The Application of STATA's Multiple Imputation Techniques to Analyze a Design of Experiments with Multiple Responses

STATA Conference - San Diego 2012

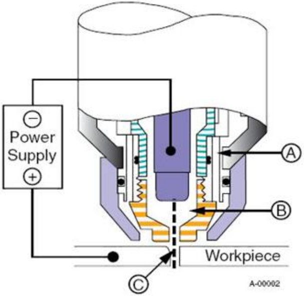
Clara Novoa, Ph.D., Bahram Aiabanpour, Ph.D., Suleima Alkusari
Texas State University - San Marcos, TX, USA
Ingram School of Engineering

Overview

- Introduction
  - Previous work
  - Motivation
- Multiple Imputation Methodology
- Multiple Imputation in STATA
- Results
- Conclusions

Introduction

Plasma Cutting Technology



Response Surface Methodology

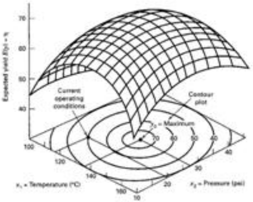
- Methodology selected for finding the best machine settings (factor levels) that optimize multiple part quality characteristics (responses)
- The usually unknown relationship between a response (y) and the affecting factors (x's) is modeled with polynomials, for example, a second-order model

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \epsilon$$

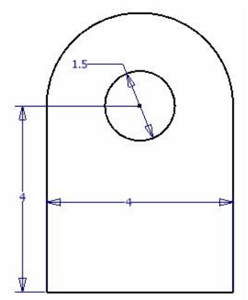
- The polynomial model can be a reasonable approximation of the true functional relationship (Montgomery and Runger, 2006)

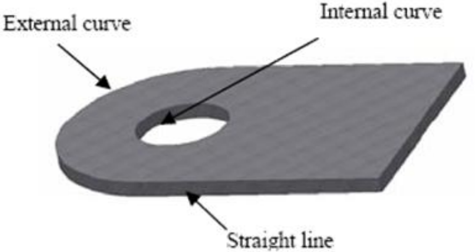
Response Surface Methodology (continuation)

- **Experimental design** permits the collection of data for the response variable at different levels of the independent variables
- **Least squares method** permits the estimation of the parameters, β 's, in the approximating polynomials
- **Linear/non-linear optimization** techniques permits the finding of an optimum point $(x_1^*, x_2^*, ..., x_k^*)$ and an optimal response value (y^*)



- All cuts were made on stainless steel sheet metal of 0.25 inch thickness





Experimental Design - Factors and Levels

Factor	Name	Low	Medium	High	Units
A	Current	40	60	80	Amps
B	Pressure	60	75	90	Psi
C	Cut Speed	10	55	100	Ipm
D	Torch height	0.1	0.2	0.3	Inch
E	Tool type *1	A	B	C	
F	Slower on curve	0	2	4	
G	Cut direction	Vertical		Horizontal	
		(G_0)		(G_1)	

*1 In experiment with missing values level names were (E_1, E_2, E_3)

*1 In experiment with imputed values names names were (E_0, E_1, E_2)

Experimental Design - Responses

- A total of 15 response variables

Surface Roughness	Flatness	Accum. Underneath	Part Geometry	Bevel Angle	Start Point Quality
(3)	(1)	(3)	(2)	(4)	(2)
Int. curve		Int. curve	x	Int. curve	Internal edge
Ext. curve		Ext. curve	y	Ext. curve	External edge
Str. line		Str. line		Left Line	
				Right line	

Experimental Design

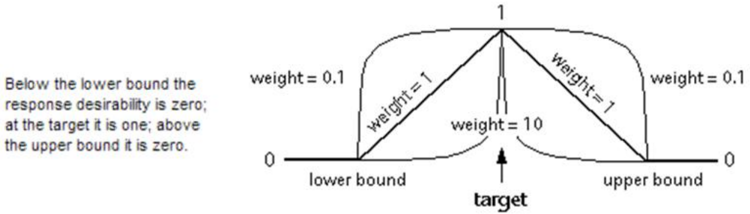
- **Taguchi orthogonal array L-18** (18 rows and 8 columns)
  - Each row represents an experimental run
  - One factor at two levels and four to seven factors at three levels
  - Economic alternative to a full factorial experiment (1458 runs if one replicate or 2916 if two-replicates)
- Design augmented with 71 additional runs to estimate two factor interactions (end with no aliases for two-factor interactions)
- Final number of runs is 89
- Objective is to fit valid models for each response (y_i) as a function of the critical factors (some of the x's). For example, a fitted second-order model

