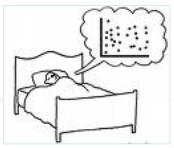


Problems Analyzing Historical Data*

*Posted at www.statease.com/webinar.html

Originally developed by Patrick J. Whitcomb
Presented by Mark J. Anderson, PE, CQE, MBA
(Email: Mark@StatEase.com)

If you are on a speaker phone, please put your microphone on mute. Thanks. But feel free to speak up with an urgent issue. However, I prefer you email questions to me. Much appreciated -- Mark

Timer by Hank Anderson

Problems Analyzing Historical Data
Lots of questions

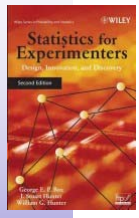
A FAQ from a Section Head of materials & testing development sent to StatHelp@StatEase.com on 9/1/09:

> "In many cases, we look to model results of existing experiments where a specific design of experiment (DOE) was not initially established. We can still, however, separate data into continuous variables and responses. For example, can we take existing variables and add interactions and square terms and test responses using linear regression? Can these be subsequently plotted using the standard contour and 3D plots?"

Answer: Press the "easy" button to avoid investing in DOE. ;)

Problems Analyzing Historical Data
Good advice

Box, Hunter and Hunter warn in regard to using least squares regression that:

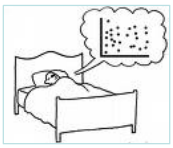


"If happenstance data are really all you can get, such analyses may be better than nothing. But they can be downright misleading as is reflected by the acronym PARC (practical accumulated records computations) and its inverse."

These problems are our focus in this webinar. The advice of these three preeminent design of experiments (DOE) experts is what we follow.*

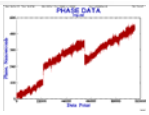
*Refer to Section 10.3 of *Statistics for Experimenters*, (2005, 2nd edition), George E. P. Box, William G. Hunter and J. Stuart Hunter, John Wiley and Sons, Inc.

Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #1



- Problems Analyzing Historical Data
  1. **Inconsistent data**
  2. Limited factor ranges
  3. Collinearity
  4. Nonsense correlation
  5. Serially correlated errors
  6. Dynamic relations
  7. Feedback control
- DOE avoids many problems

Problems Analyzing Historical Data
Inconsistent Data

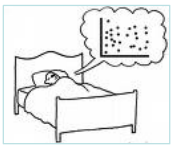


It is rare that data gathered over a long period of time is consistent and comparable.

- Standards are modified over time, instruments change, calibrations drift, operators come and go, raw materials change, processes age, ambient conditions change and there may be seasonal effects.
- Much that is relevant is unknown and not recorded.

That's all we will say on this "gotcha" – obviously with major drifts like that shown, one can get very good at predicting what happened last month. (A Joke!)

Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #2



- Problems Analyzing Historical Data
  1. Inconsistent data
  2. **Limited factor ranges**
  3. Collinearity
  4. Nonsense correlation
  5. Serially correlated errors
  6. Dynamic relations
  7. Feedback control
- DOE avoids many problems

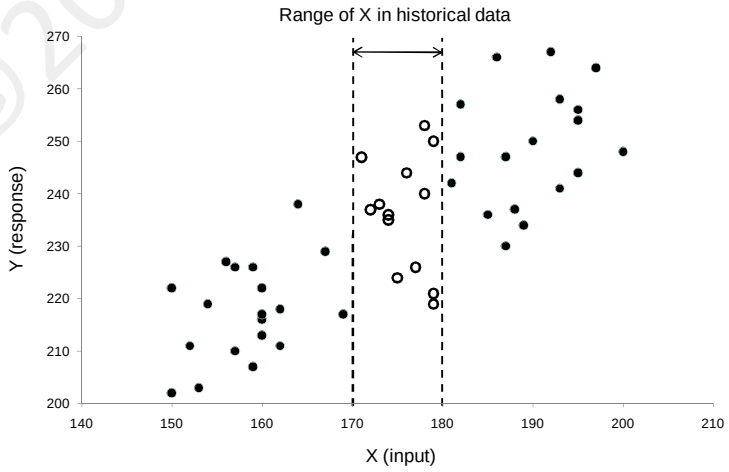
Problems Analyzing Historical Data
Limited Factor Ranges

Important factors are controlled:

- Variation about their set points is limited.
- Set points can be chosen to minimize the effect of factor variation.

