
```
Cluster Variables to Reduce Dimensions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Variable</th>
<th>Correlation</th>
<th>Lambda</th>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2</td>
<td>CHILDREN</td>
<td>0.8025</td>
<td>0.0916</td>
<td>0.2174</td>
<td>Number of children now living</td>
</tr>
<tr>
<td></td>
<td>LOUTCOME</td>
<td>0.7907</td>
<td>0.0801</td>
<td>0.2275</td>
<td>Outcome of last delivery</td>
</tr>
<tr>
<td></td>
<td>TOTALP</td>
<td>0.7398</td>
<td>0.4152</td>
<td>0.4449</td>
<td>Total pregnancies (including this one)</td>
</tr>
<tr>
<td></td>
<td>M_YrsLastLiveBirth</td>
<td>0.8245</td>
<td>0.0782</td>
<td>0.1904</td>
<td></td>
</tr>
<tr>
<td>Cluster 3</td>
<td>FEDUC</td>
<td>0.8554</td>
<td>0.1557</td>
<td>0.1713</td>
<td>Education of father (years)</td>
</tr>
<tr>
<td></td>
<td>MEDUC</td>
<td>0.8554</td>
<td>0.1737</td>
<td>0.1751</td>
<td>Education of mother (years)</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>M_terms</td>
<td>0.9329</td>
<td>0.0830</td>
<td>0.0732</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M_totalp</td>
<td>0.9329</td>
<td>0.2496</td>
<td>0.0894</td>
<td></td>
</tr>
<tr>
<td>Cluster 5</td>
<td>M_cignum</td>
<td>0.9537</td>
<td>0.0707</td>
<td>0.0499</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M_drinknum</td>
<td>0.9537</td>
<td>0.0859</td>
<td>0.0507</td>
<td></td>
</tr>
<tr>
<td>Cluster 6</td>
<td>M_fage</td>
<td>0.9257</td>
<td>0.0381</td>
<td>0.0772</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M_feduc</td>
<td>0.9244</td>
<td>0.0379</td>
<td>0.0786</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MARITAL</td>
<td>0.5320</td>
<td>0.1420</td>
<td>0.5455</td>
<td>Marital status</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>CIGNUM</td>
<td>0.8454</td>
<td>0.2511</td>
<td>0.2065</td>
<td>Average # of cigarettes daily</td>
</tr>
<tr>
<td></td>
<td>smoker</td>
<td>0.8454</td>
<td>0.0268</td>
<td>0.1689</td>
<td></td>
</tr>
<tr>
<td>Cluster 8</td>
<td>DRINKNUM</td>
<td>0.6777</td>
<td>0.0007</td>
<td>0.3225</td>
<td>Average # of alcoholic drinks per week</td>
</tr>
<tr>
<td></td>
<td>drinker</td>
<td>0.6777</td>
<td>0.2236</td>
<td>0.4151</td>
<td></td>
</tr>
</tbody>
</table>

Cluster variables on training data to reduce collinearity prior to modeling. E.g. PROC VARCLUS.

```
proc varclus data=train maxeigen=.7 short hi;
var    &IntervalandFlagVars;
run;
```
Variables RACEMOM and RACEDAD contain 9 and 10 levels respectively.

Use a Decision Tree model to optimally collapse many possible combinations of these attributes to a single 6-level variable using training data.

This step is unnecessary if you are using a decision tree as a predictive model.
Binary Target Modeling
New Modeling Routines in SAS/STAT: Decision Trees using HPSPLIT

- In SAS 9.4 some of the high performance procedures used in Enterprise Miner software for data mining are now available in SAS/STAT (now called version 14.1).

- Procedures support parallel processing and are designed to run in a distributed computing environment (across multiple servers for high speed computing).

- Procedures are available to run on a single machine or server. They are now shipped with SAS/STAT at no additional cost.

- In this section we will feature HPSPLIT for decision trees using a binary target and in a later section, HPGENSELECT for generalized linear models and mixture distributions using a continuous target.
The program fits a CART-like decision tree to low birth weight data: with surrogates, GINI splitting criterion, Cost-complexity pruning (Breiman et al.), and data splitting (PARTITION) from a single development dataset using a flag variable (SELECTED) that indicates train/validate membership.

