Predicting Low Strength Properties of Wood Composite Panels using Bayesian Logistic Regression

Xia Huang

Research Associate
Center for Renewable Carbon,
University of Tennessee
xhuang8@utk.edu

Yan Zeng

Former Graduate Research Assistant
Department of Statistics, Operations, Mgmt. Science,
University of Tennessee

Timothy M. Young

Professor
Center for Renewable Carbon,
University of Tennessee
tmyoung1@utk.edu

Frank M. Guess

Professor
Department of Statistics, Operations, Mgmt. Science,
University of Tennessee
fguess@utk.edu

Center for Renewable Carbon
University of Tennessee Institute of Agriculture

THE UNIVERSITY of TENNESSEE KNOXVILLE

Outline

- ***Background***
- ***Data***
- ***Missing Data Imputation***
  - *Imputation Methods*
  - *Imputation Comparison*
- ***Predictive Model***
  - *Variable Explanation*
  - *Bayesian Logistic Regression*
  - *Results*
- ***Summary***

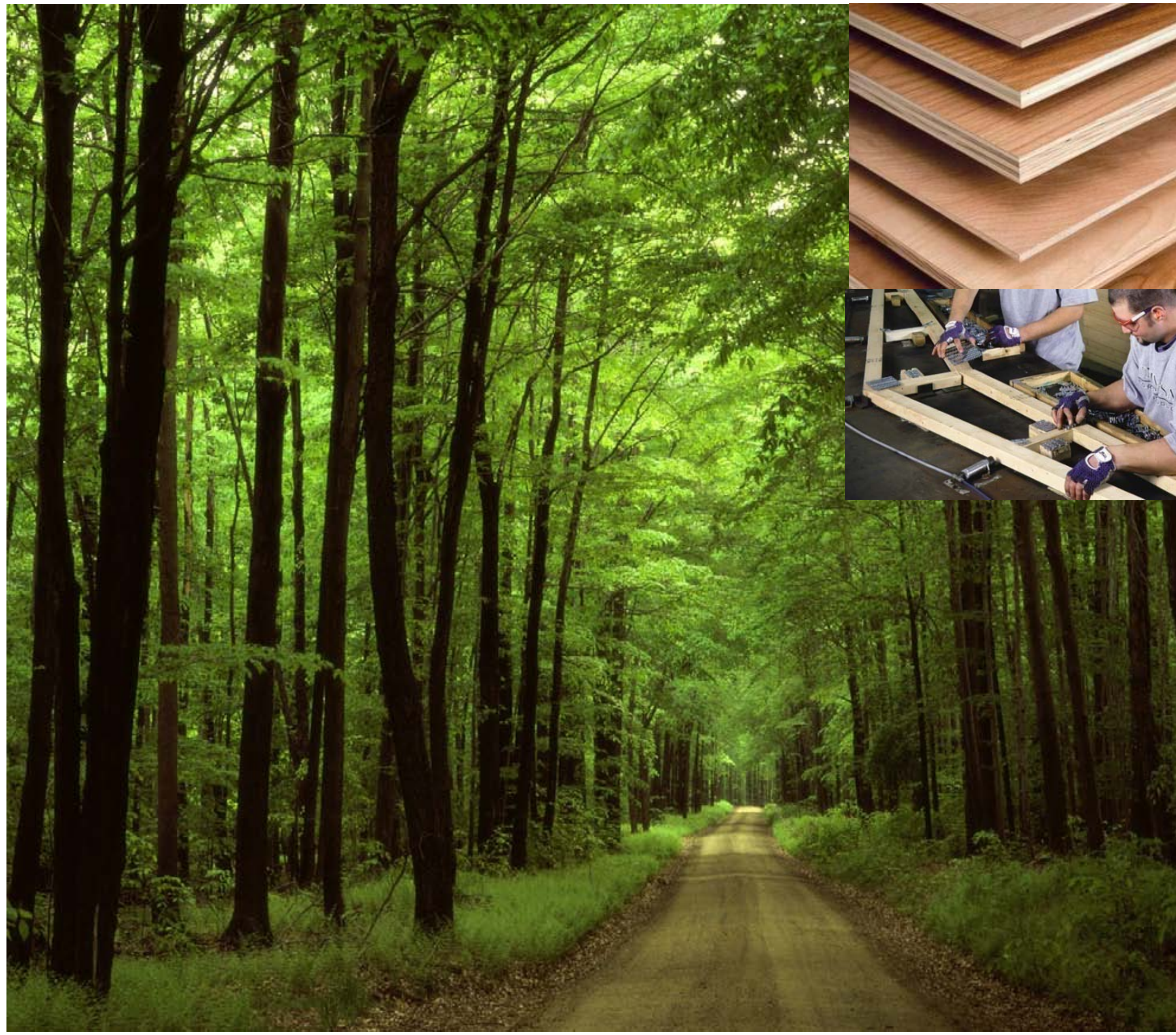
Background

Forest Product Industry

Background

- 5% of the total U.S. manufacturing gross domestic product (GDP)
- $175 billion a year in sales
- Approximately 900,000 employees, earning $50 billion in annual payroll
- Top 10 manufacturing sector employers in 48 states

(American Forest and Paper Association,2012)



Research Goal & Objectives

Study Background

- *The study was performed for a large-capacity wood composite panel manufacturing factory in the southeastern U.S.*
  - *Explore the prediction of Modulus of Rupture (MOR) and Internal Bond (IB) as a remedy to maintain product specification and minimize costs*
  - *Information loss due to sensor malfunction or data "send/retrieval" problems*

Research Goal

- *Predicting low strength properties (MOR and IB) of wood composite panels*

Objectives

- *Focus on data quality and consistency in the use of imputation methods.*
- *Identify predictors influencing low strength properties*
- *Develop predictive model*

Data

Raw Data

Predictors

- *Sensor-collected process data*