This can lead to false conclusions. For example, from the historical data we conclude that a factor has no effect on the response; when in reality the factor is tightly controlled because it has a large effect.

Limited Factor Ranges
Full Range and Limited (by control) Range



Limited Factor Ranges
Demo: Building a Design for Historic Data (part 1/2)

Watch* how one can build a one-factor response surface method (RSM) historical data design using Design-Expert® software:

1. From the "Response Surface" tab choose "Historical Data".
2. For "1" numeric factor called "X-input", enter Min = "150", Max = "200" and Rows = "50". Then "Continue >>"
3. Enter "2" responses: "Y-unlimited" & "Y-limited" & "Continue"

**Later, refer to these instructions to try this yourself!*

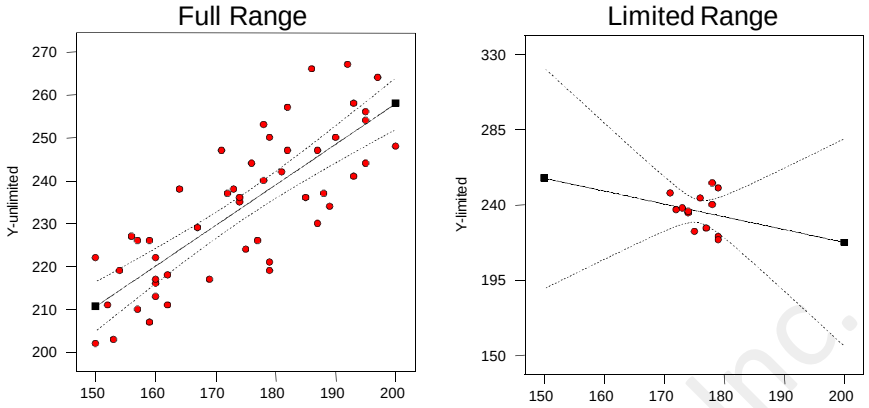
Limited Factor Ranges
Enter Data and Analyze

Copy historical data from Microsoft Excel:

- Open* "Limited factor range.xls".
- Copy and paste "X-input", "Y-unlimited" and "Y-limited" data from Excel to Design-Expert.
- Analyze "Y-unlimited" using suggested linear model.
- Analyze "Y-limited" by force fitting a linear model.

**File available on request so you to try this yourself later!*

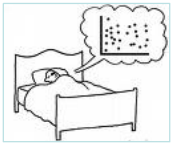
Limited Factor Ranges
Full Range and Limited Range



Double whammy when range limited:
- *Sample size n reduced => power reduced*
- *Signal generated is far less => hard to see (given same noise)*

The solution – work offline so range can be expanded and/or collect more data (increase n).

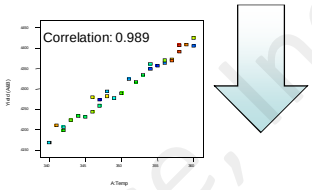
Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #3



- Problems Analyzing Historical Data
  1. Inconsistent data
  2. Limited factor ranges
  3. **Collinearity**
  4. Nonsense correlation
  5. Serially correlated errors
  6. Dynamic relations
  7. Feedback control
- DOE avoids many problems

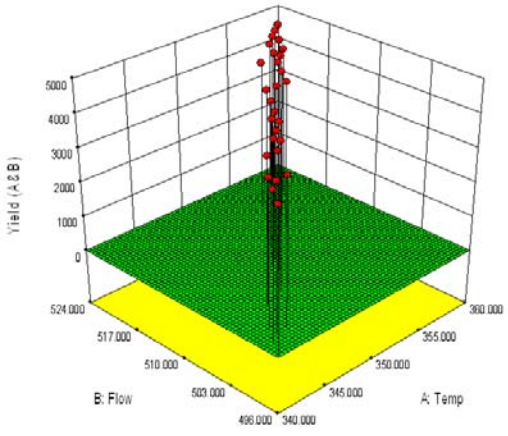
Problems Analyzing Historical Data
Collinearity

Process control systems can compensate for the change in one input by changing another. This creates collinearity, or semi confounding, among the factors. For example, in a continuous flow-through reactor, when temperature increases, the computer increases flow proportionally. In other words, these two factors are highly correlated, that is, collinear. What will happen then if temperature and flow are entered as input factors in the historical design feature for RSM in Design-Expert and yield is recorded as the response?