CHAID-like and machine learning-like trees are also possible.

Tree plots are subset using a ZOOMEDTREE option. Node details are printed using a NODES options.

Minimum leaf and class variable sizes, maximum branches, and maximum depth are set using MINLEAFSIZE, MINCATSIZE, MAXBRANCH, and MAXDEPTH options.

Missing values are assigned to the branch with largest sample size (ASSIGNMISSING=POPULAR).

Data Step scoring code is saved to a file using the CODE statement, available in PROCs: GENMOD, GLIMMIX, GLM, GLMSELECT, LOGISTIC, MIXED, REG, HPSPLIT, HPGENSELECT and others…
Cost-Complexity Pruning with the PRUNE Statement

- Though the full tree has over 80 leaves, Cost-Complexity Pruning on misclassification rate yields a 16 leaf tree that minimizes misclassification rate.
The Final Tree

- Nodes and leaves are identified with numbers and letters.
- The width of the curves are proportional to amount of data passing through each part of the tree.
- Red indicates low bwt classification while blue indicates normal weight.
Plot Tree Sections Using the ZOOMEDTREE option

- Classifications using a cutoff of 0.5 are given at the top of each node (e.g., Node T is classified as low birth weight while Node U is classified as normal weight).

- Samples sizes for Train/Validate data, as well as proportions of target 1 in each are shown.

- Red indicates low bwt classification while blue indicates normal weight.
## Display Leaf Rules using the NODES Option

<table>
<thead>
<tr>
<th>ID</th>
<th>Path</th>
<th>Training Data</th>
<th>Validation Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Root Node</td>
<td>11096</td>
<td>0.4963</td>
</tr>
<tr>
<td></td>
<td>RACEDAD = 2,3,8,9</td>
<td>4858</td>
<td>0.3895</td>
</tr>
<tr>
<td></td>
<td>PRENATAL &lt; 0.09</td>
<td>192</td>
<td>0.1250</td>
</tr>
<tr>
<td>5</td>
<td>Root Node</td>
<td>11096</td>
<td>0.4963</td>
</tr>
<tr>
<td></td>
<td>RACEDAD = 0,1,4,5,6,7 or Missing</td>
<td>6238</td>
<td>0.5795</td>
</tr>
<tr>
<td></td>
<td>smoker = 1</td>
<td>1071</td>
<td>0.4006</td>
</tr>
<tr>
<td>7</td>
<td>Root Node</td>
<td>11096</td>
<td>0.4963</td>
</tr>
<tr>
<td></td>
<td>RACEDAD = 2,3,8,9</td>
<td>4858</td>
<td>0.3895</td>
</tr>
<tr>
<td></td>
<td>PRENATAL &gt;= 0.09 or Missing</td>
<td>4666</td>
<td>0.4003</td>
</tr>
<tr>
<td></td>
<td>PRETERM = 1</td>
<td>147</td>
<td>0.1088</td>
</tr>
<tr>
<td>A</td>
<td>Root Node</td>
<td>11096</td>
<td>0.4963</td>
</tr>
<tr>
<td></td>
<td>RACEDAD = 0,1,4,5,6,7 or Missing</td>
<td>6238</td>
<td>0.5795</td>
</tr>
<tr>
<td></td>
<td>smoker = 0 or Missing</td>
<td>5167</td>
<td>0.6166</td>
</tr>
</tbody>
</table>

... etc etc.

- Rules details for all leaves are reported. Asterisk indicates selected target level.