- *237 predictor variables in different units (i.e., fiber moisture, line speed, mat temperature, press pressure, etc.)*
- Collected roughly by time order at different time intervals

Responses

- *Obtained through destructive tests*
- *Modulus of Rupture (MOR) and Internal Bond (IB), both measured in kPa (kilopascal)*
- *Average MOR and IB*

Data Structure

Data

Ranges from 3,447 to 14,926 kPa

	MOR	V1	V2	V3	V4	V5	V6	…	V237	
1	X	X	X		X	X	X	…		
2	X	X			X			…	X	
3	X	X			X			…		
4	X	X				X		…		
5	X	X	X	X	X	X	X	…	X	Value
6	X	X			X			…	X	
7	X				X	X		…	X	
8	X	X	X	X	X	X	X	…	X	Observation
9	X	X					X	…	X	
…	…	…	…	…	…	…	…	…	…	
4522	X	X			X	X		…	X	

Problems

- Missing values are in ***random pattern***
- Statistical packages such as JMP, SAS, R would remove observations with even one missing value when building prediction models, which causes great ***information loss***

Data Structure

Data

Ranges from 69 to 1,750 kPa

	IB	V1	V2	V3	V4	V5	V6	…	V237	
1	X	X	X		X	X	X	…		
2	X	X			X			…	X	
3	X	X			X			…		
4	X	X				X		…		
5	X	X	X	X	X	X	X	…	X	Value
6	X	X			X			…	X	
7	X				X	X		…	X	
8	X	X	X	X	X	X	X	…	X	Observation
9	X	X					X	…	X	
…	…	…	…	…	…	…	…	…	…	
4522	X	X			X	X		…	X	

Problems

- Missing values are in ***random pattern***
- Statistical packages such as JMP, SAS, R would remove observations with even one missing value when building prediction models, which causes great ***information loss***

Summary of Missing Values

Data

Predictors

1. ***Every*** predictor variable has missing value, ranging from 2.4% to 81%
2. ***14*** predictor variables had more than 20% of data missing

Observations

1. ***Every*** observation has missing fields, ranging from less than 0.5% to 90%
2. ***Only Six observations*** have a missing rate above 20%

Responses

11 observations with response variable missing

Pre-screened Data

Data

Observations with no response are removed

Predictor variables and observations with more than 20% missing rate are excluded