Collinearity
Raw Data (simulated) -- A "Picket Fence"

Term	StdErr	VIF
A	4.06	184
B	4.07	184
AB	84.83	22206
A^2	43.76	5888
B^2	41.40	5381

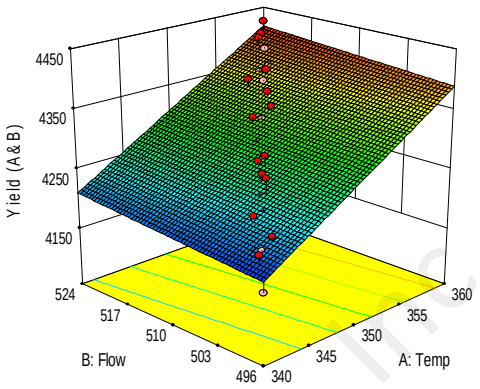


Collinearity
Model with both A-Temp and B-Flow

Adj R-Squared 0.9775
Pred R-Squared 0.9739

Estimated coefficients
+10.28 * Temp
+ 0.97 * Flow

Simulated (true) coefficients
+5.00 * Temp
+5.00 * Flow

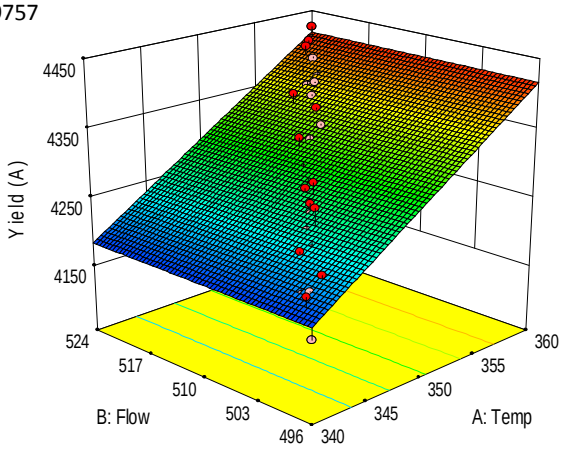


Collinearity
Model with Only A-Temp

Adj R-Squared 0.9782
Pred R-Squared 0.9757

Estimated coefficients
+11.63 * Temp

True coefficients
+5.00 * Temp
+5.00 * Flow

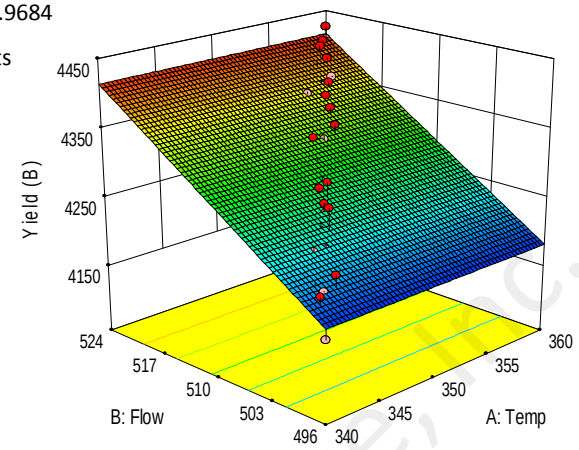


Collinearity
Only Modeling B-Flow

Adj R-Squared 0.9716
Pred R-Squared 0.9684

Estimated coefficients
+8.28 * Flow

True coefficients
+5.00 * Temp
+5.00 * Flow

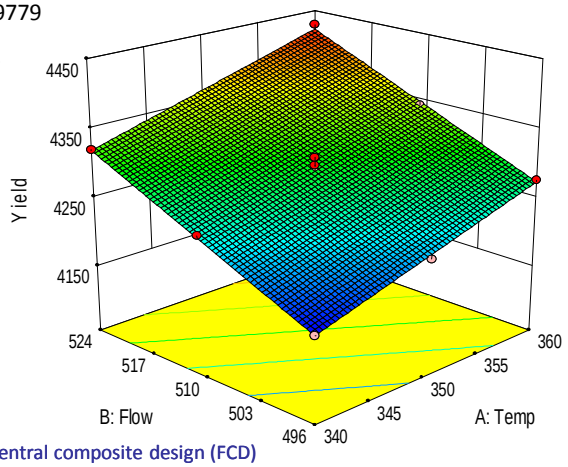


No Collinearity
Achieved by Planned RSM Experiment*

Adj R-Squared 0.9822
Pred R-Squared 0.9779

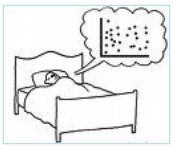
Estimated coefficients
+5.07 * Temp
+5.24 * Flow

True coefficients
+5.00 * Temp
+5.00 * Flow



*From a face-centered central composite design (FCD)

Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #4



- Problems Analyzing Historical Data
  1. Inconsistent data
  2. Limited factor ranges
  3. Collinearity
  4. **Nonsense correlation**
  5. Serially correlated errors
  6. Dynamic relations
  7. Feedback control
- DOE avoids many problems

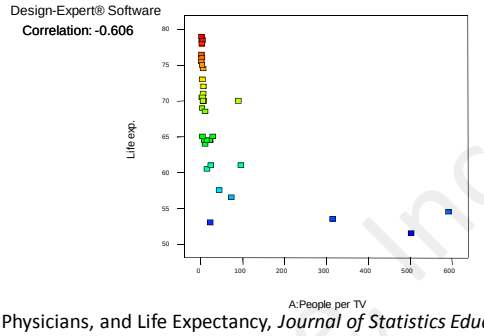
Problems Analyzing Historical Data
Nonsense Correlation

There are always lurking variables – factors that are not observed and sometimes are unknown.