Explore Variable Importance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Label</th>
<th>Training</th>
<th>Validation</th>
<th>Relative</th>
<th>Importance</th>
<th>Relative</th>
<th>Importance</th>
<th>Relative</th>
<th>Ratio</th>
<th>Count</th>
<th>Times Used as a Surrogate</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACEMOM</td>
<td>Race of mother</td>
<td>1.0000</td>
<td>1.0000</td>
<td>15.7420</td>
<td>10.9894</td>
<td>1.0000</td>
<td>1</td>
<td>4</td>
<td></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>RACEDAD</td>
<td>Race of father</td>
<td>0.9258</td>
<td>0.9661</td>
<td>14.5744</td>
<td>10.6169</td>
<td>1.0435</td>
<td>1</td>
<td>2</td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MARITAL</td>
<td>Marital status</td>
<td>0.7789</td>
<td>0.8156</td>
<td>12.2611</td>
<td>8.9634</td>
<td>1.0472</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CIGNUM</td>
<td>Average # of cigarettes daily</td>
<td>0.6946</td>
<td>0.7792</td>
<td>10.9339</td>
<td>8.5631</td>
<td>1.1219</td>
<td>3</td>
<td>1</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>smoker</td>
<td></td>
<td>0.6556</td>
<td>0.7416</td>
<td>10.3209</td>
<td>8.1502</td>
<td>1.1312</td>
<td>0</td>
<td>3</td>
<td></td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>drinker</td>
<td></td>
<td>0.5687</td>
<td>0.6257</td>
<td>8.9527</td>
<td>6.8757</td>
<td>1.1001</td>
<td>0</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CHILDREN</td>
<td>Number of children now living</td>
<td>0.5918</td>
<td>0.4677</td>
<td>9.3157</td>
<td>5.1398</td>
<td>0.7903</td>
<td>3</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TOTALP</td>
<td>Total pregnancies (including this one)</td>
<td>0.5625</td>
<td>0.4445</td>
<td>8.8553</td>
<td>4.8843</td>
<td>0.7901</td>
<td>0</td>
<td>3</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>MAGE</td>
<td>Age of mother</td>
<td>0.5500</td>
<td>0.4154</td>
<td>8.6587</td>
<td>4.5652</td>
<td>0.7552</td>
<td>0</td>
<td>7</td>
<td></td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>PRETERM</td>
<td>Prev. preterm/small</td>
<td>0.3226</td>
<td>0.3099</td>
<td>5.0788</td>
<td>3.4058</td>
<td>0.9606</td>
<td>1</td>
<td>0</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DRINKNUM</td>
<td>Average # of alcoholic drinks per week</td>
<td>0.3172</td>
<td>0.3047</td>
<td>4.9936</td>
<td>3.3486</td>
<td>0.9606</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PRENATAL</td>
<td>Month of preg. prenatal care began</td>
<td>0.3598</td>
<td>0.2947</td>
<td>5.6632</td>
<td>3.2365</td>
<td>0.8186</td>
<td>2</td>
<td>0</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>TERMS</td>
<td>Number of other terminations</td>
<td>0.2155</td>
<td>0.2727</td>
<td>3.3926</td>
<td>2.9972</td>
<td>1.2655</td>
<td>0</td>
<td>2</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MFNUC</td>
<td>Education of mother (years)</td>
<td>0.2124</td>
<td>0.1767</td>
<td>3.3440</td>
<td>1.9420</td>
<td>0.8319</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FEDUC</td>
<td>Education of father (years)</td>
<td>0.1879</td>
<td>0.1594</td>
<td>2.9586</td>
<td>1.7513</td>
<td>0.8479</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RENAL</td>
<td>Renal disease</td>
<td>0.1151</td>
<td>0.1255</td>
<td>1.8113</td>
<td>1.3790</td>
<td>1.0006</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YrsLastLiveBirth</td>
<td></td>
<td>0.2375</td>
<td>0.1070</td>
<td>3.7390</td>
<td>1.1761</td>
<td>0.4506</td>
<td>2</td>
<td>0</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>BDEAD</td>
<td>Number previous live births now dead</td>
<td>0.1606</td>
<td>0.0963</td>
<td>2.5284</td>
<td>1.0583</td>
<td>0.5996</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Race of mother and father as well as marital status and smoking behavior are the top most important variables.
### Fit Statistics for Selected Tree

<table>
<thead>
<tr>
<th></th>
<th>N Leaves</th>
<th>ASE</th>
<th>Mis-class</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Entropy</th>
<th>Gini</th>
<th>RSS</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>16</td>
<td>0.2278</td>
<td>0.3731</td>
<td>0.6264</td>
<td>0.6274</td>
<td>0.9331</td>
<td>0.4556</td>
<td>5055.7</td>
<td>0.6674</td>
</tr>
<tr>
<td>Validation</td>
<td>16</td>
<td>0.2291</td>
<td>0.3688</td>
<td>0.6346</td>
<td>0.6277</td>
<td>0.9355</td>
<td>0.4568</td>
<td>2541.6</td>
<td>0.6627</td>
</tr>
</tbody>
</table>

#### ROC Curve for lbwt

- Training AUC: 0.87
- Validation AUC: 0.86

- Training and Validation assessment measures and overlayed ROC curves are output.