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j} \sum \hat{\beta}_{ij} x_i x_j$$

Optimization using Desirability Functions - Derringer and Suich (1980)

- There are 3 types of desirability functions. Response must hit the target (T), response is to be minimized or response is to be maximized

Examples of desirability functions (d_i) for the case response (y_i) must hit a target



Desirability Function - Target is Best

$$d_i(\hat{Y}_i(x)) = \begin{cases} 0 & \hat{Y}_i(x) < L_i \\ \left(\frac{\hat{Y}_i(x) - L_i}{T_i - L_i}\right)^s & L_i < \hat{Y}_i(x) < T_i \\ \left(\frac{\hat{Y}_i(x) - U_i}{T_i - U_i}\right)^t & L_i < \hat{Y}_i(x) < T_i \\ 0 & \hat{Y}_i(x) > U_i \end{cases}$$

- The desirability function "target is best" transforms the response values to values between 0 and 1, zero if below a lower bound (L) or one if above an upper bound (U)
- The shape of the desirability function is determined by the values of the weight parameters s and t (function exponents)
- Settings for independent variables or factors affect the predicted response and the desirability function values

Optimizing the Overall Desirability

maximize

$$D = \left(\prod_{i=1}^{n} d_i(\hat{Y}_i(x))^{w_i}\right)^{\frac{1}{\sum_{i=1}^{n} w_i}}$$

subject to

$$Low \leq x \leq High$$

- This is a non-linear deterministic optimization model with objective function to maximize the overall desirability. Weights w_i represent the importance given to response y_i
- x is the vector of model decision variables corresponding to the non-categorical experimental factors (current, pressure, cut speed, torch height, and slower on curves)
- Constraints in the model say that decision variables x's must to take values within the experimented region (Low-High). Categorical factors tool type and cut direction are fixed to each one of their 6 possible levels. Thus, six different optimization models need to be solved in this study

Research Motivation

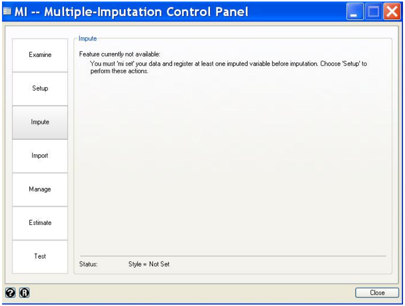
- 43 experimental conditions had missed responses (36 had all responses missing and other 7 had some responses missing)
- Analysis of the experiment done through general linear regression model (GLM) ignoring the missing responses
- **Is multiple imputation (MI) an effective method for completing and analyzing this experimental design?**

Multiple Imputation (MI)

- Method proposed by Rubin (1987). It is a simulation-based approach for analyzing incomplete data (Manchenko, 2010)
- Each missing value is replaced with a random sample of simulated values that represent the uncertainty about the right value (Rubin, 1987)
- User specifies the size of the random sample (number of imputations to add)
- Includes 3 steps: imputing, conducting analysis with each complete set of data, and analyzing aggregate results
- Variances of the parameter estimates are estimated more accurately than in single-imputation reducing the type I error
- In contrast to single-imputation, MI permits to estimate the impact of missing information on parameter estimation (McKnight, et al., 2007)

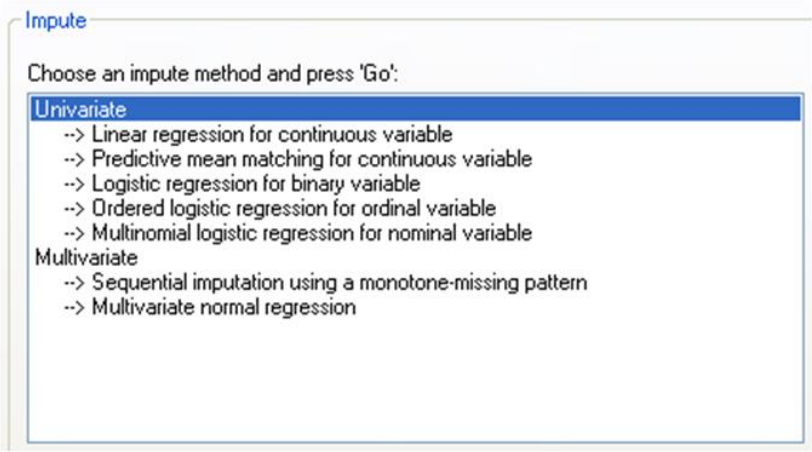
MI in STATA 11 - Multiple Imputation Control Panel