	Response	V1	V2	V3	…	V222
1	X	X	X		…	
2	X	X			…	X
3	X	X	X	X	…	X
4	X	X			…	X
5	X	X	X	X	…	X
6	X	X			…	X
…	…	…	…	…	…	…
4411	X	X			…	X

3,647 observations with at least two or more fields missing

Collinearity

- A routine step of data quality assessment
- Correlation matrix and variation-inflation factors (VIF)
  - suggest some highly correlated predictors in the pre-screened data set
- Would affect later selection of statistical/modeling methods

Standardization

$$\frac{x_i - \bar{x}}{\hat{s}}$$

where $\bar{x}$ is the average of non-missing values, $\hat{s}$ is the standard deviation of non-missing value

Missing Data Imputation

Variable Selection

Why?

Due to the constraint of computation resources required by iterated computation when imputing missing values, i.e., statistical packages SAS® or R ® can become slow (EM)or may not converge (MCMC) on imputation results

Reduce the calibration model training time

Improve prediction performance for highly correlated data

LASSO

(least absolute shrinkage and selection operator)

Proposed by Tibshirani (1996)

A constrained version of the ordinary least squares (OLS) estimator, to achieve ***shrinkage*** and ***variable selection*** simultaneously

Sacrifice little variance for ***less bias*** in estimators

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1)$$

or

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad (2)$$

$$Subject\ to \sum_{j=1}^{p} |\beta_j| \le t$$

Variable Selection Results

(non-imputed datasets)

Imputation

	MOR	V1	V2	V3	…	V107
1	X	X	X		…	
2	X	X			…	X
3	X	X	X	X	…	X
4	X	X			…	X
5	X	X	X	X	…	X
6	X	X			…	X
…	…	…	…	…	…	…
4411	X	X			…	X

***1,073* complete** observations

Missing Pattern

Arbitrary

MAR
(Missing at Random)

***1,194* complete** observations

	IB	V1	V2	V3	…	V86
1	X	X	X		…	
2	X	X			…	X
3	X	X	X	X	…	X
4	X	X			…	X
5	X	X	X	X	…	X
6	X	X			…	X
…	…	…	…	…	…	…
4411	X	X			…	X

Imputation Methods

Imputation

Substitution | Mean/Median

Replace missing fields with the mean/median of the same predictor variable

"Hot-Deck" Method | The simple random imputation method

Replace the missing value with a randomly selected value from in the same predictor variable

LOCF | Last Observation Carried Forward

Replace the missing value with the last known value (observation) of the variable in a time-ordered data set

EM algorithm | Expectation-Maximization algorithm

"Expectation" Step - Given the observed data (including response variables), use available mean vector and covariance matrix for a multivariate normal distribution to calculate the conditional expectation of the complete-data log-likelihood
"Maximization" Step – Maximize above log-likelihood multivariate normal distribution to calculate the conditional, updating mean and covariance matrix
Use updated parameters to "impute" data, update mean and covariance matrix, iterate above two steps until convergence

MI procedure | Multiple Imputation using Markov Chain Monte Carlo (MCMC)

Replace each missing value with a set of plausible values that represent the uncertainty of the correct value
MCMC - combined with Bayesian inference of prior information to stimulate posterior distribution
Samples (estimated values for missing fields) - drawn from posterior distribution
M "complete datasets - Iterate above process for M times (e.g., 3 to 5 times)
A single point estimate – Average the values across M complete datasets

Method Comparison

Ten-fold Cross Validation

1. Partition ***all complete observations*** in non-imputed data as ***a matrix*** into ten subsets

2. Retain one subset as ***validation set*** and intentionally ***remove as missing***

3. Use rest of available data in ***non-imputed dataset*** to impute all missing fields ***including earlier removed part***

4. Compare imputed results with validation data, calculating Root Mean Square of Error (RMSE)

5. Repeat above process for ***each of ten subsets***

Use two-fold as an example		
X	X	X
X		X
X	X	
X	X	X
X		
X	X	X
X	X	X

Use two-fold as an example		
X	X	X
X		X
X	X	
X	X	X
X		
X	X	X
X	X	X

Use two-fold as an example		
X		X
X		X
X	X	
	X	
X		
X	X	
		X

Use two-fold as an example		
X	X_F	X
X	X_F	X
X	X	X_F
X_F	X	X_F
X	X_F	X_F
X	X	X_F
X_F	X_F	X_F
X_F	X_F	X