Score Data using HPSPLIT Scoring Code

```sas
data scoredtreeval(keep=p_tree p_ladjtree idnum);
  set DEV00;
  %include 'C:\data\EDU\TALKS\HUG2015\lbwtTREE.sas'/source2;
  P_1ADJtree=(P_lbwt1*(1-\&RHO1)*\&PRIOR1)/
    (1-P_LBWT1)*\&RHO1*(1-\&PRIOR1)+(P_LBWT1*(1-\&RHO1)*\&PRIOR1));
  if selected=0;
  p_Tree=p_lbwt1;
run;
```

Etc

```sas
data scoredtreetest(keep=p_tree p_ladjtree idnum);
  set test01;
  %include 'C:\data\EDU\TALKS\HUG2015\lbwtTREE.sas'/source2;
  P_1ADJtree=(P_lbwt1*(1-\&RHO1)*\&PRIOR1)/
    (1-P_LBWT1)*\&RHO1*(1-\&PRIOR1)+(P_LBWT1*(1-\&RHO1)*\&PRIOR1));
  p_Tree=p_lbwt1;
run;
```

Etc

- Validation and Test data are scored using Decision Tree Scoring code.
- Probabilities can be adjusted for oversampling (P_1ADJtree) if desired though this is not required for ROC curve assessments.
- Validation and Test scores from the tree model are match-merged back to the corresponding imputed sets to be used for regression (not shown here).
The SCORE statements allow for scoring of new data (Validation and Test) and adjusts oversampled data back to the population prior (PRIOREVENT=0.072).

The ALLVAL and ALLTEST sets containing decision tree predictions are supplied in the first regression run. The same datasets are re-scored (SCO_VALIDATE, SCO_TEST), and prediction variables renamed, so that predictions for all four models are in the same set for comparisons.
In general, predictive models fit to larger datasets tend to have more parameters than more theoretically informed explanatory models in health.

Odds ratios for previous premature babies (PRETERM), renal disease (RENAL), and chronic hypertension (HYPERCH) and are particularly large.

The collapsed version of mother and father race from an initial decision tree (TREE_RACE) appears in the model.
Model Assessments for Binary Targets

** - Where Predicted 1=(Pred Prob > Cutoff)

Accuracy   = (TP+TN)/n
Sensitivity = TP/AP
Specificity = TN/AN
Lift            = (TP/PP)/π₁
Assessment Charts for Binary Targets

Lift Charts

ROC Charts

Explore measures across a range of cutoffs
**Receiver Operator Curves**

- A weak model is one that discriminates between target class levels poorly, with areas under the curve ranging from 0.5 to 1.0.
- A strong model is one that discriminates well, with areas under the curve nearing 1.0.

A concordance statistic: for every pair of observations with different outcomes (LBWT=1, LBWT=0) AuROC measures the probability that the ordering of the predicted probabilities agrees with the ordering of the actual target values.

...Or the probability that a low birth weight baby (LBWT=1) has a higher predicted probability of low birth weight than a normal birth weight baby (LBWT=0).
Predict the Future with Data Splitting

- Models are fit to Training Data, compared and selected on Validation and tested on a future Test set.
Assess Models using ROC Curves

```sas
title "Comparing Four Models on Validation Data";
ods html;
ods graphics on;
proc logistic data=sco_validate;
   model lbwt(event='1')=p_all p_allint p_early p_tree / nofit;
   roc "Backward Regression" p_all;
   roc "Backward Interactions Regression" p_allInt;
   roc "Backward Early Warning Regression" p_early;
   roc "Early Warning Decision Tree" p_Tree;
   roccontrast "Comparing the Four Models: Validation Data "/estimate=allpairs;
run;
```

- The dataset with all four predictions (SCO_VALIDATE) is supplied to PROC LOGISTIC.

- The ROCCONTRAST statements provides hypothesis tests for differences between ROC curves, for model results specified in the three ROC statements.

- To generate ROC contrasts, all terms used in the ROC statements must be placed on the model statement. The NOFIT option suppresses the fitting of the specified model.

