



**THE  
POWER  
TO KNOW®**

# Predictive Modeling with SAS

---

Lorne Rothman, PhD, P.Stat.  
Principal Statistician  
[Lorne.Rothman@sas.com](mailto:Lorne.Rothman@sas.com)

# Purpose of Predictive Modeling

- ✓ To Predict the Future
- ✗ To identify statistically significant attributes or risk factors
- ✗ 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.
- ✗ To enable decision making and influence policy through publications and presentations

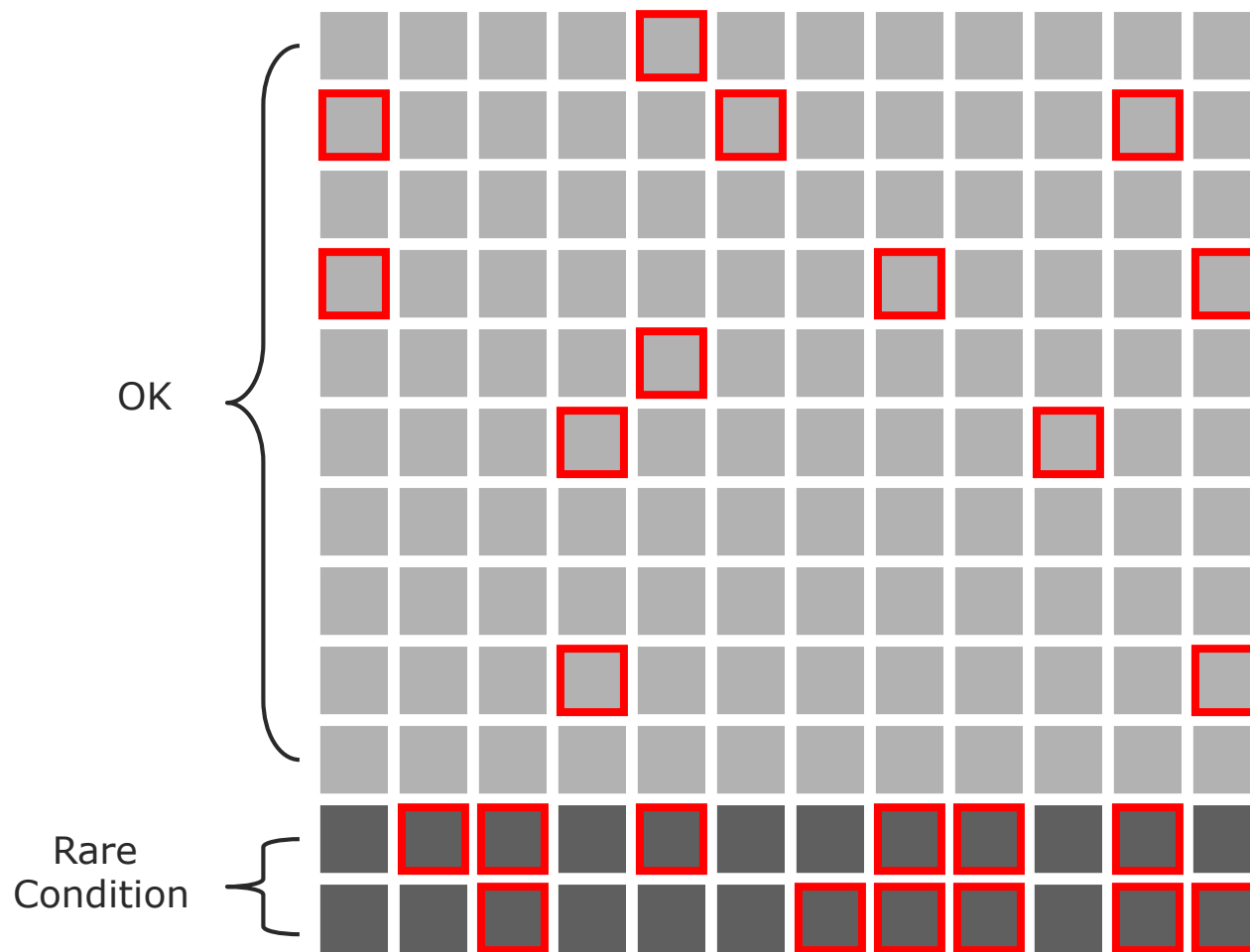
# Challenges: Opportunistic Data

	<u>Experimental</u>	<u>Opportunistic</u>
<b>Purpose</b>	Research	Operational
<b>Value</b>	Scientific	Commercial
<b>Generation</b>	Actively controlled	Passively observed
<b>Size</b>	Small	Massive
<b>Hygiene</b>	Clean	Dirty
<b>State</b>	Static	Dynamic

# Data Deluge

OHIP CIHI Population Census Canadian  
Cancer Registry Canadian Community Health  
Survey Canadian MIS Database Canadian  
Joint Replacement Registry Canadian Health  
Replacement Registry Canadian Organ  
Use Monitoring Registry Canadian Tobacco  
Database Disease Surveillance Canadian  
General Social Survey Discharge Abstract  
Activity Limitation Survey – Cycle 6 Health and  
Access Survey Hospital Mental Health and  
Hospital Morbidity Database Health Services  
Experiences Survey Natchem/Particulate  
Matter Database Natchem/Precipitation  
Database National Alcohol and Drug Survey  
National Pollutant Release Inventory  
(Household) National Health Survey  
Reporting System National Rehabilitation  
Survey on Aging and Independence Survey  
on Smoking in Canada TRAIID Vital  
Statistics