Results (MOR)

RMSEs from Imputations for ***Standardized*** Dataset with ***MOR*** as Response

RMSE	Mean Substitution	Median Substitution	Single Random Imputation	LOCF	EM	MI - MCMC
1	1.92	0.17	1.87	2.14	0.14	0.09
2	4.54	2.28	5.01	1.84	0.70	0.37
3	4.43	1.92	2.66	1.41	0.92	0.59
4	3.47	1.39	3.14	0.96	0.07	0.26
5	2.16	0.27	2.52	0.54	0.12	0.07
6	2.01	0.40	2.60	0.93	0.24	0.48
7	2.18	0.74	0.86	0.84	0.27	0.25
8	4.08	1.58	3.04	2.54	0.87	0.86
9	3.63	1.58	5.12	1.48	0.19	0.28
10	5.23	2.87	2.62	1.83	0.79	0.84
Average	3.37	1.32	2.94	1.45	0.43	0.41

Results (IB)

Imputation

***RMSE*s from Imputations for *Standardized* Dataset with *IB* as Response**

RMSE	Mean Substitution	Median Substitution	Single Random Imputation	LOCF	EM	MI - MCMC
1	3.08	0.77	4.92	0.08	0.33	0.33
2	4.55	1.15	1.44	0.92	0.65	0.79
3	2.48	1.46	0.57	1.70	0.27	0.25
4	4.16	1.08	2.84	2.31	0.98	0.90
5	3.92	0.34	2.61	1.43	0.12	1.41
6	2.47	1.09	2.55	2.06	0.10	0.00
7	2.24	1.63	1.99	0.74	0.11	0.15
8	2.26	0.79	1.48	0.75	0.43	0.66
9	2.96	0.51	1.46	1.79	0.64	0.45
10	3.49	0.05	5.04	0.05	0.59	0.37
Average	3.16	0.89	2.49	1.18	0.42	0.53

Results Summary

- **EM** and **MI-MCMC** *achieve better results*
- **No** *apparent differences between EM and MI-MCMC*
- **EM** *a bit faster than MI-MCMC*
  (Computation time for both is tolerable,20-30-min CPU time each. EM is 10% to 20% faster)
- **EM** *does better job for pre-screened data without variable-selection*
  (MI-MCMC wouldn’t converge when imputing the pre-screened data without variable selection)
- *Choose* **EM** *for imputation*
  - *Final Data with* **222** *predictor variables and* **4,411** *complete observations*

Predictive Model

Descriptive Statistics for Responses
(MOR and IB)