- When analyzing historical data, establishing correlation between response y and factor x does not provide assurance of cause and effect.
- In a DOE randomization, blocking and orthogonal arrays are used to overcome lurking factors.

Nonsense Correlation
Television Improves Life Expectancy?

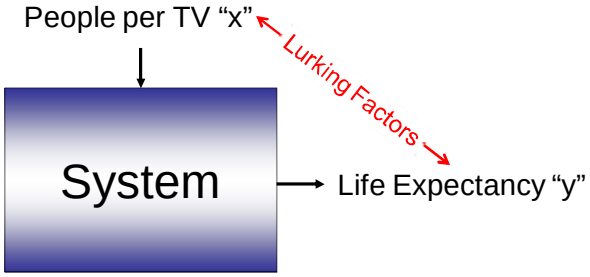
Rossman* provides an enlightening example of nonsense correlation. He observed that life expectancy in various countries (lowest in Ethiopia, Tanzania, Sudan. Bangladesh, Zaire and Myanmar) apparently varies with the number of people per television (TV) set.



* Rossman, Allan. Televisions, Physicians, and Life Expectancy, *Journal of Statistics Education* 2, no. 2 (1994). Also see *RSM Simplified* by Anderson & Whitcomb, pp 40-41.

Nonsense Correlation
Television Improves Life Expectancy?

Should we funnel our foreign aid into boat loads of TVs shipped to third-world countries?

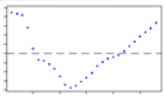


Or – Are there lurking factors that affect both life expectancy and TV purchases?

Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #5

- Problems Analyzing Historical Data
  1. Inconsistent data
  2. Limited factor ranges
  3. Collinearity
  4. Nonsense correlation
  5. **Serially correlated errors**
  6. Dynamic relations
  7. Feedback control
- DOE avoids many problems

Problems Analyzing Historical Data
Serially Correlated Errors

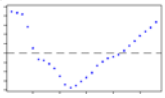


In least squares regression it is assumed the errors (*residuals*) are normal, independent and identically distributed (NIID).