- Because of the presence of the ROC and ROCCONTRAST statements, ROC plots are generated when ODS GRAPHICS are enabled.

- The identical process is repeated with the scored Test set, SCO_TEST. Can the model predict the future?
Compare ROC Curves on Validation Data

ROC Curves for Comparisons

ROC Curve (Area)
- Backward Regression (0.7266)
- Backward Interactions Regression (0.7264)
- Backward Early Warning Regression (0.6822)
- Early Warning Decision Tree (0.6627)
## Compare AuROC Curves on Validation Data

### ROC Association Statistics

<table>
<thead>
<tr>
<th>ROC Model</th>
<th>Area</th>
<th>Standard Error</th>
<th>95% Wald Confidence Limits</th>
<th>Somers' D</th>
<th>Gamma</th>
<th>Tau-a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward Regression</td>
<td>0.7266</td>
<td>0.00670</td>
<td>0.7135 - 0.7397</td>
<td>0.4532</td>
<td>0.4540</td>
<td>0.2266</td>
</tr>
<tr>
<td>Backward Interactions Regression</td>
<td>0.7264</td>
<td>0.00670</td>
<td>0.7133 - 0.7396</td>
<td>0.4529</td>
<td>0.4536</td>
<td>0.2265</td>
</tr>
<tr>
<td>Backward Early Warning Regression</td>
<td>0.6822</td>
<td>0.00708</td>
<td>0.6684 - 0.6961</td>
<td>0.3645</td>
<td>0.3645</td>
<td>0.1823</td>
</tr>
<tr>
<td>Early Warning Decision Tree</td>
<td>0.6627</td>
<td>0.00719</td>
<td>0.6486 - 0.6768</td>
<td>0.3253</td>
<td>0.3644</td>
<td>0.1627</td>
</tr>
</tbody>
</table>

### ROC Contrast Test Results

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing the Four Models: Validation Data</td>
<td>3</td>
<td>147.7898</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

### ROC Contrast Estimation and Testing Results by Row

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Wald Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward Regression - Backward Interactions Regression</td>
<td>0.00164</td>
<td>0.00107</td>
<td>-0.00193 - 0.00226</td>
<td>0.0234</td>
<td>0.8784</td>
</tr>
<tr>
<td>Backward Regression - Backward Early Warning Regression</td>
<td>0.0443</td>
<td>0.00422</td>
<td>0.0361 - 0.0526</td>
<td>110.1907</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Backward Regression - Early Warning Decision Tree</td>
<td>0.0639</td>
<td>0.00565</td>
<td>0.0528 - 0.0750</td>
<td>127.8547</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Backward Interactions Regression - Backward Early Warning Regression</td>
<td>0.0442</td>
<td>0.00433</td>
<td>0.0357 - 0.0527</td>
<td>103.8641</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Backward Interactions Regression - Early Warning Decision Tree</td>
<td>0.0638</td>
<td>0.00561</td>
<td>0.0528 - 0.0747</td>
<td>129.2086</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Backward Early Warning Regression - Early Warning Decision Tree</td>
<td>0.0196</td>
<td>0.00444</td>
<td>0.0109 - 0.0283</td>
<td>19.4494</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Compare ROC Curves on Test Data

ROC Curves for Comparisons

ROC Curve (Area)
- Backward Regression (0.7173)
- Backward Interactions Regression (0.7176)
- Backward Early Warning Regression (0.6747)
- Early Warning Decision Tree (0.6550)
# Compare AuROC Curves on Test Data

## ROC Association Statistics

<table>
<thead>
<tr>
<th>ROC Model</th>
<th>Mann-Whitney</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Area</td>
<td>Standard Error</td>
<td>95% Wald Confidence Limits</td>
<td>Somers' D</td>
<td>Gamma</td>
<td>Tau-a</td>
</tr>
<tr>
<td>Backward Regression</td>
<td>0.7173</td>
<td>0.00392</td>
<td>0.7097</td>
<td>0.7250</td>
<td>0.4347</td>
<td>0.4353</td>
</tr>
<tr>
<td>Backward Interactions Regression</td>
<td>0.7176</td>
<td>0.00392</td>
<td>0.7100</td>
<td>0.7253</td>
<td>0.4353</td>
<td>0.4359</td>
</tr>
<tr>
<td>Backward Early Warning Regression</td>
<td>0.6747</td>