- The MI control panel can be accessed from the main menu under the Statistics option



- Some relevant steps needed are, registering the variables that will be imputed (. mi register imputed), looking at the summary of missing data (.mi misstable summarize), looking at the data statistics (.mi describe), looking at some patterns for missing information (.mi misstable patterns), deciding on the format to save the imputations (for example .mi set mlong)

Impute Options in STATA 11



Impute Command - Example

```
. mi impute pmm newFlatness = Current Pressure Cut_speed Torch_height Slowoncurv
> es E_0 E_1 G_0, noconstant add(5)

Univariate imputation                       Imputations =        5
Predictive mean matching                          added =        5
Imputed: m=1 through m=5                        updated =        0

             |            Observations per m
             |----------------------------------------------
    Variable |   complete   incomplete   imputed |     total
-------------+-----------------------------------+----------
 newFlatness |         53           36        36 |        89
------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled in observations.)
```

- In this example, the number of imputations for each missing value, m, is 5 and the imputation method selected was predictive mean matching (pmm)

Impute Options - Predictive Mean Matching (pmm)

- Preferred to linear regression when the normality of the underlying model is suspect
- Introduced by Little (1988) based on Rubin (1986)
- Prediction of linear regression is used as a distance measure to form the set of nearest neighbors or donors for the imputation
- Randomly draws a value from the set of nearest neighbors to impute the missing value
- By drawing from the observed data ppm preserves the original distribution of the observed values
- Estimates of the model parameters are simulated from their joint posterior distribution

Estimate Command - Example - Output 1

```
. mi estimate : regress newFlatness Cut_speed E_0

Multiple-imputation estimates                   Imputations     =          5
Linear regression                               Number of obs   =         89
                                                Average RVI     =     0.4335
                                                Complete DF     =         86
DF adjustment:   Small sample                   DF:     min     =      20.04
                                                        avg     =      34.43
                                                        max     =      50.17
Model F test:       Equal FMI                   F(   2,    22.0) =      4.65
Within VCE type:          OLS                   Prob > F        =     0.0207
```

newFlatness	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Cut_speed	-.0000614	.0000325	-1.89	0.067	-.0001275	4.62e-06
E_0	.0060596	.0027282	2.22	0.038	.0003694	.0117498
_cons	.0273574	.0022366	12.23	0.000	.0228654	.0318495

- The first time the command mi estimate was invoked, a regression (regress) for newFlatness as dependent variable and all the possible terms in a second order polynomial model on the factors (current, pressure, cut speed torch height, slow on curve, tool type and cut direction) was performed. Quadratic terms and second order interactions were included except those involving categorical variables
- By performing iteratively the command mi estimate, we eliminated from the model the non-significant factors one at a time until obtaining a **final regression model with only significant factors for each response**

Estimate Command - Example - Output 2

```
. mi estimate, vartable nocitable

Multiple-imputation estimates                    Imputations      =          5

Variance information
```

	Imputation variance Within	Between	Total	RVI	FMI	Relative efficiency
Cut_speed	8.0e-10	2.1e-10	1.1e-09	.317904	.262372	.950142
E_0	4.8e-06	2.2e-06	7.4e-06	.549906	.391906	.927316
_cons	4.2e-06	6.4e-07	5.0e-06	.181203	.163194	.968393

Note: FMIs are based on Rubin's large-sample degrees of freedom.

$$efficiency = \frac{1}{1 + \frac{\gamma}{m}}$$

$$\gamma = \frac{r + 2/(df + 3)}{r + 1}$$

$$r = \frac{(1 + m^{-1})B}{\bar{U}}$$

$$df = (m - 1)(1 + \frac{m\bar{U}}{(m + 1)B})^2$$

RVI = Relative variance increase due to non-response
FMI = Fraction of missing information
The smaller the RVI and FMI values the better
RVI can be greater than 1

Relative efficiency value, the closer to 1 the better

Deterministic Optimization Model

- The multi-response non-linear optimization model was laid out in Excel
- Risk Solver Platform (RSP) software from Frontline Systems was used for the optimization step.
- The optimization technique used by RSP to solve the non-linear non-smooth optimization problem is genetic algorithms (GA)
- Solve times were less than 1 minute 43 seconds in all runs and the mean was 55.74 seconds