- The errors are normally distributed
- Independent (*not correlated*)
- Identically distributed (*constant variance*)

The IID part is more important than the N (*normality*) part.

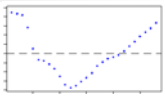
Problems Analyzing Historical Data
Serially Correlated Errors



For sets of happenstance data that are serially collected over time it is reasonable to expect the errors are not independent – they may be autocorrelated:

- Positively – when the error e_t is high, the error e_{t+1} (next time period) also tends to be high (or low-low, that is – in the same direction). See this pictured.
- Negatively – when the error e_t is high, the error e_{t+1} tends to be low (or low-high, that is – opposite).

Problems Analyzing Historical Data
Serially Correlated Errors

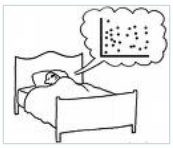


When autocorrelation exists, the estimates of the standard errors of the regression coefficients can be wrong by an order of magnitude! This can be virtually undetectable.

Randomization breaks serial links in the error structure, but this may not be an option (such as in the Longley case).

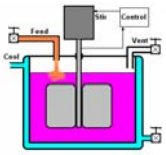
Fortunately, data with serial correlation can often be successfully modeled by including an appropriate time-series model for the errors. For more details, see Chapter 14 in *Statistics for Experimenters.*

Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #6



- Problems Analyzing Historical Data
  1. Inconsistent data
  2. Limited factor ranges
  3. Collinearity
  4. Nonsense correlation
  5. Serially correlated errors
  6. **Dynamic relations**
  7. Feedback control
- DOE avoids many problems

Problems Analyzing Historical Data
Dynamic Relations



In serially-collected data there may be unaccounted-for dynamic relationships. Let's look at a simple example:

- The output of a process is measured every 20 minutes at the outlet from a continuously stirred tank reactor.
- There is one input factor which has a value of x_t at time t and x_{t+1} one interval later.
- The usual regression model $y_t = \beta_0 + \beta x_t + \varepsilon$ implies that y_t at time t is related only to the x_t at that same time.

Problems Analyzing Historical Data
Dynamic Relations

In this situation the output y_t is likely to be dependent on the recent history of x not just x_t. If an exponential memory model is appropriate, then

$$y_t = \beta_0 + \beta(x_t + rx_{t-1} + r^2x_{t-2} + r^3x_{t-3} + \cdots) + \varepsilon$$

where *r* is a weighting factor between zero and one.

With an exponential process the usual regression model

$$y_t = \beta_0 + \beta x_t + \varepsilon$$

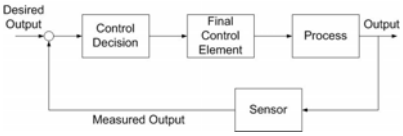
will produce misleading results.

I've used exponential memory models like this for forecasting seasonal sales. Merry Christmas!

Problems Analyzing Historical Data
Seven Issues to Keep You Awake at Night -- #7

- Problems Analyzing Historical Data
  1. Inconsistent data
  2. Limited factor ranges
  3. Collinearity
  4. Nonsense correlation
  5. Serially correlated errors
  6. Dynamic relations
  7. **Feedback control**
- DOE avoids many problems

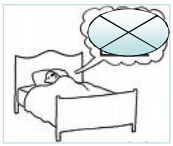
Problems Analyzing Historical Data
Feedback control



Feedback control (*where an input factor is adjusted based on a reading of the output*) creates a relationship between the output and the input. Regressing y on x in this case reflects the programmed control equation (*specifying* how x will be adjusted based on y) and not about how the level of x drives the level of y.

Try to avoid doing this! (That's all we are going to say.)

Problems Analyzing Historical Data
How to get a good night's sleep!