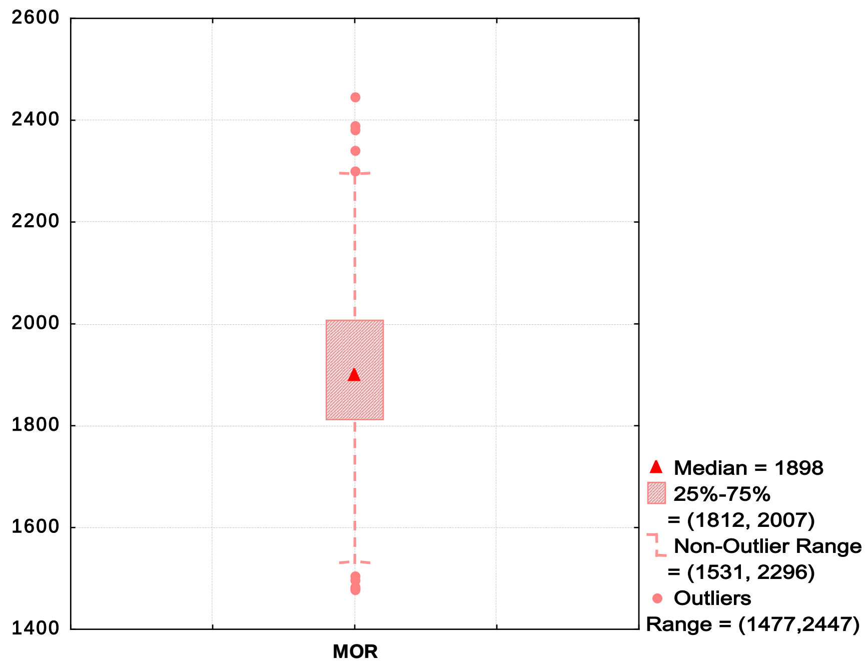
Predictive Model

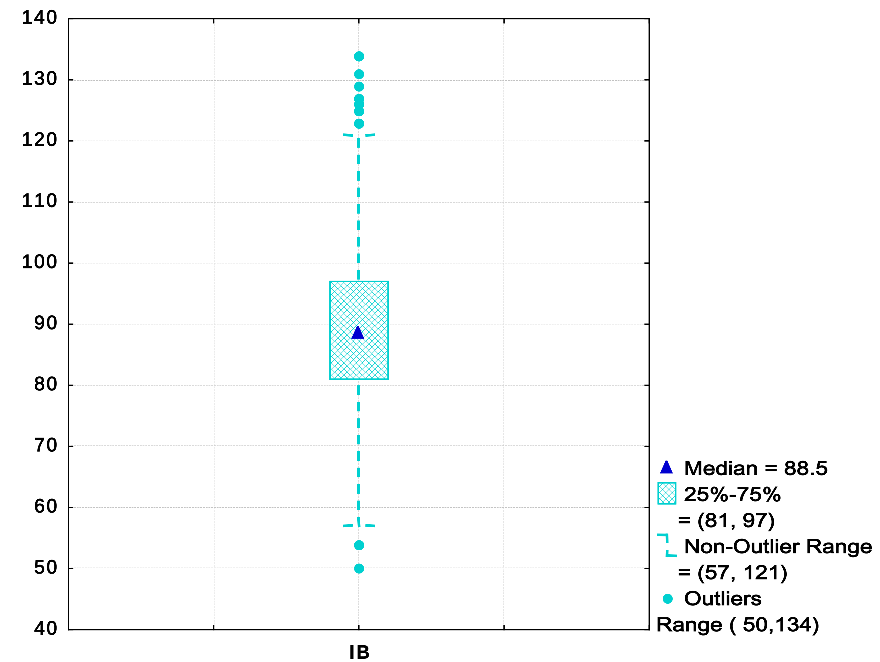
Data

Products : ½ UPINE ¾ UPINE 3/8 UPINE, 5/8 UPINE,11/16 UPINE

189 standardized predictor variables and ***1,084*** complete observations

Average ***MOR*** and ***IB*** as responses





Variable Explanation

Predictive Model

$y_{MOR}=0$ — 25%

$y_{MOR}=1$ — 75%

MOR	V1	V2	V3	…	V189
1,477	X	X	X	…	X
	X	X	X	…	X
1,812	X	X	X	…	X
…	X	X	X	…	X
	X	X	X	…	X
2,007	X	X	X	…	X
…	…	…	…	…	…
2,447	X	X	X	…	X

$y_{IB}=0$ — 25%

$y_{IB}=1$ — 75%

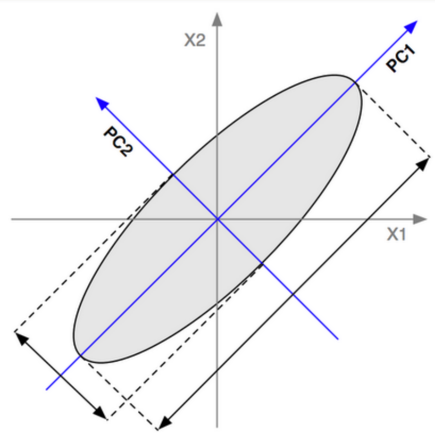
IB	V1	V2	V3	…	V189
50	X	X	X	…	X
	X	X	X	…	X
81	X	X	X	…	X
…	X	X	X	…	X
	X	X	X	…	X
97	X	X	X	…	X
…	…	…	…	…	…
134	X	X	X	…	X

