AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Analytical Modeling: A Technique for Handling Big (and small) Data

ITEA's 17th Test Instrumentation Workshop
21 May 2014
Las Vegas, NV

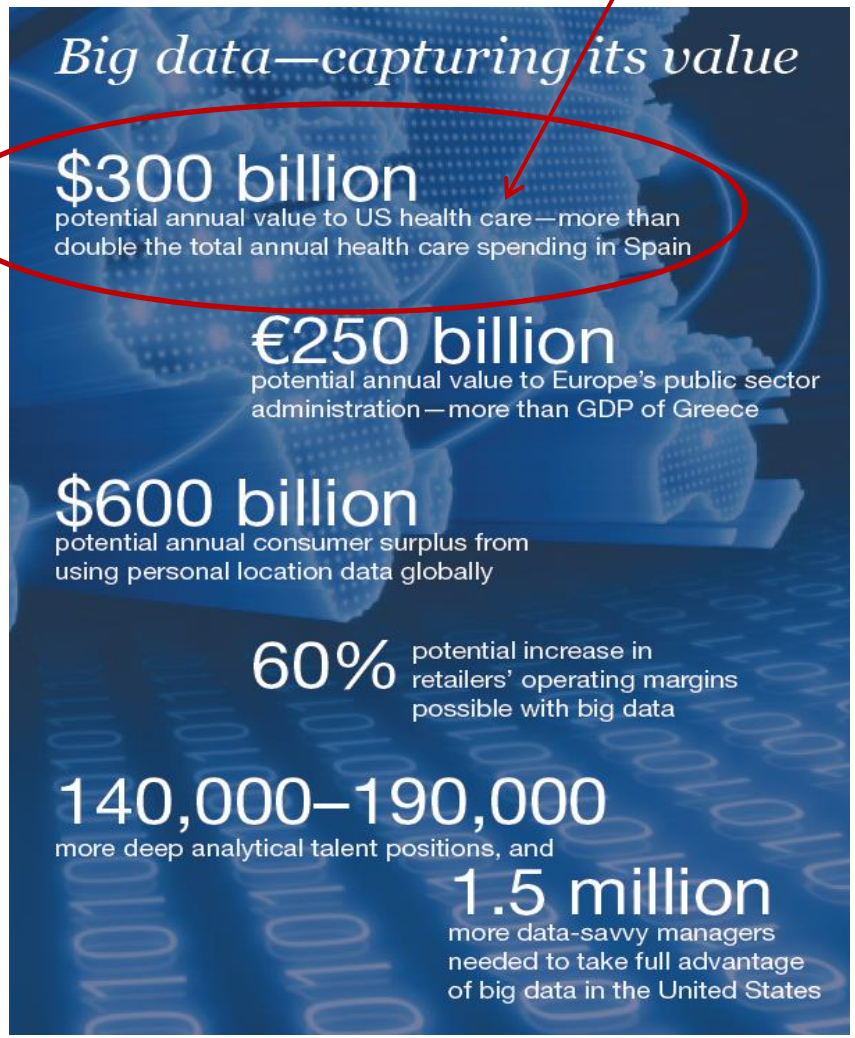
14-ANALMOD-5A

Mark J. Kiemele, Ph.D.
President and Co-Founder
Air Academy Associates

Office: 719-531-0777
Cell: 719-337-0357
mkiemele@airacad.com
www.airacad.com

The Case for Big Data*

The case study we address will come from the healthcare sector.



*** Source: McKinsey Global Institute**

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Motivation for Big Data Analytics

- The primary goal of big data analytics is to help organizations **make better decisions** by analyzing a large amount of data involving a large number of factors.
- Big data analytics can be done with software tools commonly used as part of advanced analytics disciplines such as **predictive analytics** and **data mining**.
- Today's topic, **Analytical Modeling**, falls within both of these categories and **uses the regression analysis tool**.

Air Academy Associates
Simplify, Perfect, Innovate

Motivation for Modeling from Dr. Gilmore (DOT&E)

(from his 26 June 2013 memo on Flawed Applications of DOE)

1. One of the most important goals of operational testing is to **characterize** a system's effectiveness over the operational envelope.