<td>0.00412</td>
<td>0.6666</td>
<td>0.6828</td>
<td>0.3494</td>
<td>0.3494</td>
</tr>
<tr>
<td>Early Warning Decision Tree</td>
<td>0.6550</td>
<td>0.00416</td>
<td>0.6468</td>
<td>0.6631</td>
<td>0.3100</td>
<td>0.3477</td>
</tr>
</tbody>
</table>

## ROC Contrast Test Results

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing the Four Models: Test Data</td>
<td>3</td>
<td>401.7402</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

## ROC Contrast Estimation and Testing Results by Row

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Wald Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward Regression - Backward Interactions Regression</td>
<td>-0.00029</td>
<td>0.000570</td>
<td>-0.00141</td>
<td>0.000828</td>
<td>0.2584</td>
</tr>
<tr>
<td>Backward Regression - Backward Early Warning Regression</td>
<td>0.0426</td>
<td>0.00248</td>
<td>0.0378</td>
<td>0.0475</td>
<td>294.6697</td>
</tr>
<tr>
<td>Backward Regression - Early Warning Decision Tree</td>
<td>0.0624</td>
<td>0.00338</td>
<td>0.0557</td>
<td>0.0690</td>
<td>341.1706</td>
</tr>
<tr>
<td>Backward Interactions Regression - Backward Early Warning Regression</td>
<td>0.0429</td>
<td>0.00253</td>
<td>0.0380</td>
<td>0.0479</td>
<td>287.3259</td>
</tr>
<tr>
<td>Backward Interactions Regression - Early Warning Decision Tree</td>
<td>0.0627</td>
<td>0.00334</td>
<td>0.0561</td>
<td>0.0692</td>
<td>351.3665</td>
</tr>
<tr>
<td>Backward Early Warning Regression - Early Warning Decision Tree</td>
<td>0.0197</td>
<td>0.00269</td>
<td>0.0144</td>
<td>0.0250</td>
<td>53.8845</td>
</tr>
</tbody>
</table>
Compare Lift Charts on Test Data

Early Warning Decision Tree
- Individuals in the top 5% most likely to have low birth weight babies are about 2.5 x more likely than average to have a lbwt baby.

All Effects Regression
- Individuals in the top 5% most likely to have low birth weight babies are about 3.5 x more likely than average to have a lbwt baby.
Continuous Target Modeling
Continuous Target: Birth Weight

Distribution of bwtkg

Percent

bwtkg

Copyright © 2006, SAS Institute Inc. All rights reserved.
Build Regression Models with GLMSELECT

- GLMSELECT fits continuous target models (under GLM assumptions) and can process validation and test datasets, or perform cross validation for smaller datasets. It can also perform data partition using the PARTITION statement.

- GLMSELECT supports a class statement similar to PROC GLM but is designed for predictive modeling.

- Selection methods include Backward, Forward, Stepwise, LAR and LASSO.
Least Angle Regression

- Standardize inputs and response. All coefficients are zero.

- The predictor, X1 that is most correlated with the current residual (*makes the least angle with the residual*) is determined and a step is taken in the direction of this predictor (*X1 is added to model*). The length of this step (*the coefficient magnitude*) is chosen so that some other predictor, X2 and the current predicted response have the same correlation with the current residual (*equiangular*).

- At this point, the predicted response moves in the direction that is equiangular between (*equally correlated with*) these two predictors. Moving in this direction ensures that these two predictors continue to have a common correlation with the current residual.

- The predicted response moves in this direction until a third predictor, X3 has the same correlation with the current residual as the two predictors already in the model. A new direction is determined that is equiangular between these three predictors and the predicted response moves in this direction until a fourth predictor joins the set having the same correlation with the current residual.

- This process continues until all predictors are in the model.