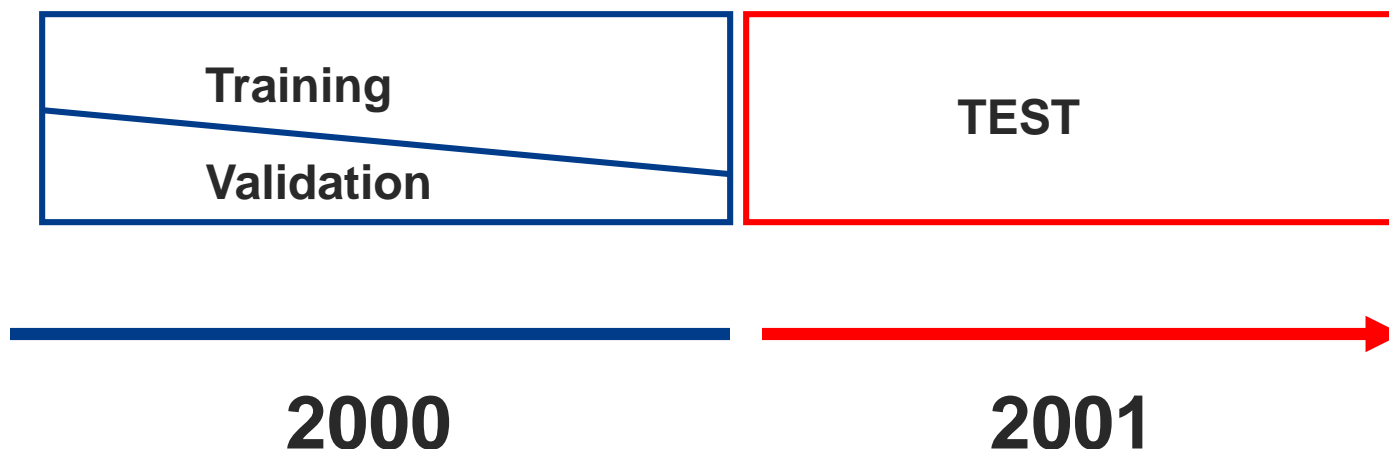
# Challenges: Rare Events



# Methodology: Empirical Validation

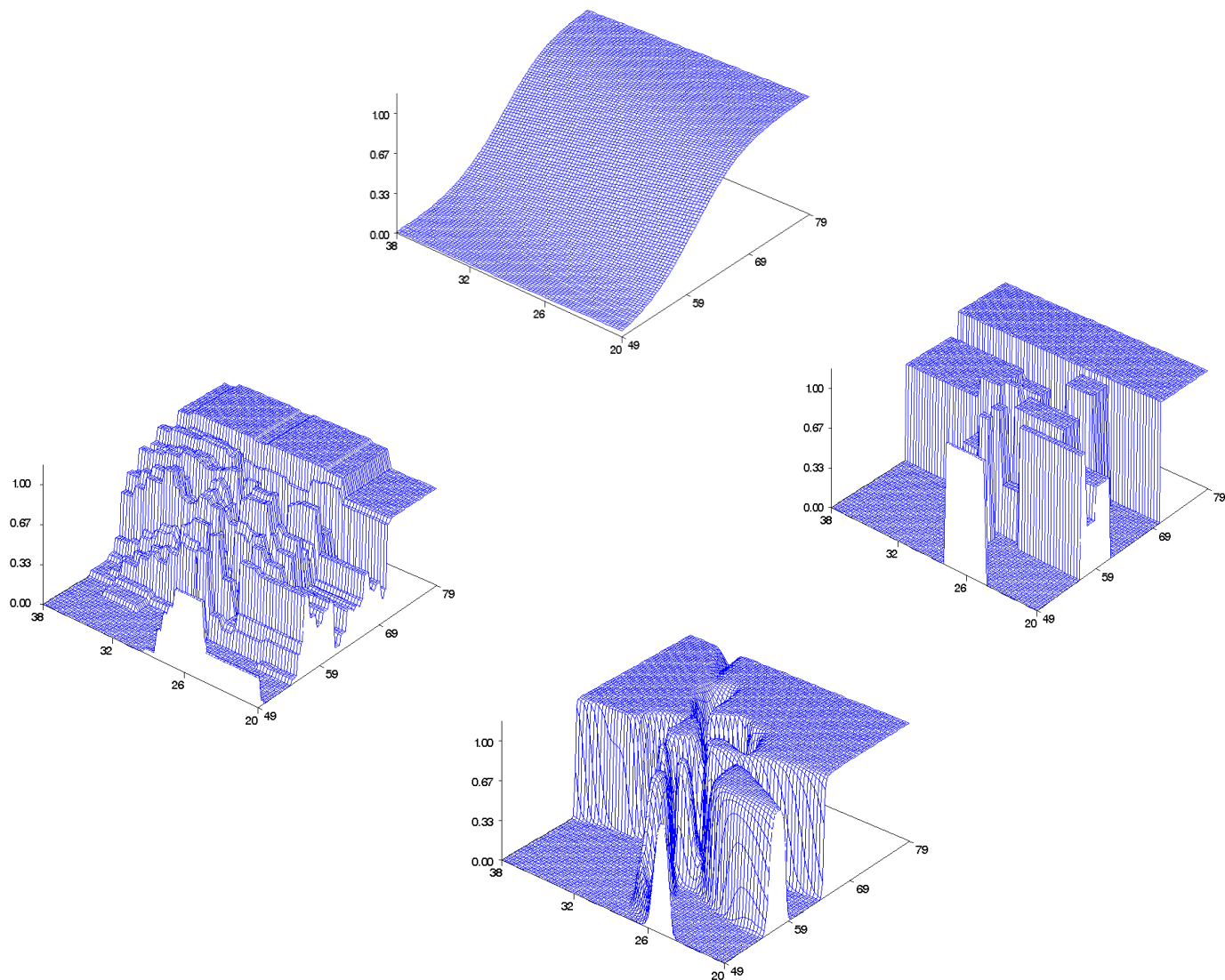


# Predicting the Future with Data Splitting



- ❖ 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.
- Data Scientist = (a) What statistician's call themselves when looking for a job. (b) A statistician from California.

# The North Carolina Birth Records Data

- North Carolina Birth Records from North Carolina Center for Health Statistics: 122,550 from 2000, and 120,300 from 2001.
- 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 pregnancy related data

•Medical history for this pregnancy

- Obstetric procedures
- Events of Labor
- Method of delivery
- New born characteristics

Time



**THE  
POWER  
TO KNOW®**

## Variable Preparation

# Key Features of Variable Preparation: Missing Indicators

```
data bwt00;
set bwt00;
array vars{*} fage mage feduc meduc totalp bdead terms loutcome
prenatal marital children racemom racedad 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_racemom M_racedad
M_cignum M_drinknum M_anemia M_cardiac M_aclung M_diabetes M_herpes
M_hydram 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;
```

- ❖ Create missing indicators to capture associations between missingness and the target.
- ❖ The process is repeated for Test data if present.

# Key Features of Variable Preparation: Imputation

```
proc stdize data=train reponly method=median
            out=train outstat=med;
    var _numeric_;
run;

proc stdize data=valid out=valid
            reponly method=in(med);
var _numeric_;
run;

proc stdize data=pm.test01 out=test
            reponly method=in(med);
var _numeric_;
run;
```

- ❖ STDIZE will do missing value replacement (REONLY) 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.