- Problems Analyzing Historical Data
  1. Inconsistent data
  2. Limited factor ranges
  3. Collinearity
  4. Nonsense correlation
  5. Serially correlated errors
  6. Dynamic relations
  7. Feedback control
- **DOE avoids many problems**

Problems Analyzing Historical Data
DOE Avoids Many Problems

1. Inconsistent data – Blocking, randomization and an effort to hold all factors not in the design constant.
2. Limited factor ranges – Experimenter chooses factor ranges and checks power.
3. Collinearity – Use of orthogonal arrays such as factorial designs.
4. Nonsense correlation – Randomization and blocking.
5. Serially correlated errors – Randomization and blocking.
6. Dynamic relations – Measure at steady state.
7. Feedback – Disable control during DOE.

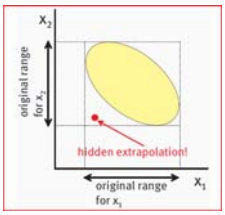
Advantages Analyzing DOE Data
Using Least Squares Regression (Concluding slide)

Least squares regression can handle:

1. Botched design; *for example, if a run is done incorrectly*.
2. Editing factor levels to match what was actually run; *such as when certain factors levels can not be achieved*.
3. Non-orthogonal designs; *for example, when linear constraints on the factors are imposed*.
4. Augmenting an existing design.
5. Calculating diagnostics and influence statistics.

PS. Advice from Montgomery, Peck & Vining*

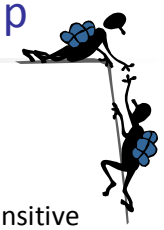
- (3.6) Danger of "hidden extrapolation" where all individual inputs (x1, x2, …) to the regression model are within their observed ranges and yet the coordinate falls outside of it.



- (15.3) Advantages of planned experiments (orthogonal design):
  - Multicollinearity is no longer a problem.
  - Factors can be selected so all important ones are included within "appropriate" ranges.
  - Data collection will be done in a way that minimizes "wild" observations with relatively small measurement errors.

*_Introduction to Linear Regression Analysis_, 3rd Edition, Wiley. (4th edition published 2006).

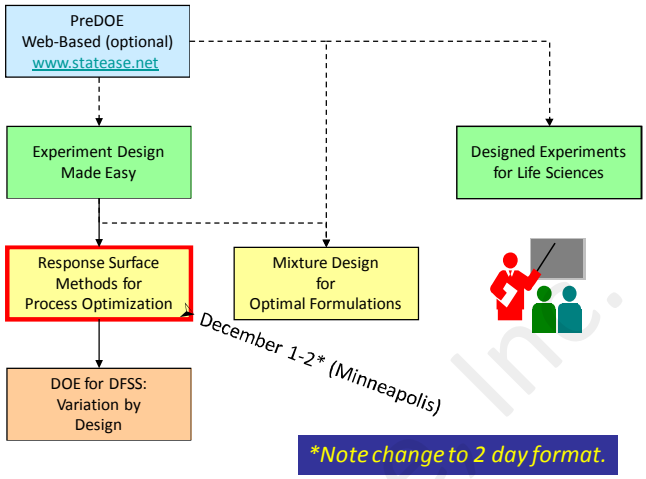
How to get help



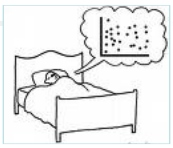
- Search publications posted at www.statease.com.
- In Stat-Ease software press for Screen Tips , view reports in annotated mode, look for context-sensitive Help (right-click) or search the main Help system.
- Explore Experiment Design Forum http://forum.statease.com and post your question (if not previously answered).
- E-mail stathelp@statease.com for answers from Stat-Ease's staff of statistical consultants.
- Call 612.378.9449 and ask for "statistical help."

Stat-Ease Training:
Computer-Intensive Statistical Workshops

Shari Kraber,
Workshop Manager
& Master Statistician
shari@statease.com



Statistics Made Easy®



If all fails for fitting nightmarishly scattered data, there's always the "black thread method"!

Best of luck for your model fitting!
Thanks for listening!

-- Mark
mark@statease.com*

**Pdf of this Powerpoint presentation posted at www.statease.com/webinar.htm.*
*For future webinars,** subscribe to DOE FAQ Alert at www.statease.com/doealert.html.*

**PS. Next webinar may be via a new VOIP system (PC speaker and microphone) with teleconference optional at long-distance phone charges. Stay tuned!