Principal Component Analysis (PCA) Removes Severe Collinearity among Predictor Variables

Use an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of ***linearly uncorrelated*** variables

Advantage – Reduce the number of variables, but incorporate as much information as possible

Final Predictor Factors after PCA – ***13*** independent factors, preserving ***80%*** variation



Variable Explanation

Predictive Model

Factors	Description	Variation Proportion
Factor1	Actual Line Speed; Actual Value Distance (01-28) left/right Forming Line Mat weight Set Point	33.70%
Factor2	ThCt pressure frame (05-07) left ThCt pressure frame (11-15,18-21,23) left/right	13.18%
Factor3	Top/Bottom Face Former Pounds per square Foot MPot (01-05,07-09)pressure Track 4 ThCt pressure frame 22 left/right Percent of speed 1 Water Injection Control Output	11.25%
Factor4	MPot (01-06) pressure Track 1 + 7	5.26%
Factor5	ThCt pressure frame 05-06 right	3.89%
Factor6	Steam Injection Control Output Top Face Former feet per Minute #1/#2 Dry Refiner Infeed Chip Temperature Top/Bottom Core Former feet per Minute	3.06%
Factor7	Face Resin GPM	2.30%
Factor8	Core Blender Motor current in percent	2.02%
Factor9	# 2 Dry Refiner Infeed Chip Temperature	1.41%
Factor10	Face Ratio Of Shavings Setpoint	1.35%
Factor11	Out Of Press Board Width	1.23%
Factor12	Press Temperature Zone (2-3) Core Resin Usage in Percent	1.19%
Factor13	Core Resin Percent Solids OD Wood	0.96%

Model

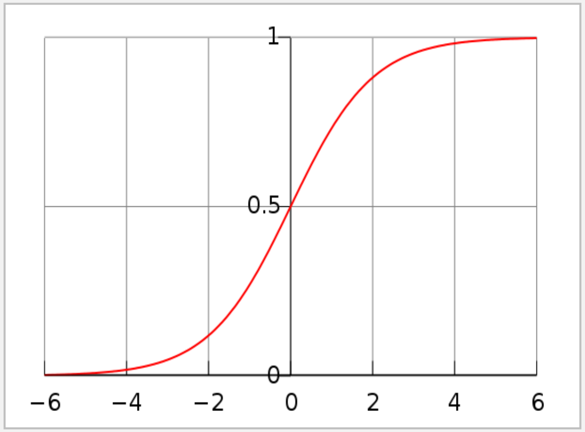
Predictive Model

Logistic Regression

$$\mathrm{logit}[\theta_i] = \log_e \frac{\theta_i}{1-\theta_i} \text{ where } \theta_i = \text{probability of occurrence of the event}$$

$$\log_e \frac{\theta_i}{1-\theta_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

$$\theta_i = \frac{1}{1+e^{-[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n]}}$$



- *No assumption on linearity and normality*

Bayes' Theorem

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)\,p(y \mid \theta)}{p(y)}$$

- *Prior and posterior probability distributions*
- *Extend logistic regression model in a Bayesian framework (Xu and Akella 2008)*
- *Use Bayesian Inference Methods for coefficient estimates (β)*