2. I advocate the use of DOE to ensure that test programs are able to **determine the effect of factors on** a comprehensive set of operational mission-focused and **quantitative response variables**.

3. Future test plans must state clearly that data are being collected to measure a particular response variable (possibly more than one) in order to **characterize** the system's performance by examining the effects of multiple factors … and clearly delineating what statistical **model** (e.g., main effects and interactions) is motivating … the variation of the test.

4. Confounding factors must be avoided.

5. Another pitfall to avoid is relying on binary metrics as the primary response variable.

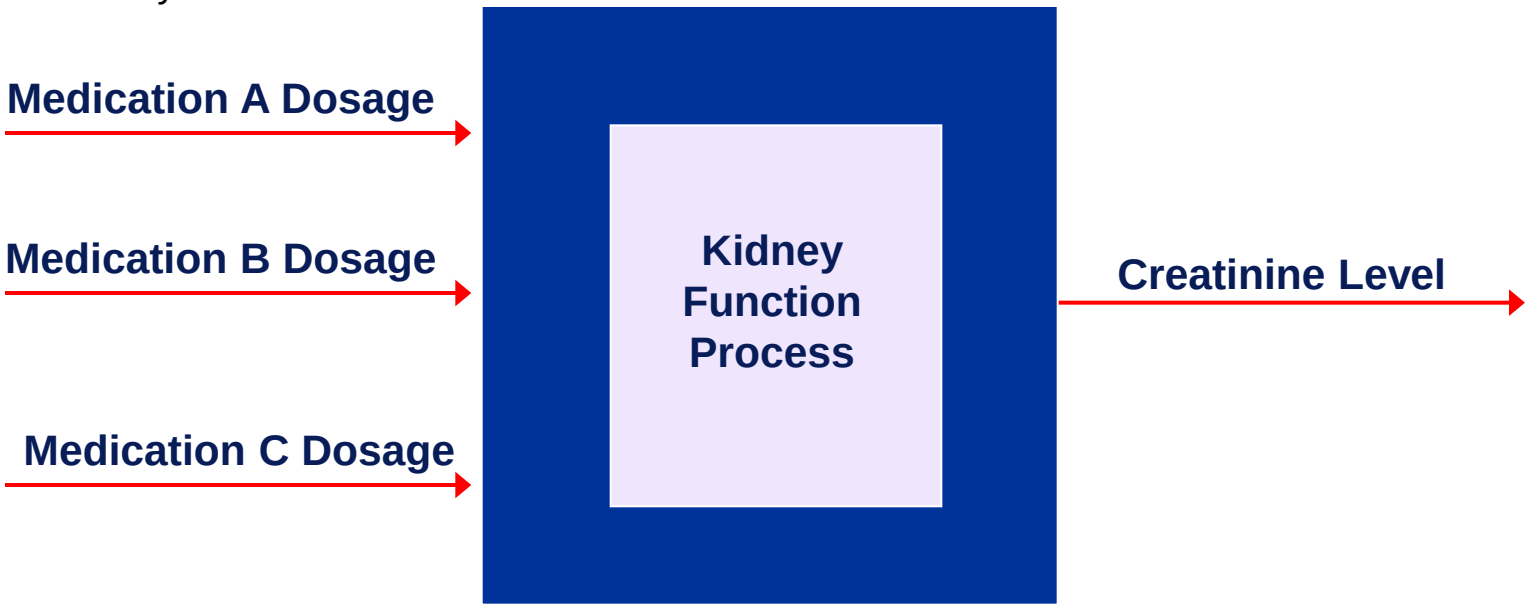
AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Analytical Modeling of Kidney Performance
(After a Kidney Transplant)

The input variables are dosages of three different medications given to a patient who has just received a kidney transplant.* The output (y) variable is the amount of Creatinine which should be minimized to avoid rejection. There are other important output variables as well, but we will look only at Creatinine in this case study to demonstrate the technique.

The data on the following page represent 38 consecutive days of post-operative treatment. Use regression analysis to build a model that will predict y (creatinine level) as a function of the input variables A, B, and C. Examine what effect each medication has on the response variable. Medication A was an experimental drug at the time. What can you say about its effect on y?