Excel - Risk Solver Platform Deterministic Optimization Model

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Desirability Optimization Model for Optimizing the Automated Plasma Cutting Process																
2	**Constraints**																
3		Lower	Upper														
4	Current	40	80														
5	Pressure	60	90														
6	Cut speed	10	110														
7	Torch Height	0.1	0.3														
8	Slower on Curves	0	4														
9	Overall D =	**0.9221012**															
10		Estimated resp.	**10.34**	**10.00**	**9.82**	**0.02**	**17.50**	**10.00**	**9.96**	**4.00**	**6.00**	**4.97**	**8.62**	**-0.50**	**4.38**	**9.89**	**9.89**
11		Function	Max	Max	Max	Min	Max	Max	Max	Target	Target	Target	Target	Target	Target	Max	Max
12		Lower	**5**	**5**	**5**	0.015	**5**	**5**	**5**	**3.5**	**5.5**	**-20**	**-20**	**-20**	**-30**	**5**	**5**
13		Upper	10	10	10	**0.1**	10	10	10	**4.5**	**6.5**	**20**	**20**	**20**	**30**	10	10
14		Target	**10**	**10**	**10**	**0**	**10**	**10**	**10**	**4**	**6**	**0**	**0**	**0**	**0**	**10**	**10**
15		t	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16		s	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17		Desirability, di	1.00	1.00	0.96	0.77	1.00	1.00	0.99	0.99	1.00	0.75	0.57	0.97	0.85	0.98	0.98
18		Weight	4	4	4	1	4	4	4	2	2	3	3	3	3	2	2
19		Sum weights	45														
20		Desirability ^ (weig	**1.00**	**1.00**	**0.87**	**0.77**	**1.00**	**1.00**	**0.97**	**0.98**	**1.00**	**0.42**	**0.18**	**0.93**	**0.62**	**0.96**	**0.96**
21			**Coefficients from Clean Regressions (input data)**														
22			**New_Rough_IC**	**New_Rough_EC**	**New_Rough_SL**	**New_Flatness**	**New_Acum_IC**	**New_Acum_EC**	**New_Accum_SL**	**New_Geom_Accur_x**	**New_Geom_Accur_y**	**New_bevel_int_cur**	**New_bevel_ext_cur**	**New_bevel_left_sl**	**New_bevel_**	**New_Startp_inter**	**New_Startp_exter**
23																	
24	Constant	1	0	0	6.768983	0.0273574	7.450695	2.225542	2.86539	4.030143	6.31409	-15.01253	-11.59085	-13.493	0	11.88003	8.062731
25	Current	79.5049903	0.2321414	0.2212003							-0.0122	0.606645				0.023298	
26	Pressure	90														-0.10854	
27	Cut speed	64.83694566				-6.14E-05	0.1280647	0.1820017	0.1318934	-0.0005453			0.151244	0.03286			
28	Torch Height	0.273264827	6.236696	10.46248	10.01399								38.09256	112.743			
29	Slower on Curves	0	-0.2198405				-30.20556										
30	E_0	0	-0.9734168	-0.9127137	-0.480179	0.0060596				-0.0176562	-0.03815			-3.317	9.452		-0.6893232
31	E_1	1						0.945579	1.056404								
32	E_2	0															
33	G_0	0	-0.6539388	-0.7728354													-1.026018
34	G_1	1															
35	Current^2	6321.043482	-0.0016838	-0.0012405							0.0001						
36	Pressure^2	8100	-0.0003931		-0.000252		-0.0005057										-0.0004163
37	Cut_speed^2	4203.829523	-0.0004677	-0.0003235	-0.000347		-0.0008573	-0.0011832	-0.0009277							-0.00028	-0.0004047
38	Torch Height^2	0.074673666								-0.7538644		-199.9436		-267.12		-45.6996	

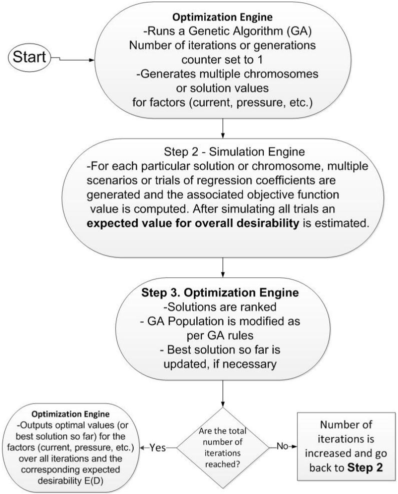
Numerical Results

Factor	Experiment no imputation	Experiment with MI
Current	80	80
Pressure	90	90
Cut Speed	**55**	**65**
Torch height	0.3	0.3
Slower on Curves	0.4	0
Tool Type	Third tool	Second tool
Cut direction	Horizontal	Horizontal