# Key Features of Variable Preparation: Dimension Reduction

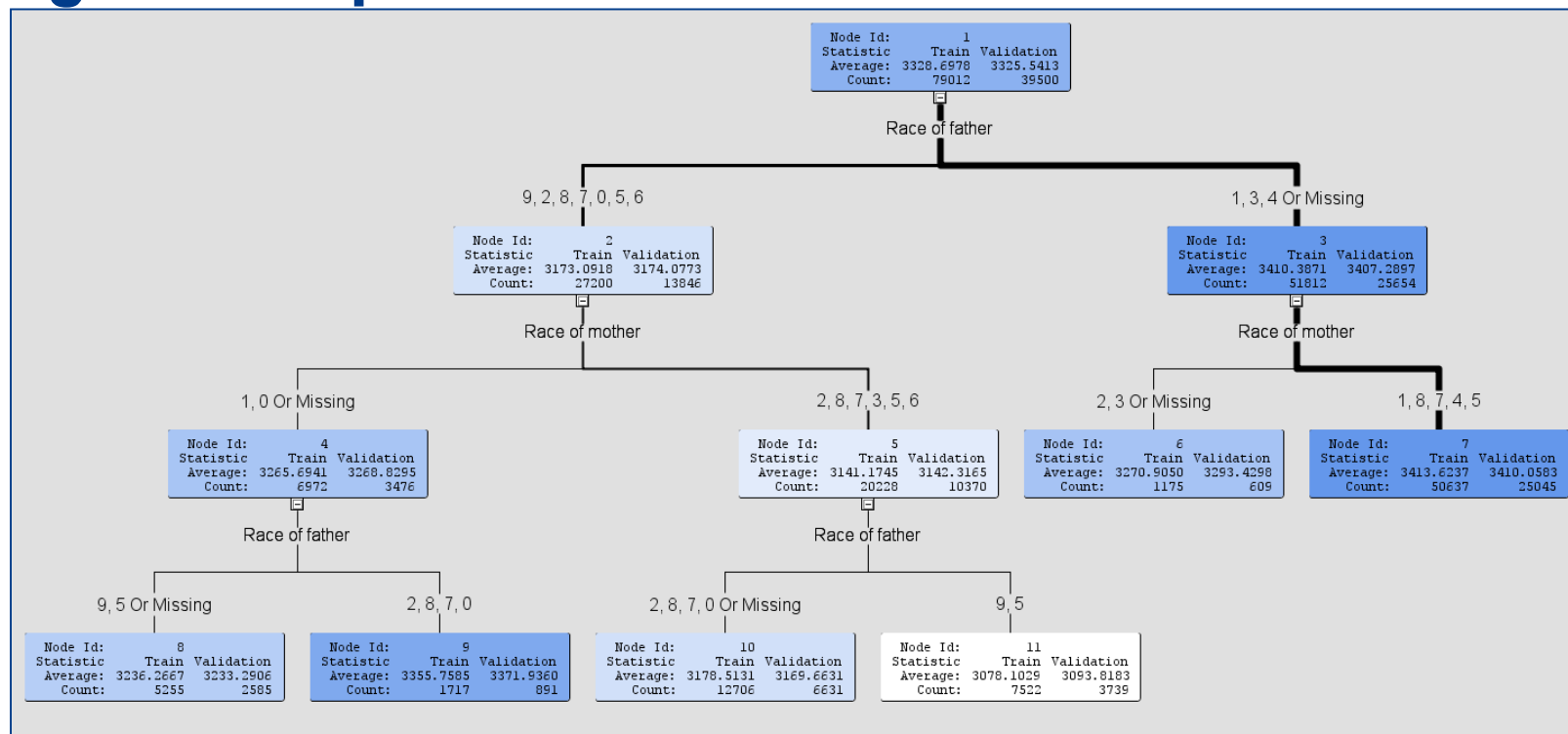
Cluster 2	CHILDREN	0.8025	0.0916	0.2174	Number of children now living
	LOUTCOME	0.7907	0.0801	0.2275	Outcome of last delivery
	TOTALP	0.7398	0.4152	0.4449	Total pregnancies (including this one)
	M_YrsLastLiveBirth	0.8245	0.0782	0.1904	
Cluster 3	FEDUC	0.8554	0.1557	0.1713	Education of father (years)
	MEDUC	0.8554	0.1737	0.1751	Education of mother (years)
Cluster 4	M_terms	0.9329	0.0830	0.0732	
	M_totalp	0.9329	0.2496	0.0894	
Cluster 5	M_cignum	0.9537	0.0707	0.0499	
	M_drinknum	0.9537	0.0859	0.0507	
Cluster 6	M_fage	0.9257	0.0381	0.0772	
	M_feduc	0.9244	0.0379	0.0786	
	MARITAL	0.5320	0.1420	0.5455	Marital status
Cluster 7	CIGNUM	0.8454	0.2511	0.2065	Average # of cigarettes daily
	smoker	0.8454	0.0268	0.1589	
Cluster 8	DRINKNUM	0.6777	0.0007	0.3225	Average # of alcoholic drinks per week
	drinker	0.6777	0.2236	0.4151	

- ❖ 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;
```



# Key Features of Variable Preparation: Collapsing Categorical Inputs



- ❖ Variables RACEMOM and RACEDAD contain 9 and 10 levels respectively.
- ❖ Use a Decision Tree model to optimally collapse the hundreds of possible combinations (over 50 parameters in a regression) to a single 6-level variable using training data.



**THE  
POWER  
TO KNOW®**

## Binary Target

---

# Oversampling

```
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;
```

lbwt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8260	49.63	8260	49.63
1	8383	50.37	16643	100.00

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

# Empirical Validation via Data Partitioning

```
proc surveyselect
    data=bwt00s
    samprate=.67
    out=lbwt.develop
    seed=4444
    outall;

run;

data lbwt.train lbwt.valid;
    set lbwt.develop;
    if selected then output lbwt.train;
    else output lbwt.valid;
run;
```

- ❖ 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.
- ❖ Stratification on the target, LBWT ensures equal representation of low birth weight cases in training and validation sets.

# Model Selection and Scoring

```

title "Early Warning";
proc logistic data=train;
  class tree_race(param=ref ref='8');
  model lbwt(event='1')=&numvars2 tree_race/ selection = backward slstay=.01;
  score data=valid out=sco_validate(rename=(p_1=p_early)) priorevent=.072;
  score data=test out=sco_test(rename=(p_1=p_early)) priorevent=.072;
run;

title "All Inputs";
proc logistic data=train;
  class tree_race(param=ref ref='8');
  model lbwt(event='1')=&numvars tree_race/selection = backward slstay=.01;
  score data=sco_validate out=sco_validate(rename=(p_1=p_all)) priorevent=.072;
  score data=sco_test out=sco_test(rename=(p_1=p_all)) priorevent=.072;
run;