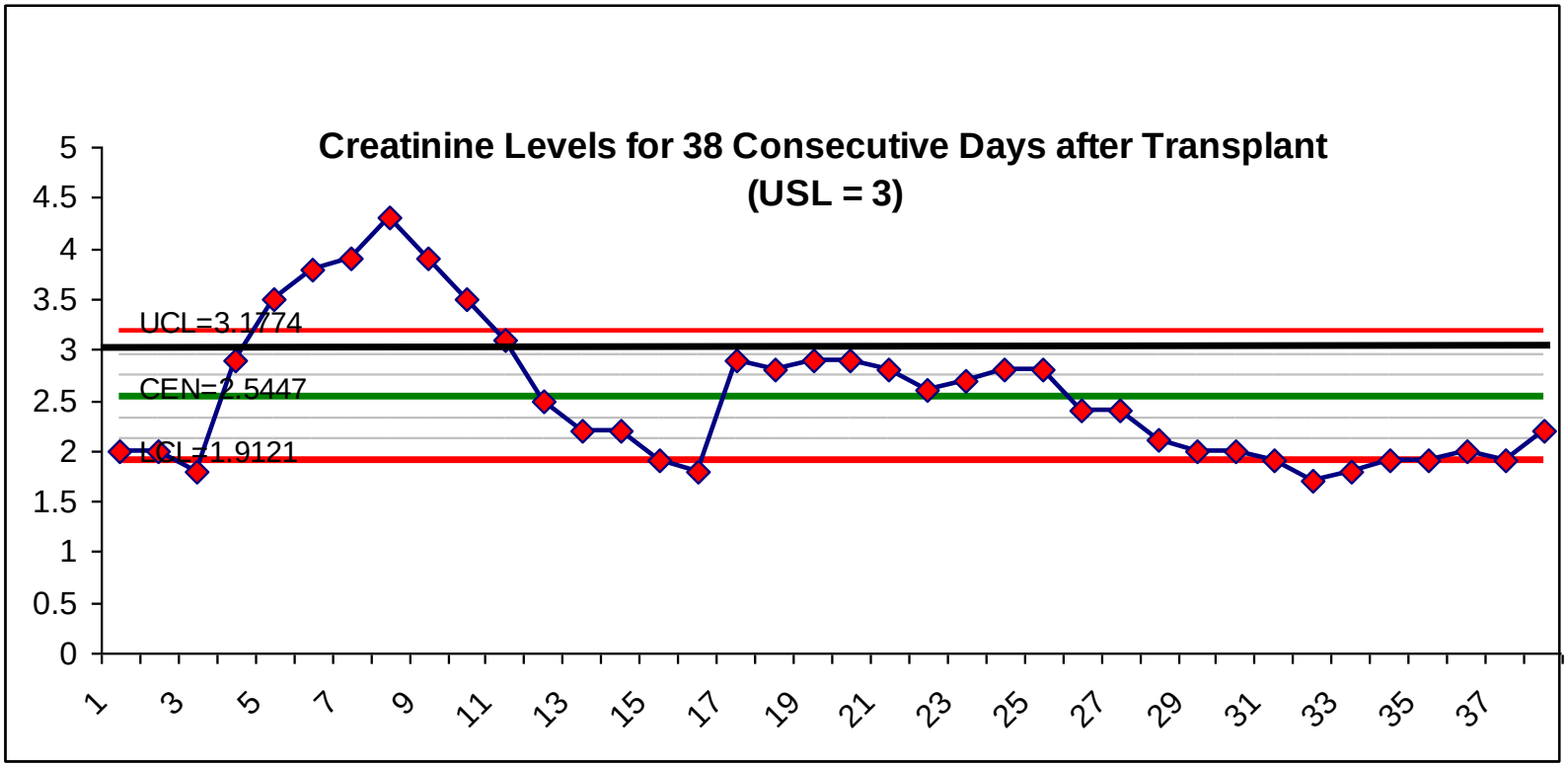
*More than 70,000 kidney transplants are performed each year worldwide.