Conclusions and Further Research

- **MI under STATA** proved to be effective to analyze the plasma cutting experiment with missing values
- After MI, it was discovered that a setting with **slightly higher speeds** do not negatively affect response variables and overall desirability
- MI reports on the variability of the estimates of the regression coefficients. This variability may be included in a **stochastic simulation optimization model** that Risk Solver Platform (RSP) can solve
    - The stochastic optimization model objective function is now to minimize the expected overall desirability under the same constraints as in the deterministic optimization model
    - β's in the regression models are now random variables with a given mean and standard error. Desirability's will depend on responses which will be a function of the factors (x's) and the realizations for the β's

Steps in Stochastic SimulationOptimization Model



References

- 1. Asiabanpour, B., Vejandla, D. T., Jimenez, J., and Novoa, C., 2009, Optimizing the Automated Plasma Cutting Process by Design of Experiments, International Journal of Rapid Manufacturing 1(1), 19−40.
- 2. Vejandla, D. T., 2009, Optimizing the Automated Plasma Cutting Process by Design of Experiments, Masterś Thesis, Texas State University, Department of Engineering Technology.
- 3. Montgomery, D. C., 2001, Design and Analysis of Experiments, 5th Edition, John Wiley & Sons, Inc., New York.
- 4. Montgomery, D.C., Runger, G.C., 2006, Applied Statistics and Probability for Engineers, 4th Edition, John Wiley & Sons, Inc., New York.
- 5. Godolphin, J. D., 2006, Reducing the Impact of Missing Values in Factorial Experiments Arranged in Blocks, Quality and Reliability Engineering International 23, 669−682.
- 6. Yuan, Y. C., Multiple Imputation for Missing Data: Concepts and New Development (version 9.0), SAS Institute Inc., Rockville, MD.
- 7. Castillo, E. D., Montgomery, D., and McCarville, D., 1996, Modified Desirability Functions for Multiple Response Optimization, Journal of Quality Technology, 28(3), 337−345.
- 8. NIST−SEMATECH, 2003, Section 5.5.3.2.2: Multiple Responses: The Desirability Approach in e-Handbook of Statistical Methods, Engineering Statistics Handbook. Online. http://www.itl.nist.gov/div898/handbook/pri/section5/pri5322.htm. Last date accessed Jan 16, 2012.

References - Continuation

- 9. ArtilesLeon, N., NovoaRamirez, C. M. & Domenech, C., 1995. Improving fabric finishing through experimental design, ASQC 49th Annual Quality Congress Proceedings, USA, pp. 952-961.
- 10. Koller-Meinfelder, F., 2009. Analysis of Incomplete Survey Data Multiple Imputation via bayesian Bootstrap Predictive Mean Matching, Germany: Otto-Friedrich-Universitat Bamberg. Online. Last date Accessed July 23, 2012.
- 11. Manchenko, Y., 2010. Mutipleimputation using Stata's mi command. Online. http://www.stata.com/meeting/boston10/boston10_marchenko.pdf. Last date accessed July 23, 2012.
- 12. McKnight, P. E., McKnight, K. M., Sidani, S. & Fuigueredo, J. A., 2007. Missing Data A Gentle Introduction. 1st ed. N.Y.: The Guilford Press.
- 13. Mealli, F. & Rubin, D. B., 2002. Assumptions when Analyzing Randomized Experiments with Noncompliance and Missing Outcomes. Health Services & Outcomes Research Methodology, Volume 3, pp. 225−232.
- 14. Rubin, D. B., 1987. Multiple Imputation for Nonresponse in Surveys. 1st ed. New York: Wiley.
- 15. Rubin, D. B., 1996. Multiple imputation after 18 years (with discussion). Journal of the American Statistical Association, 91(434), pp. 473−489.
- 16. STATA. Multiple imputation for missing data. Online. http://www.stata.com/stata11.mi.html. Last Date Accessed July 23, 2012.

References - Continuation

- 17. Schafer, J. The multiplie imputation frequently asked page. Online. Available at: http://sites.stat.psu.edu/j̃ls/mifaq.html.Last Date July 23, 2012
- 18. Verbeke, G. & Mohenberghs, G., 2000. Linear mixed models for longitudinal data. New York: SpringerVerlag.
- 19. Wayman, J. C., 2003, Multiple Imputation For Missing Data: What Is It And How Can I Use It? , Annual Meeting of The American Education Research Association, Chicago, IL., available at http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf (last date accessed Jan 16, 2012)

Questions