title "All Inputs with Interactions";
proc logistic data=train;
  class tree_race(param=ref ref='8');
  model lbwt(event='1')= HYPERCH HYPERPR CERVIX BDEAD CIGNUM ECLAMP HEMOGLOB HYDRAM MEDU
    PINFANT PRENATAL PRETERM RHSEN TOTALP UTERINE drinker M_YrsLastLiveBirth
    M_smoker MARITAL smoker Tree_Race HYPERCH|HYPERPR|CERVIX|BDEAD|CIGNUM|ECLAMP|
    HEMOGLOB|HYDRAM MEDUC|PINFANT|PRENATAL|PRETERM|RHSEN|TOTALP|UTERINE|drinker|M_YrsLast
    M_smoker|MARITAL|smoker|Tree_Race @2/ selection = forward slentry=.01 include=21 ;
  score data=sco_validate out=sco_validate(rename=(p_1=p_AllInt)) priorevent=.072;
  score data=sco_test out=sco_test(rename=(p_1=p_AllInt)) priorevent=.072;
run;

```

- ❖ The SCORE statements allows for scoring of new data (Validation and Test) and adjusts oversampled data back to the population prior (PRIOREVENT=0.072).
- ❖ The same dataset is re-scored (sco\_validate, sco\_test) so that predictions for all three models are in the same set for comparisons.

# Model Selection and Scoring

- Predictive models tend to have more parameters than theoretically driven explanatory models.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1363	0.1695	0.6469	0.4212
HYPERCH		1	0.9184	0.1800	26.0366	<.0001
BDEAD		1	0.3607	0.1215	8.8093	0.0030
CIGNUM		1	0.0133	0.00438	9.1896	0.0024
FEDUC		1	-0.0471	0.00913	26.5769	<.0001
MAGE		1	0.0119	0.00399	8.8973	0.0029
PINFANT		1	-2.4707	0.2474	99.7645	<.0001
PRENATAL		1	-0.1077	0.0151	50.9375	<.0001
PRETERM		1	1.7561	0.1672	110.2543	<.0001
TOTALP		1	0.0653	0.0184	12.6044	0.0004
drinker		1	0.5215	0.1980	6.9387	0.0084
M_YrsLastLiveBirth		1	0.6452	0.0521	153.2474	<.0001