Select Models with GLMSELECT

<table>
<thead>
<tr>
<th>title1 'BACKWARD: select SL choose validate'</th>
</tr>
</thead>
<tbody>
<tr>
<td>ods graphics on;</td>
</tr>
<tr>
<td>proc glmselect data=lbwt.train valdata=lbwt.valid /<em>testdata=test01</em>/</td>
</tr>
<tr>
<td>plots(stepAxis=number)=ASEPlot;</td>
</tr>
<tr>
<td>class tree_race;</td>
</tr>
<tr>
<td>model bwtkg = &amp;numvars tree_race</td>
</tr>
<tr>
<td>/selection=backward(choose = validate select = sl slstay=0.00001)</td>
</tr>
<tr>
<td>showpvalues;</td>
</tr>
<tr>
<td>run;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title1 'BACKWARD: select SBC choose SBC'</th>
</tr>
</thead>
<tbody>
<tr>
<td>ods graphics on;</td>
</tr>
<tr>
<td>proc glmselect data=lbwt.train valdata=lbwt.valid</td>
</tr>
<tr>
<td>plots=all;</td>
</tr>
<tr>
<td>class tree_race;</td>
</tr>
<tr>
<td>model bwtkg = &amp;numvars tree_race</td>
</tr>
<tr>
<td>/selection=backward(choose = sbc select = sbc) showpvalues;</td>
</tr>
<tr>
<td>run;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title1 'LAR: Choose Validation ASE'</th>
</tr>
</thead>
<tbody>
<tr>
<td>ods graphics on;</td>
</tr>
<tr>
<td>proc glmselect data=lbwt.train valdata=lbwt.valid</td>
</tr>
<tr>
<td>plots=all;</td>
</tr>
<tr>
<td>class tree_race;</td>
</tr>
<tr>
<td>model bwtkg = &amp;numvars tree_race marital</td>
</tr>
<tr>
<td>/selection=lar(choose = validate stop=none);</td>
</tr>
<tr>
<td>run;</td>
</tr>
</tbody>
</table>

- **SELECTION=LAR** requests Least Angle Regression.
- **SELECT=** determines the order in which effects enter or leave the model. Options include, for example: ADJRSQ, AIC, SBC, CP, CV, RSQUARE and SL. SL uses the traditional approach of significance level. SELECT is not available for LAR and LASSO.
- Models can be tuned with the **CHOOSE=** option to select the step in a selection routine using e.g. AIC, SBC, Mallow's CP, or validation data error. **CHOOSE=VALIDATE** selects that step that minimizes Validation data error.
Backward Model Tuning using Validation ASE

Progression of Average Squared Errors by Role for bwtkg

- **Average Squared Error**
  - 0.340
  - 0.338
  - 0.336
  - 0.334

- **Step**
  - 0
  - 10
  - 20
  - 30

- **Graph**
  - Training (blue)
  - Validation (green)

Selected Step
Backward Model: Select and Choose using SBC
LAR Model Tuning using Validation ASE
Assess Final Model

GLMSELECT does not provide model diagnostics.

The model selected by GLMSELECT can be refit in PROC GLM.

PLOTS=DIAGNOSTICS requests diagnostic plots. With larger datasets the user may have to increase the number of allowable plotting points using the MAXPOINTS= option.
PROC GLM Model Estimates

- Tuning a model on Validation data error, especially when using a backward regression does not guarantee that all terms in the final model will be significant at the 5% level.