Bayesian Logistic Regression

Predictive Model

In Mathematics

$$p(y=0 \mid x, M, D) = \int_\beta p(y=0, \beta \mid x, M, D)\,\partial\beta$$

$$= \int_\beta p(y=0 \mid x, \beta, D)\,p(\beta \mid M, D)\,\partial\beta$$

where

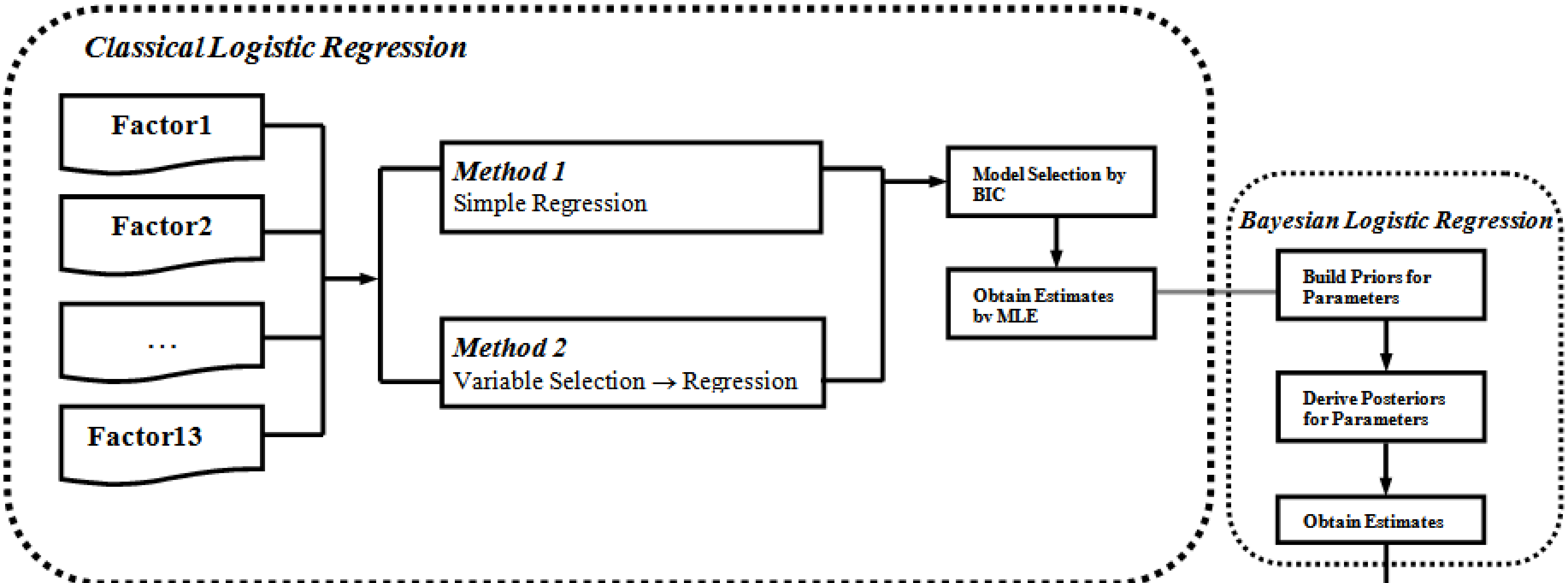
$$p(y=0 \mid x, \beta, D) = \{1+\exp[-\beta^T x]\}^{-1}$$

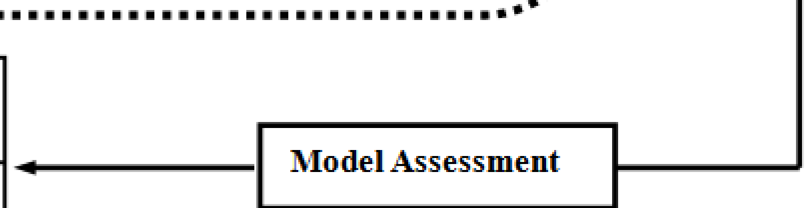
Priors

- ***Non-informative prior:***
  - *Prior 1: Uniform prior distribution* $p(\beta) \propto constant$
- ***Informative prior:***
  - *Prior 2: Gaussian prior distribution* $p(\beta \mid \mu, \sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(\beta-\mu)^2}{2\sigma^2}\right)$

Logic

Predictive Model





Predicted Value / Actual Value	Low MOR/IB (0)	High MOR/IB (1)
Low MOR/IB (0)	*True Prediction*	*False Prediction*
High MOR/IB (1)	*False Prediction*	*True Prediction*

Results

Predictive Model

Significant Factors

Significant Factors	MOR	IB
Factor1	+	-
Factor3		+
Factor4	-	+
Factor5	+	
Factor6	+	-
Factor8		+
Factor12	-	