MARITAL	1	1	-0.1093	0.0276	15.7155	<.0001
smoker		1	0.6233	0.0717	75.5892	<.0001
Tree_Race	6	1	0.0777	0.2608	0.0888	0.7657
Tree_Race	7	1	-0.2579	0.0843	9.3686	0.0022
Tree_Race	9	1	-0.2259	0.1597	2.0002	0.1573
Tree_Race	10	1	0.5841	0.0885	43.5252	<.0001
Tree_Race	11	1	0.6674	0.0936	50.7890	<.0001
Tree_Race	99	1	0.2783	0.2040	1.8609	0.1725

# Model Assessments for Binary Targets

		1	Predicted**	0	
Actual	1	TP		FN	AP
	0	FP		TN	AN
		PP		PN	n

**Accuracy =**  
 $(TP+TN)/n$

**Sensitivity =**  
 $TP/AP$

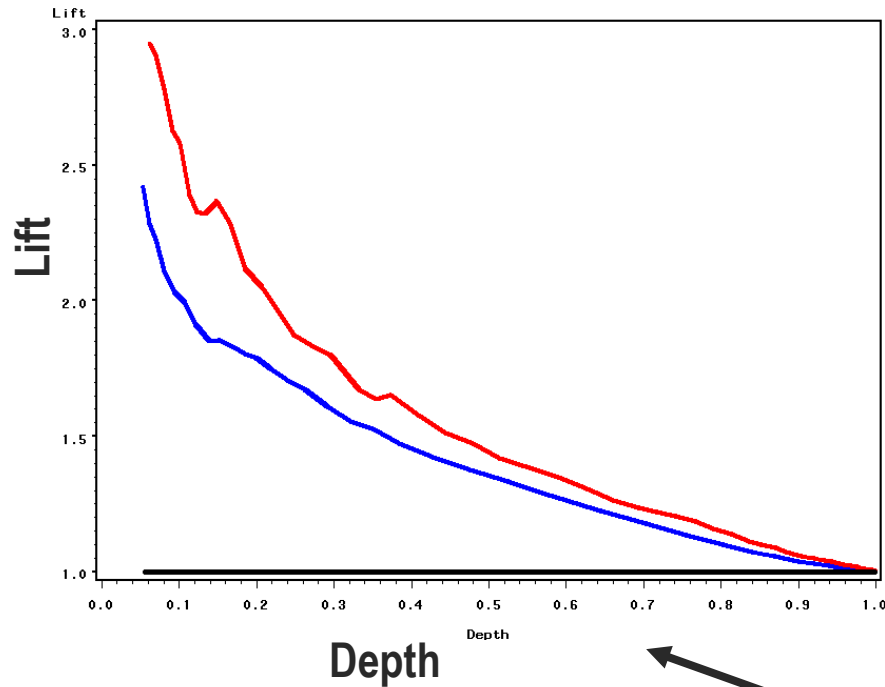
**Specificity =**  
 $TN/AN$

**Lift =**  
 $(TP/PP)/\pi_1$

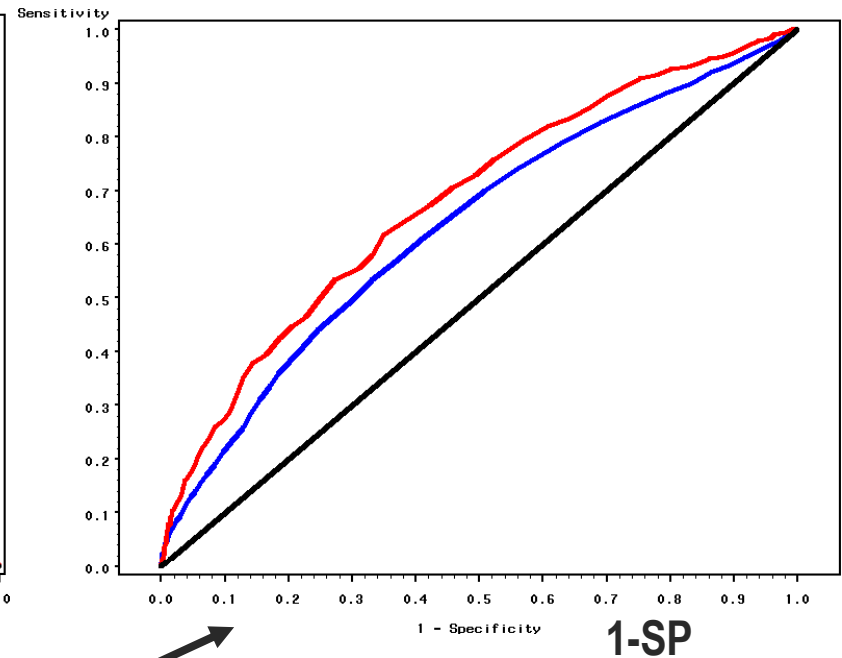
**\*\* - Where Predicted 1=(Pred Prob > Cutoff)**

# Assessment Charts for Binary Targets

Lift Charts



ROC Charts



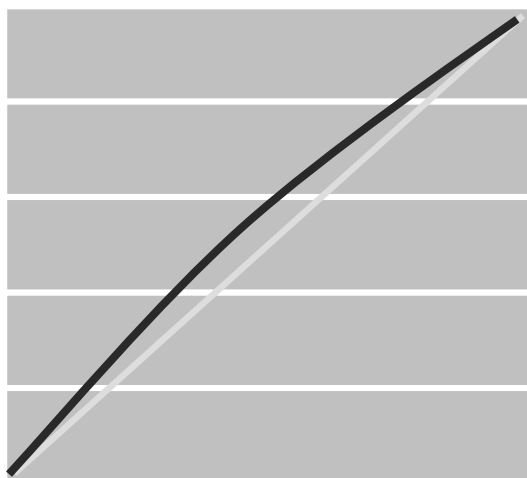
SE

Explore measures across a range of cutoffs

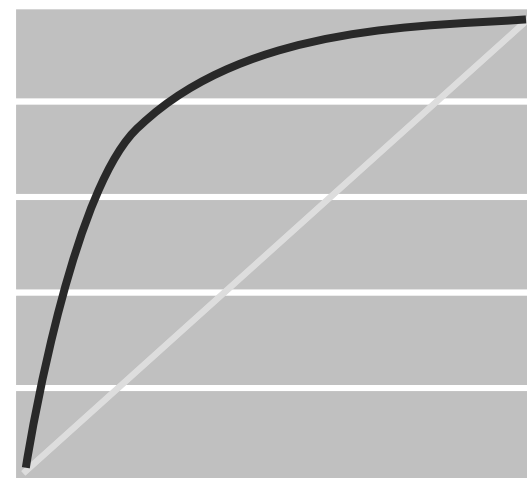
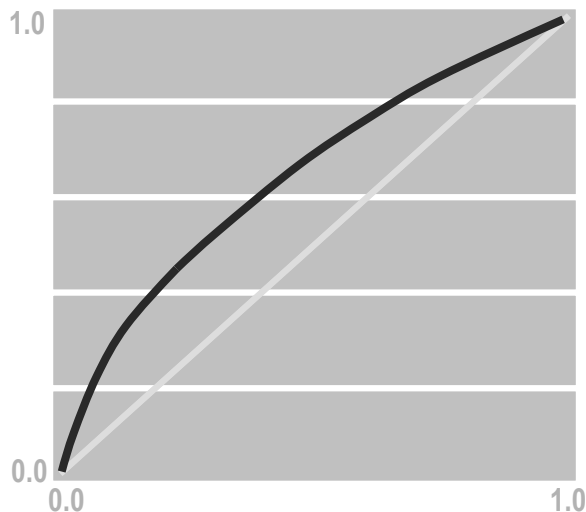
TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
FP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP	TN



# Receiver Operator Curves



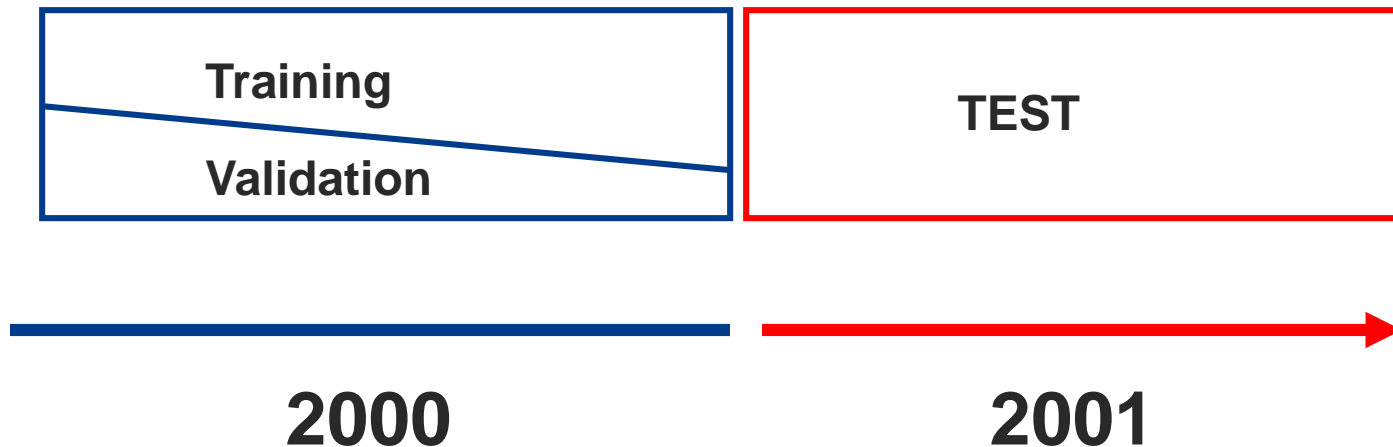
**weak model**



**strong model**

- ❖ A measure of a model's predictive performance, or model's ability to discriminate between target class levels. Areas under the curve range from 0.5 to 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).

# Predicting the Future with Data Splitting



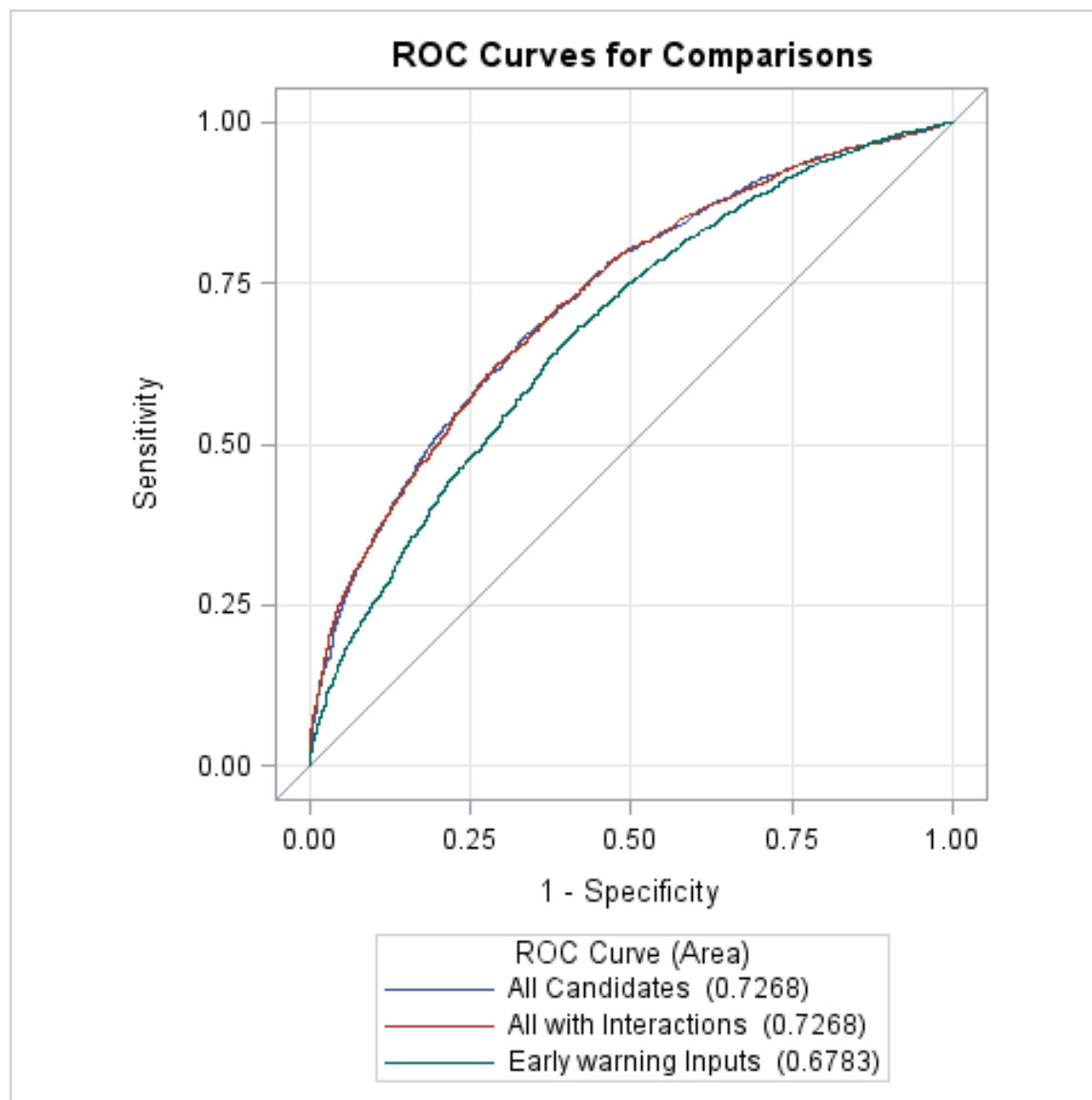
- ❖ Models are fit to Training Data, compared and selected on Validation and tested on a future Test set.

# Model Assessment: ROC Curves