| Parameter               | Estimate | Standard Error | t Value | Pr > |t| |
|-------------------------|----------|----------------|---------|------|---|
| Intercept               | 3.393074829 | 0.09321956 | 36.40   | <.0001 |
| HYPERCH                 | -0.316437355 | 0.07078175 | -4.47   | <.0001 |
| CHILDREN                | 0.176907601 | 0.03535468 | 5.00    | <.0001 |
| CIGNUM                  | -0.009032741 | 0.00241774 | -3.74   | 0.0022 |
| DIABETES                | 0.063329386 | 0.04046288 | 1.57    | 0.1176 |
| FEDUC                   | 0.004833165 | 0.00356407 | 1.36    | 0.1751 |
| LOUTCOME                | 0.043116199 | 0.02452329 | 1.76    | 0.0788 |
| MEDUC                   | 0.12849662 | 0.00346726 | 3.71    | 0.0022 |
| PINFANT                 | 0.470098794 | 0.07583535 | 6.20    | <.0001 |
| PRETERM                 | -0.538295956 | 0.06383069 | -8.88   | <.0001 |
| TERMS                   | 0.149955661 | 0.03661426 | 4.10    | <.0001 |
| TOTALP                  | -0.166673935 | 0.03451441 | -4.83   | <.0001 |
| YrsLastLiveBirth        | -0.005594258 | 0.00285376 | -1.96   | 0.0500 |
| drinker                 | -0.163389112 | 0.07007677 | -2.33   | 0.0197 |
| M_bdead                 | -0.577327101 | 0.36256303 | -1.59   | 0.1113 |
| M_children              | 0.883328150 | 0.42624567 | 2.07    | 0.0383 |
| M_terms                 | 1.225645880 | 0.40683560 | 3.01    | 0.0026 |
| M_totalp                | -1.396395899 | 0.41368468 | -3.38   | 0.0007 |
| M_YrsLastLiveBirth      | -0.116547142 | 0.01980040 | -5.89   | <.0001 |
| smoker                  | -0.124818184 | 0.03223567 | -3.87   | 0.0001 |
| MARITAL                 | -0.027463817 | 0.01855666 | -1.48   | 0.1389 |
| Tree_Race 6             | -0.087179833 | 0.11168668 | -0.78   | 0.4351 |
| Tree_Race 7             | 0.049543737 | 0.07514436 | 0.66    | 0.5097 |
| Tree_Race 8             | 0.013329560 | 0.07930082 | 0.17    | 0.8665 |
| Tree_Race 9             | 0.053123518 | 0.08561883 | 0.62    | 0.5350 |
| Tree_Race 10            | -0.209024226 | 0.07644620 | -2.73   | 0.0063 |
| Tree_Race 11            | -0.222655587 | 0.07842122 | -2.84   | 0.0045 |
| Tree_Race 99            | 0.000000000 | 0.00000000 | -    | -     |
PROC GLM Statistical Graphics Diagnostics

- ODS GRAPHICS ON and PLOTS=DIAGOSTICS.
Generalize Linear Predictive Modeling
The Gamma Distribution for Right-Skewed Targets: Charity Donation Data

- Donation amount, as is often the case for monetary or usage outcomes, is right skewed.
- An alternative to the Lognormal is the Gamma Distribution from the Exponential Family.

```sas
proc univariate data=dev;
var targetd;
histogram /gamma(alpha=est sigma=est theta=est color=blue w=2)
   midpoints=0 to 210 by 10;
run;
```
New Modeling Routines in SAS/STAT: Generalized Linear Models and Mixture Distributions using HPGENSELECT

- Fits Generalized Linear Models by specifying a distribution and link function to enable modeling of count data, rates, and non-normal continuous outcomes.

- Supports model selection routines
  - Backward, Forward, Stepwise selection using significance level
  - LASSO selection
  - Choice of final model using significance level, AIC, SBC

- Supports mixture distributions
  - Zero Inflated Poisson
  - Zero Inflated Negative Binomial
  - Tweedie
Mixture Models with HPGENSELECT

- The Tweedie distribution has been used extensively in insurance data modeling as it corresponds to the underlying loss generating process (e.g. total cost of claims). This mixed type distribution results from the mixing of two underlying components - the Poisson and Gamma distributions.

- In the Zero-Inflated Poisson Model, the population is considered to consist of two types of individuals. The first type gives Poisson distributed counts, which might contain zeros. The second type always gives a zero count. The ZIP model fits, simultaneously, two separate regression models. One is a logistic model that models the probability of being eligible for a non-zero count. The other models the size of that count.
A single development dataset is supplied to the procedure. The PARTITION statement requests a 67/33% split of this set.

A backward regression is run, eliminating terms based on statistical significance. The final model in the backward sequence is chosen using the Schwartz Bayesian Criterion.

A Gamma distribution is fit using a log link (though inverse is the canonical link).

```
proc hpgenselect data=dev;
  class &catvars;
  partition fraction(validate=.33 seed=12345);
  selection method=backward(slstay=.0001 choose=sbc);
  model targetd = &demvars &loggiftvars &cntvars
          &timevars &promvars &catvars/dist=gamma link=log;
run;
```
The model at step 15 minimizes SBC and is selected.

- Fit Statistics are given for Training and Validation data.
- Estimates and significance tests are provided.
Thank You
Lorne.Rothman@sas.com