MOR

Factor1	Actual Line Speed Actual Value Distance (01-28) left/right Forming Line Mat weight Set Point
Factor4	MPot (01-06) pressure Track 1 + 7
Factor5	ThCt pressure frame 05-06 right
Factor6	Steam Injection Control Output Top Face Former feet per Minute #1/#2 Dry Refiner Infeed Chip Temperature Top/Bottom Core Former feet per Minute
Factor12	Press Temperature Zone (2-3) Core Resin Usage in Percent

IB

Factor1	Actual Line Speed Actual Value Distance (01-28) left/right Forming Line Mat weight Set Point
Factor3	Top/Bottom Face Former Pounds per square Foot MPot (01-05,07-09)pressure Track 4 ThCt pressure frame 22 left/right Percent of speed 1 Water Injection Control Output
Factor4	MPot (01-06) pressure Track 1 + 7
Factor6	Steam Injection Control Output Top Face Former feet per Minute #1/#2 Dry Refiner Infeed Chip Temperature Top/Bottom Core Former feet per Minute
Factor8	Core Blender Motor current in percent

Results (MOR)

Predictive Model

Misclassification *and* ***Correct Classification rates*** *for validation dataset* with ***MOR*** as Response

Run	Classical Logistic Regression		Bayesian Logistic Regression (uniform prior)		Bayesian Logistic Regression (Gaussian prior)	
	Misclassification	Correct Classification Rate for y=0	Misclassification	Correct Classification Rate for y=0	Misclassification	Correct Classification Rate for y=0
1	0.32	0.67	0.32	0.67	0.32	0.67
2	0.29	0.76	0.29	0.76	0.29	0.76
3	0.24	0.81	**0.23**	**0.83**	0.24	0.81
4	0.22	0.8	0.22	0.8	0.22	0.8
5	0.28	0.74	0.28	0.74	0.28	0.74
6	0.3	0.74	0.3	0.74	0.3	0.74
7	0.32	0.74	**0.31**	**0.78**	**0.31**	**0.78**
8	0.34	0.72	**0.33**	0.72	0.34	0.72
9	0.34	0.61	**0.33**	**0.65**	0.34	0.63
10	0.3	0.65	0.3	0.63	0.3	0.63
Average	0.3	0.73	**0.29**	**0.73**	0.29	0.73

Results (IB)

Predictive Model

Misclassification *and* ***Correct Classification rates*** *for validation dataset* with ***IB*** as Response

Run	*Classical Logistic Regression*		*Bayesian Logistic Regression (uniform prior)*		*Bayesian Logistic Regression (Gaussian prior)*	
	Misclassification	Correct Classification Rate for y=0	Misclassification	Correct Classification Rate for y=0	Misclassification	Correct Classification Rate for y=0
1	0.3	0.69	**0.29**	0.69	0.3	0.69
2	0.22	0.72	**0.21**	**0.74**	**0.21**	**0.74**
3	0.19	0.82	**0.18**	**0.88**	0.19	0.82
4	0.26	0.75	0.26	0.73	0.26	0.75
5	0.15	0.87	**0.14**	**0.88**	0.15	0.87
6	0.21	0.86	**0.2**	0.86	**0.2**	0.86
7	0.18	0.81	0.18	0.81	0.18	0.81
8	0.19	0.77	0.19	0.77	0.19	0.77
9	0.21	0.78	0.21	0.78	0.21	0.78
10	0.21	0.8	**0.19**	**0.82**	0.21	0.8
Average	0.21	0.79	**0.2**	**0.8**	0.21	0.79

Summary

- **EM** and **MI-MCMC** *achieved more precise results for imputation*

- **Bayesian Logistic Regression** *identified significant factors influencing low strength properties*

- *On average, Bayesian logistic regression had a correct classification rate for low strength properties of* ***73%*** *for* ***MOR****, and* ***80%*** *for* ***IB***

References

American Forest and Paper Association. (2012). Economic Impact. Retrieved on 2012-10-31. http://www.afandpa.org/ourindustry.aspx?id=35

Tibshirani, R. (1006). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*. 58 (1), 267-288.

Xu, Z. and Akella, R. (2008). A Bayesian logistic regression model for active relevance feedback. *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*. 227-234.