Raw Data for 38 Consecutive Days

Factor	A	B	C		
Row #	A	B	C		Creatinine
1	4	2	36		2
2	4	2	20		2
3	5	2	20		1.8
4	5	2	20		2.9
5	6	2	20		3.5
6	9	2	80		3.8
7	9	2	50		3.9
8	11	2	20		4.3
9	12	2	20		3.9
10	13	2	20		3.5
11	15	2	20		3.1
12	16	2	20		2.5
13	17	2	20		2.2
14	19	2	20		2.2
15	20	2	20		1.9
16	20	2	20		1.8
17	18	2	20		2.9
18	16	2	20		2.8
19	13	2	20		2.9
20	14	2	0		2.9
21	13	2	0		2.8
22	12	2	0		2.6
23	13	2	0		2.7
24	15	2	0		2.8
25	16	2	0		2.8
26	17	2	0		2.4
27	18	2	0		2.4
28	9	2	0		2.1
29	17	2	0		2
30	16	2	20		2
31	16	2	20		1.9
32	14	2	16		1.7
33	14	2	16		1.8
34	14	2	14		1.9
35	13	1.5	12		1.9
36	13	0.5	12		2
37	11	1	10		1.9
38	11	1	8		2.2

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Patient Rejecting Kidney (days 5-11)



AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Analytical or Mathematical Model
(from applying regression analysis to the raw data)

Y-hat Model

Factor	Name	Coeff	P(2 Tail)	Tol	Activ
Const		-2.584	1.0000		
A	A	-9.368	1.0000	0.0000	X
B	B	5.116	1.0000	0.0000	X
C	C	-6.196	1.0000	0.0000	X
AA		-1.078	0.0058	0.7138	X
AB		8.846	1.0000	0.0000	X
AC		-13.385	1.0000	0.0000	X
BB		0.76596	0.3104	0.3172	X
BC		6.666	1.0000	0.0000	X
CC		-0.13637	0.7831	0.3613	X
ABC		12.805	1.0000	0.0000	X

R^2	0.4230
Adj R^2	0.2093
Std Error	0.6167
F	1.9794
Sig F	0.0772
F_{LOF}	3.1152
Sig F_{LOF}	0.0429

Source	SS	df	MS
Regression	7.5	10	0.8
Error	10.3	27	0.4
$Error_{Pure}$	1.4	9	0.2
$Error_{LOF}$	8.8	18	0.5
Total	17.8	37	

The Tol values inform us about the confounding of the factors. A Tol value is a value between 0 and 1. Bigger is better.

Rule of Thumb: In order to trust the p-values and coefficients, all Tols must be greater than .5 and most should be greater than .7.

Implication: We can reduce the confounding by reducing the model.

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Reduced Analytical Model

(by removing terms from the model)

Y-hat Model

	Factor	Name	Coeff	P(2 Tail)	Tol	Active
	Const		2.758	0.0000		
C	A	A	-0.21343	0.2577	0.8910	X
C	B	B	0.55961	0.0181	0.8988	X
C	C	C	0.61038	0.0273	0.8745	X
	AA		-0.99471	0.0021	0.8999	X

R^2	0.3884
Adj R^2	0.3143
Std Error	0.5742
F	5.2400
Sig F	0.0022
F_{LOF}	2.4988
Sig F_{LOF}	0.0773

Source	SS	df	MS
Regression	6.9	4	1.7
Error	10.9	33	0.3
$Error_{Pure}$	1.4	9	0.2
$Error_{LOF}$	9.5	24	0.4
Total	17.8	37	

Factor	Name	Low	High	Exper
A	A	4	20	16
B	B	0.5	2	1
C	C	0	80	20

Multiple Response Prediction

			99% Confidence Interval	
	Y-hat	S-hat	Lower Bound	Upper Bound
	1.9109	0.5742	0.188	3.634

AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

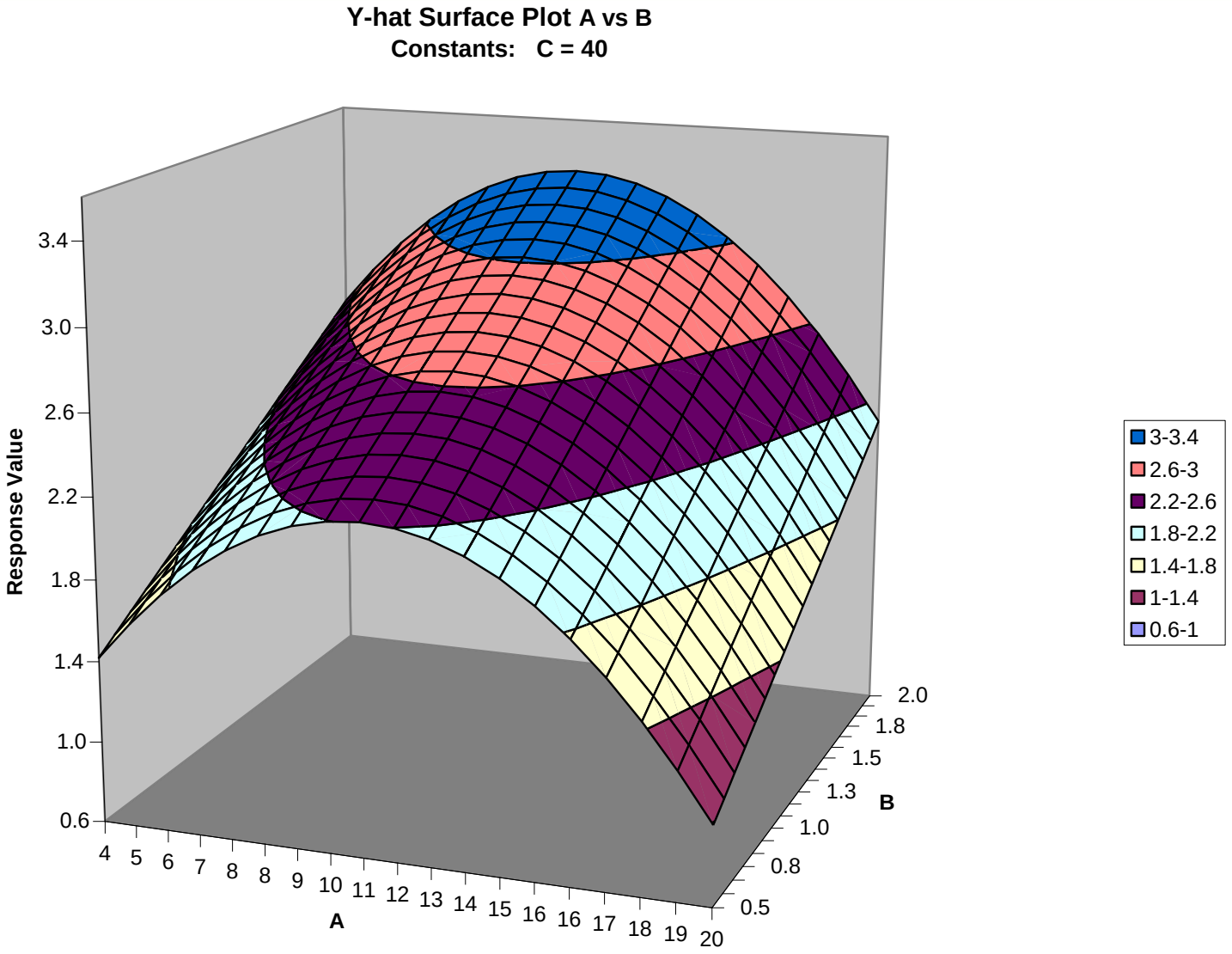
Analytical or Mathematical Model

An analytical model (for coded values of A, B, and C):