```
ods html;  
ods graphics on;  
proc logistic data=sco_validate;  
  model lbwt(event='1')=p_all p_allint p_early / nofit;  
  roc "All Candidates" p_all;  
  roc "All with Interactions" p_allint;  
  roc "Early warning Inputs" p_early;  
  rocncontrast "Comparing the Three Models: Validation Data "/estimate=allpairs;  
run;
```

- ❖ The dataset with all three predictions (Sco\_validate) is supplied to PROC LOGISTIC.
- ❖ The ROCCONTRAST statements provides statistical significance 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 process can be repeated with the Test set.

# Comparing ROC Curves on Validation Data



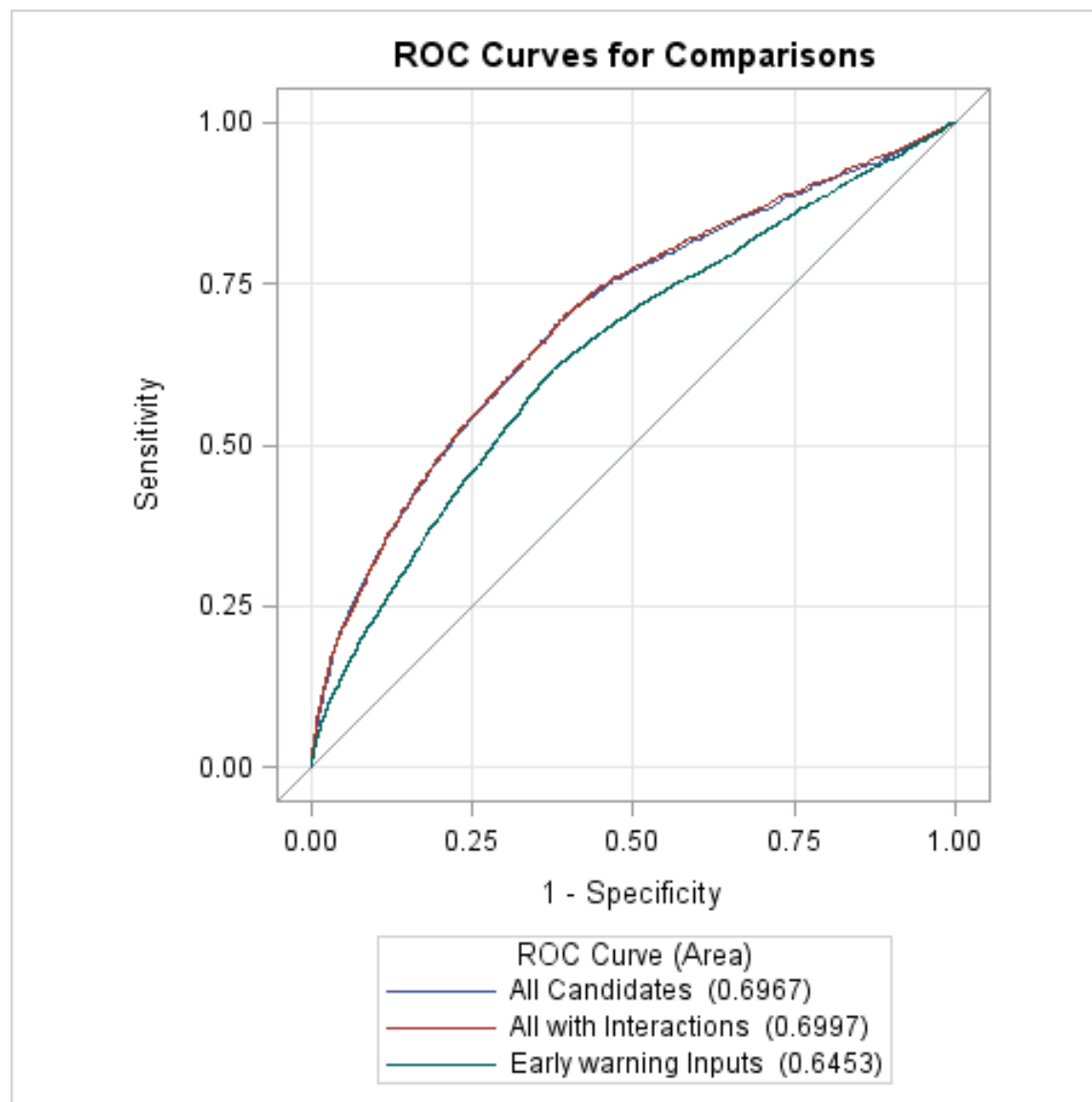
# Comparing ROC curves on Validation Data

ROC Association Statistics							
ROC Model	Mann-Whitney				Somers' D (Gini)	Gamma	Tau-a
	Area	Standard Error	95% Wald Confidence Limits				
All Candidates	0.7268	0.00670	0.7137	0.7400	0.4536	0.4544	0.2268
All with Interactions	0.7268	0.00670	0.7137	0.7399	0.4535	0.4543	0.2268
Early warning Inputs	0.6783	0.00711	0.6644	0.6923	0.3566	0.3567	0.1783

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Comparing the Three Models: Validation Data	2	126.6817	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
All Candidates - All with Interactions	0.000045	0.00170	-0.00329	0.00339	0.0007	0.9788
All Candidates - Early warning Inputs	0.0485	0.00431	0.0400	0.0569	126.5842	<.0001
All with Interactions - Early warning Inputs	0.0485	0.00459	0.0395	0.0574	111.5635	<.0001

# Comparing ROC curves on Test Data

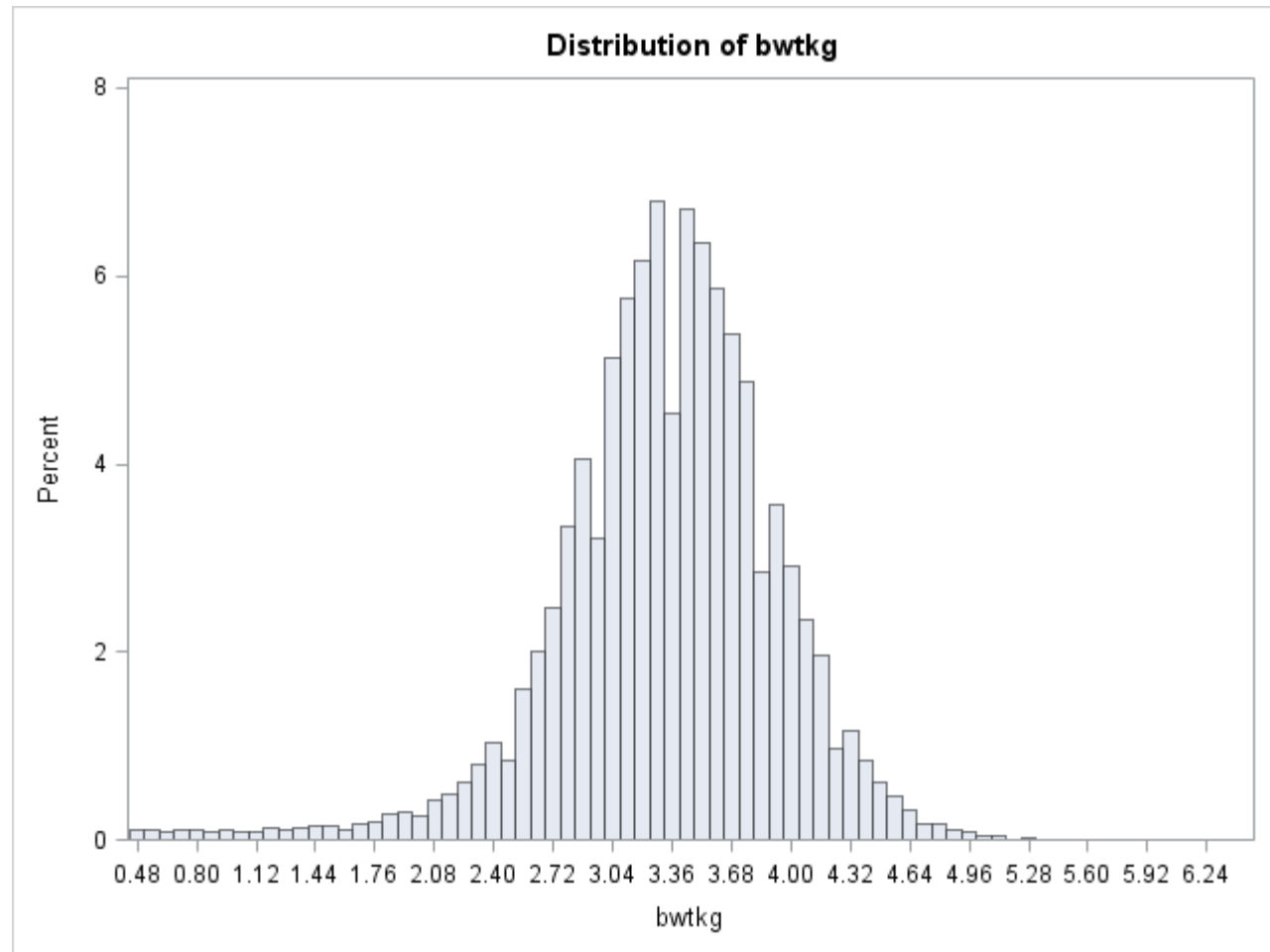




**THE  
POWER  
TO KNOW®**

## Interval Target

# Interval Target: Birth Weight





# PROC GLMSELECT!

- ❖ GLMSELECT fits interval target models 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

LARS works as follows:

- Standardize inputs and outcome. Start with no variables in your model.
- Find the variable  $X$  most correlated with the residual. (*Note that the variable most correlated with the residual is equivalently the one that makes the least angle with the residual, whence the name.*)
- Move in the direction of this variable until some other variable  $X_2$  is just as correlated.
- At this point, start moving in a direction such that the residual stays equally correlated with  $X$  and  $X_2$  (i.e., so that the residual makes equal angles with both variables), and keep moving until some variable  $X_3$  becomes equally correlated with our residual.
- And so on, stopping when we've decided our model is big enough.
- How big is enough? Complexity of the model can be optimized on validation data to minimize validation ASE.

# Least Absolute Shrinkage and Selection Operator (LASSO)

- A constrained form of ordinary least squares is used:
  - The sum of the absolute values of the regression coefficients must be smaller than a certain value.
- The LASSO coefficients  $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_m)$  are the solution to:

$$\text{Minimize } \|y - X\beta\|^2$$

$$\text{subject to } \sum_{j=1}^m |\beta_j| \leq t$$

# Model Selection Using GLMSELECT: Backward

```
proc glmselect data=lbwt.train valdata=lbwt.valid /*testdata=test01*/
    plots(stepAxis=number)=ASEPlot;
    class tree_race marital;
    model bwtkg = &numvars tree_race
        /selection=backward(choose = validate select = sl slstay=0.00001);
run;
```

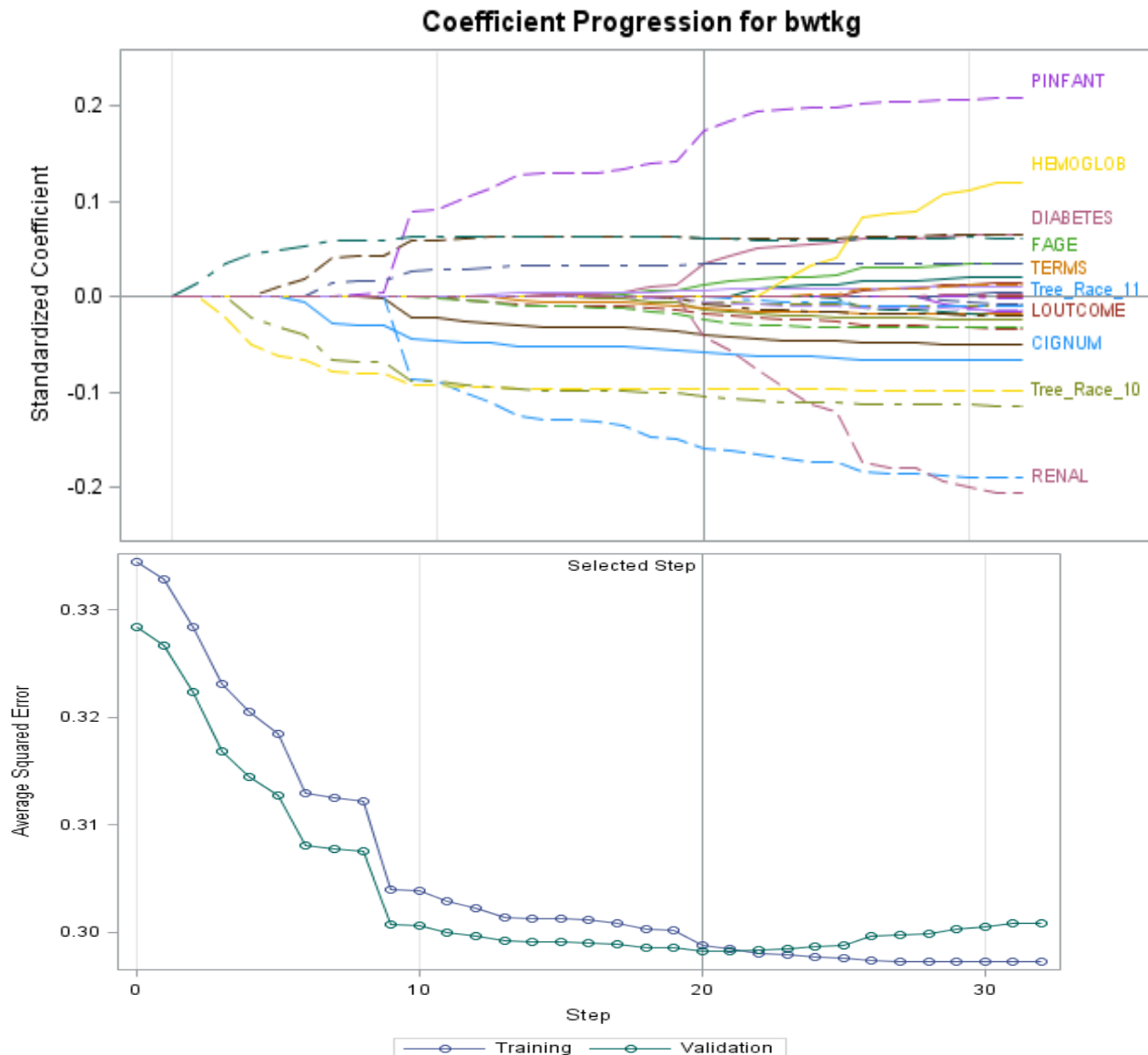
```
proc glmselect data=lbwt.train valdata=lbwt.valid /*testdata=test01*/
    plots=all;
    class tree_race marital;
    model bwtkg = &numvars tree_race marital
        /selection=lar(choose = validate stop=none);
run;
```

- ❖ SELECTION=LAR requests Least Angle Regression.
- ❖ 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.
- ❖ 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.

# Backward Model Tuning using Validation ASE



# LAR Model Tuning using Validation ASE



# Final Model Fitting and Score Code in GLM

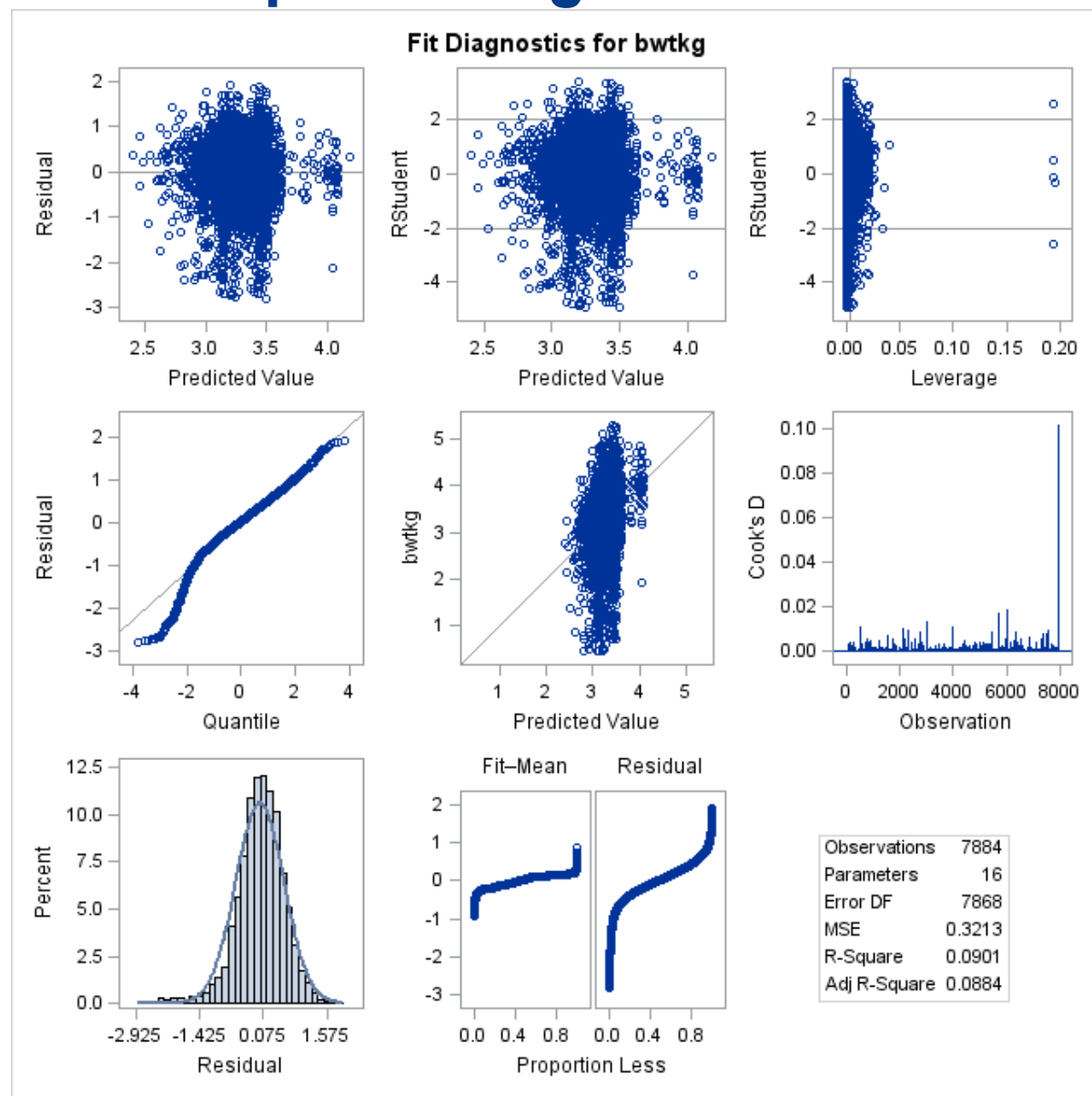
```
ods html;
ods graphics on;
proc glm data=lbwt.train plots(maxpoints=500000)=diagnostics;
  class marital tree_race;
  model bwtkg= BDEAD CIGNUM DIABETES MEDUC PINFANT PRETERM RENAL smoker
    MARITAL Tree_Race/solution;
  code file = 'C:\DATA\EDU\TALKS\HUG2015\bwtmod.sas';
run;
quit;

data scored;
  set bwt00;
  %include 'C:\DATA\EDU\TALKS\HUG2015\bwtmod.sas'/source2;
run;
```

- ❖ GLMSELECT does not provide hypothesis test results and model diagnostics.
- ❖ The model selected by GLMSELECT can be refit in PROC GLM.
- ❖ PLOTS=DIAGNOSTICS requests diagnostic plots.
- ❖ The new CODE statement requests score code that can be applied to a new set with the %INCLUDE statement. SOURCE2 prints the scoring action to the log.
- ❖ The following procedures support a CODE statement as of V12.1: GENMOD, GLIMMIX, GLM, GLMSELECT, LOGISTIC, MIXED, PLM, and REG.

# PROC GLM Statistical Graphics Diagnostics

- ❖ ODS GRAPHICS ON and PLOTS=DIANGOSTICS.





# PROC GLM Statistical Graphics Diagnostics

- ❖ Partial Score Code generated by Code statement, and resulting predictions or scores.

```

1302 **** Compute Linear Predictors;
1303 +drop _LP0;
1304 +_LP0 = 0;
1305 +
1306 +_LP0 = _LP0 + (-0.13315183718733) * BDEAD;
1307 +_LP0 = _LP0 + (-0.00869051351195) * CIGNUM;
1308 +_LP0 = _LP0 + (0.05014815893447) * DIABETES;
1309 +_LP0 = _LP0 + (0.01301679390255) * MEDUC;
1310 +_LP0 = _LP0 + (0.50368835463479) * PINFANT;
1311 +_LP0 = _LP0 + (-0.52153951669412) * PRETERM;
1312 +_LP0 = _LP0 + (-0.11160459368849) * RENAL;
1313 +_LP0 = _LP0 + (-0.12459102060591) * smoker;
1314 **** Effect: MARITAL;
1315 +_TEMP = 1;
1316 +_LP0 = _LP0 + (0.05728314885559) * _TEMP * _0_0;
1317 +_LP0 = _LP0 + (0) * _TEMP * _0_1;
1318 **** Effect: Tree_Race;
1319 +_TEMP = 1;

```

Obs	bwtkg	P_bwtkg
1	3.74213	3.31806
2	3.57204	3.30109
3	3.43029	3.33493
4	3.17514	3.20153
5	3.85553	3.10129
6	3.79883	3.48333
7	3.14679	3.42741



# Thank You

---

**THE  
POWER  
TO KNOW®**