$$\text{Creatinine Level} = 2.8 - .21A + .56B + .61C - .99A^2$$

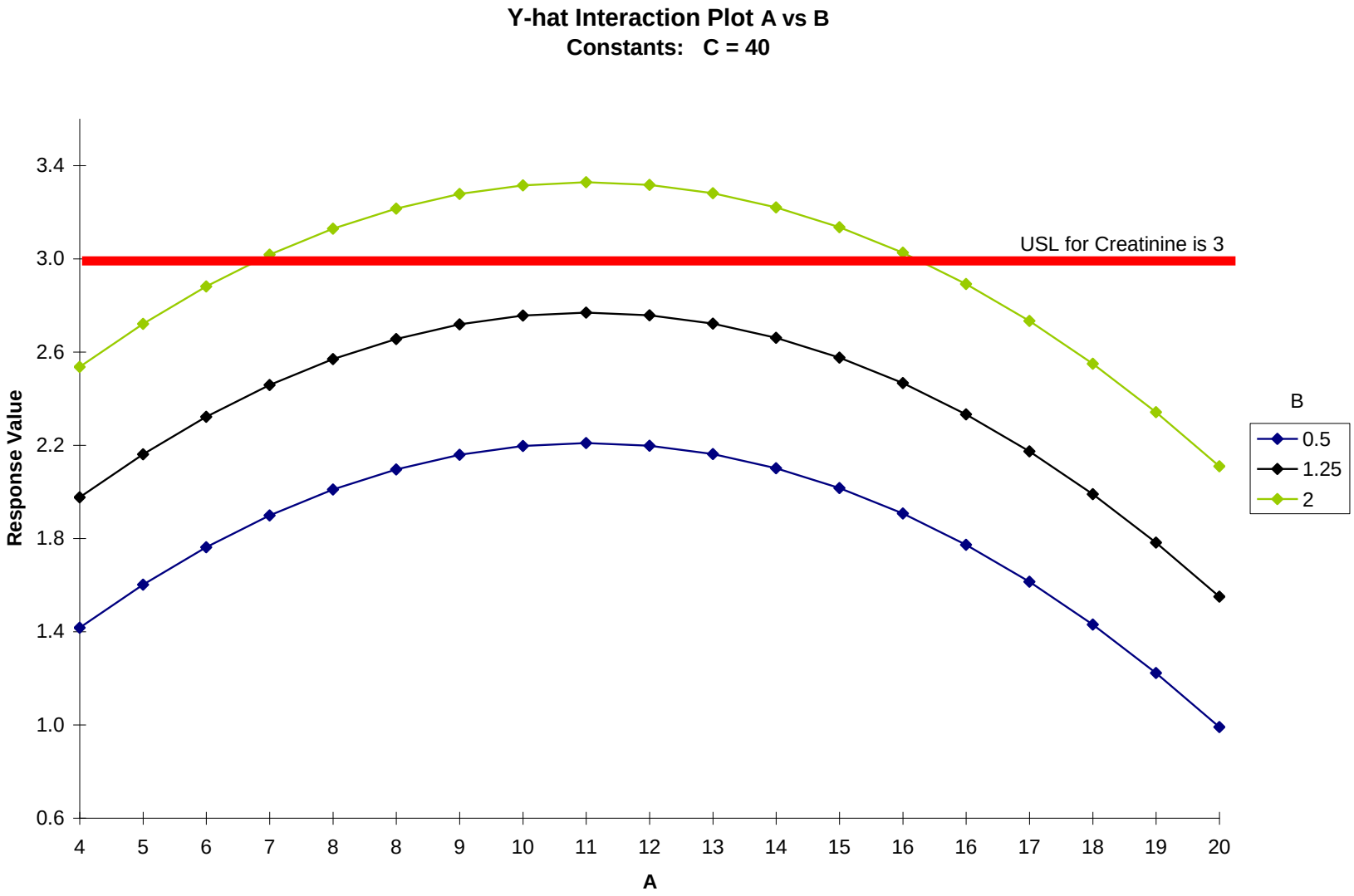
Air Academy Associates
Simplify, Perfect, Innovate

Surface Plot of the Model
(the geometry)



Air Academy Associates
Simplify, Perfect, Innovate

Interaction Plot
(obtained from the surface plot)



AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Summary of Big Data Analytics

The Bad News:

- There is much more noise in big data (historical data) than in data that would be collected under a DOE format.
- The more factors there are to be investigated, the more potential interactions there are and the more confounding there will be among the factors and their interactions.

The Good News:

- Regression analysis provides a technique to analyze big data (or small data as in our case study) for the purpose of identifying key relationships that exist among the factors, their interactions, and the response of interest. That means building analytical models, or characterizing the process, for the purpose of prediction and risk assessment.
- In T & E, this technique can be used to analyze test data that is combined from various test sites and/or to analyze existing test data to see which factors need further testing, i.e., factors for a DOE.

Thank You

Colorado Springs, Colorado

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate