JMP reveals the truth: Statistics can be fun!

William Gjertsen, SAS Institute Inc., Cary, NC

Abstract: JMP is an interactive statistical analysis program that promotes the visualization and discovery of important patterns and outliers in your data. It accomplishes this through unique graphics and the dynamic linking of points. JMP's charter is to harness both statistics and graphics together at one time so that data analysis becomes understandable, enjoyable, and even fun! The use of leverage plots, desirability and interaction profilers, parallel axis graphs, Gabriel bi-plots, the interactive linking of graphs, and the use of animation, magnifying. lasso, and slider tools will be shown with examples from science, engineering and finance.

Introduction: JMP is to data analysis what X-rays and Ultrasound images are to medicine. Sall (1990) in describing leverage plots gives this analogy, "consider the case of a doctor who can learn a lot by examining the patient from the outside, but a series of X-ray pictures that show how things look inside is often essential for the best diagnosis. What the linear-model doctor needs is an inside picture of the data - one that shows how each observation functions in producing parameter estimates and test statistics". This analogy in fact extends to other design features in JMP. It is the interactive and revealing nature of JMP that many times makes it an excellent tool for data diagnosis.

Example 1: The cautious bull (Dynamic Linking)

An investor owns an index fund that is indexed to the Dow Jones Industrials. He is a cautious optimist and is willing to sell part of his fund whenever there's an indication of a downturn. He uses the CumHST oscillator proposed in Daughtry, Pinder, Tang, and Gjertsen (1995) as his technical indicator Figures 1a, 1b, and Table 1. By plotting both the indicator and the Dow Jones index simultaneously any buy or sell signals generated by the indicator get reflected in the plot of Dow Jones Industrials. As an extension if he had multiple indicators then by using different colors and markers he could look for confirmations in signals. The signals in a period from October 1995 to March 1996 is shown. Here we only have four round-turn transactions. The first and third result in very small losses reflecting the cautious nature of the investor. The second and fourth represent substantial gains (see Table 1).

Example 2: Am I Normal? (Lillifors confidence bands, magnifying tool, kernel slider, stem & leaf plot)

Figure 2a represents an x-y measurement for 200 chips on a wafer. The engineer wants to perform capability analysis to determine if these circuits meet specification. Is this x-y dimension normally distributed? Figure 2a gives no real indication one way or the other but Table 2 shows that the answer is emphatically "no", in fact there are less than 7 chances in 1000 that this data is coming from a normal population! We also have our first clues from both the smooth curve (with kenel smoother of .000683) and the normal quantile plot of why it's non normal. In the region from 1.495 to 1.5 there appear to be substantially more values than what we could expect from a normal distribution. Using the magnifying glass tool we can "go inside" and see where normality is violated (Figure 2b). By highlighting and coloring we can see the exact point on the stem and leaf (Figure 2c). Strictly speaking, in it's current form this data should not be subjected to capability analysis.

Example 3: Counting T-Cells (Non-parametric density estimation)

There is a tremendous amount of investigation going on worldwide to understand AIDS and the HIV virus. What is the mechanism by which one becomes HIV and what is the transition to AIDS and what markers do we have for measurement? The Hertzenberg Laboratory at Stanford University is a leader in using flow cytometry to address some of these questions. Sample data from their laboratory is shown in Figures 3a and 3b. A flow cytometer separates and counts different kinds of white blood cells in a person's blood sample by the differential scattering of laser light on the sample. A scatter of 30,000 points from a patient blood sample is shown. Here we are plotting the differential laser scattering of CD8 (Helper TCells) vs. CD3 (Killer TCells). Notice in these figures the equal probability (quantile density) contours for counts in each of the sub populations. Using the magnifier and lasso tools we can get a get a good count of how many and where 90% of the high CD3s and high CD8s reside. AIDS patients will have different TCell profiles than those that just are HIV^{+}so investigators can use graphs and contours like these in AIDS related clinical trials as progression of disease markers. Version 3.2 (Macintosh) now reads flow cytometry (FACS) data directly as an import option.

Example 4: Portfolio Pruning and Stock Selection (Matching Variable plots, Gabriel Bi-plots, Mahalanobis outlier detection plots)

For this example put yourself in the shoes of a portfolio manager. The policy he must follow based on company guidelines is to "prune" or sell any stock after a quarter year period if it's relative strength with respect to its original price is less than 1 and buy more of the best performing stock(s) instead. He decides to use a parallel coordinate plot (matching variable analysis) to quickly decide if any pruning needs to take place see Figure 4a. This example shows that for this 13 week period all stocks ended up by at least 6% except for SGP which didn't meet the pruning cutoff criteria. This portfolio manager true to the company's policy sells SGP and buys more LLY stock. Incidentally although this plot is shown only for a few stocks it is particularly useful for portfolios of hundreds or thousands of stocks!

Imagine also that in this same office we have a second portfolio manager whose interest is in semiconductors and computer networking companies. He has information gleaned from Investors Business Daily on these two market sectors (this is actual data for semiconductors on 12/29/94 and for computer networking companies on 12/28/94). He then wants to filter his selection of stocks by looking for interesting outliers. Using the spin plot in JMP and putting x, y, and z on the first three principal components we get a Gabriel bi-plot which depicts the variables (columns) as spokes, the first three principal components as axes, and the stocks as points (Figure 4b). We can immediately see that Cisco and Micron Technology were standouts in that they were high in Return on Equity and Net % Profit Margin but low in %Debt. To look at any other outliers he decides to look at any 5 dimensional outliers (in Net % Profit Margin, Return on Equity, %Debt, Beta, and Undervalued Index) in the jackknifed Mahalanobis distance plot (Figure 4c). Five other stocks, Three Com, Ascend. Cheyenne Systems, Microdyne and Apertus Tech appear for consideration. (Four of these stocks, Cisco, Three Com, Ascend, and Micron Technology all increased more than 200% within a year after the Dec. '94 analysis was done).

Example 5: It varies (Variability plots, Gage R & R report)

Many manufacturers need to make their products and make measurements with gages (i.e., calipers, voltmeters, etc.) The trouble is that these gage measurements are made across production lines, operators, parts (i.e.. wafers or steppers if you are a chip manufacturer), day of the week, and other factors, and each of these entities within the process introduces a possible source of variation measured by the gage. As an example suppose XYZ Corporation produces widgets (parts) and has several workers (operators) producing these widgets. The burning question is if there is significant variation, is it due to part#, Operator, or possibly their interaction. The appropriate Gage R &R analysis is shown in Figures 5a and Table 5. Since part variation is explaining 81.61% of the tolerance, we can easily examine that variation in Figure 5b by just interchanging the specification of part# and Operator. We see that this widget production line is quite variable (Parts 1, 5 and 10 are low) and that part# and the part# x Operator interaction needs attention.

Example 6: Just DOE it! (Prediction Profiler, Scatter Plots of Interactions, Contour Profiler)

A unique mix of prediction profiles, effect and interaction plots and overlay contours makes JMP very useful to a company wanting to do both the generation and the analysis of designed experiments. In this example a chip manufacturer takes multiple chip measurements across several wafers in an attempt to find out what settings should be made for Thickness (thickness of photoresist), PEB (post-exposed bake time), PUDDLE (developer puddle time on the wafer), Time (exposed time), and Focus (lens setting) to meet a tolerance setting for the chips in both the wx and yz directions and yet simultaneously minimize the standard deviation in chip size in both directions. The desirability profiler in JMP will give good, robust values that are reasonably optimal. The engineer can then overlay contours to see the "sweet spot" regions of acceptable designs. In this example we set our desirabilities for the means to 1.38 with a tolerance of + or - .02 and simultaneously want "smaller is better" for the standard deviations. With a full response surface model we get preferred settings for each of the independent variables (Fig 6a). The profiler, together with the interaction plot for SD ckt yz, and a contour plot showing the "sweet spot" region for Thickness vs. Time are shown in Figures 6a, 6b, and 6c.

Examples 7,8: "Maxed out" (Leverage Plots)

Leverage plots for General Linear Hypotheses have been described extensively by Sall (1990). Here we first show the utility of leverage plots with an artificial example from Huber (1981) Figure 7a. If we fit Y with both the linear and quadratic models for X and look at the residual plots they are not revealing observation 6's true outlying nature in Fig. 7c. But a leverage plot for Y vs. X exposes point 6 as being far along the slanted fitted line (Fig. 7d). So whereas residual plots expose interior data points as the vertical deviations from the reference predicted value line the leverage

plot pictures the points along a sloped line which represents that line where the variable is constrained to be in the model. Highly leveraged points are useful for showing outliers that are near the beginning or end of the data and are exposed not with residual plots but with leverage plots.

Lastly we look at a treatment which may or may not increase oxygen consumption. Patients are randomly assigned to treatment "T" or control "C". They are then asked to run or jog 1 mile on a treadmill. Oxygen Consumption (Oxy), Runtime, RunPulse, MaxPulse, RestPulse, Age, Wt. and Sex are recorded. The stepwise model for Oxy vs. all effects except Name is shown with leverage plots for effects. In this Aerobics data set "Max" stands out not so much as an outlier but as a very highly leveraged point for Runtime, and it's collinear partners RunPulse and MaxPulse.

Summary:

Using JMP is a refreshing way to make discoveries (a few mouse clicks) and bring joy back to data analysis (supporting graphs included with the analysis).

References :

Sall, J. (1990). "*Leverage Plots for General Linear Hypotheses*," The American Statistician, 44, 308-314

Daughtry, J., Pinder J., Tang E., and Gjertsen W. (1995) "*Interactive Visualization and Statistical Interpretation of Market Share, Competive Price, Stock and Mortgage Loan Data Using JMP® 3.1*" Proceedings of the 20th SAS Users Group International Conference

Huber, P. (1981), *Robust Statistics*, New York: Wiley

Date	DJI	Cum HST	Gain/Loss
10/12/95	4764.88	17439	
10/25/95	4753.68	16082	-11.20
10/30/95	4756.57	15369	
12/18/95	5075.21	18690	318.64
12/20/95	5059.32	19802	
1/10/96	5032.94	20560	-26.38
1/16/96	5088.22	21087	
4/8/96	5594.37	24443	506.15

Table 1: Trades by the "Cautious Bull"

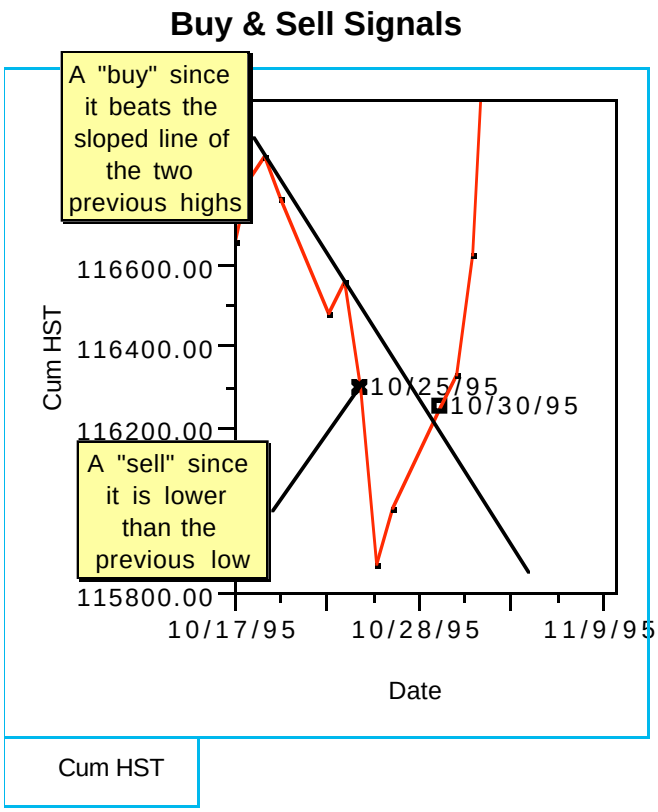

Figure 1a: Buy and sell signals occur whenever the indicator "beats" the previous high (buy signal) or low (sell signal) or "beats" the sloped line of two previous highs (buy signal) or two previous lows (sell signal).

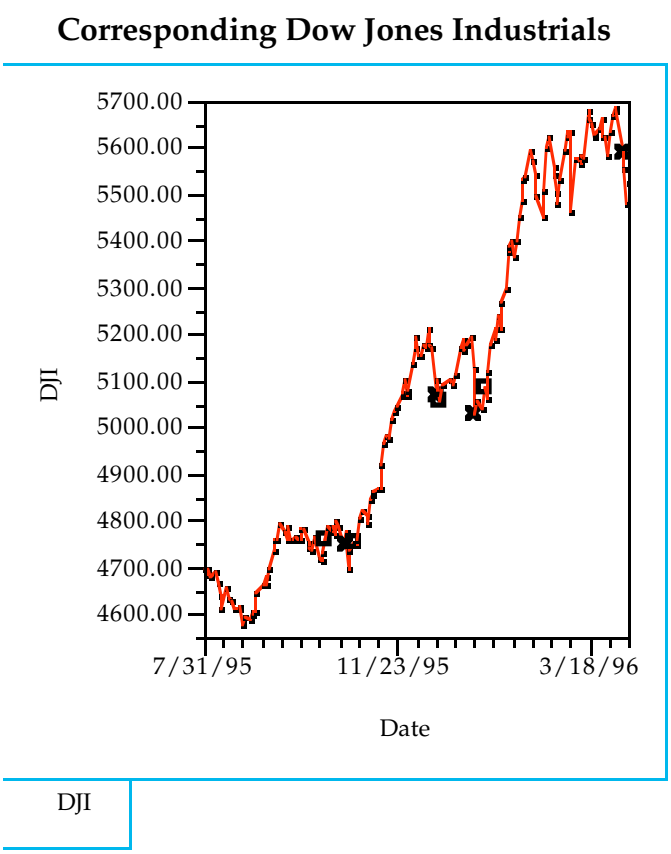

Figure 1b. Due to dynamic linkage we see the result of each buy or sell dictated by our indicator CUMHST.

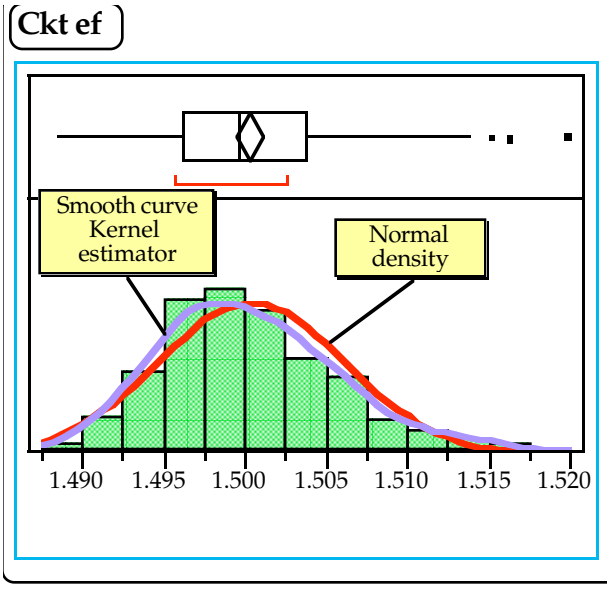

Figure 2a. Is this data normal?

Test for Normality

Shapiro-Wilk W Test

W	Prob<W
0.967302	0.0068

Table 2: There's only a .7% chance that Ckt ef is normally distributed!

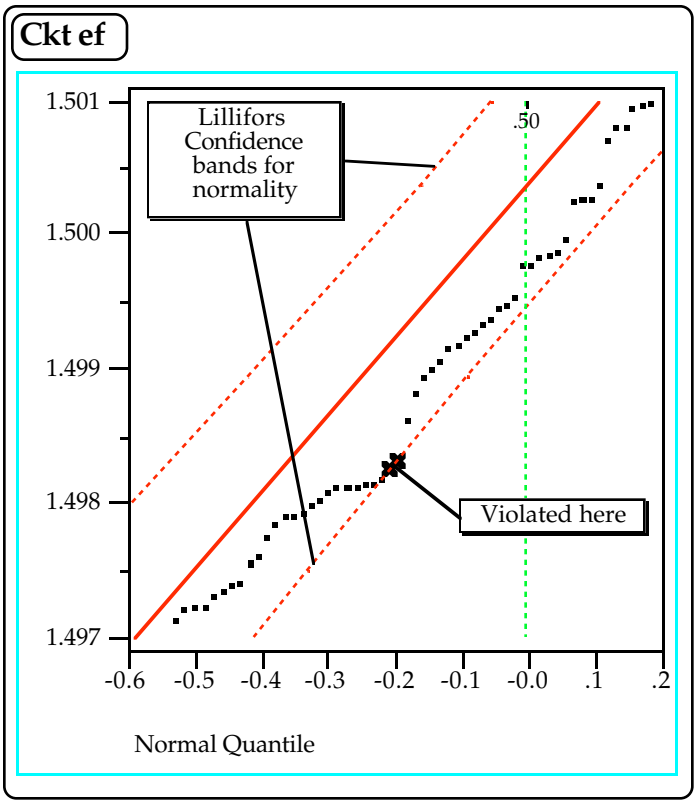

Figure2b. Magnified region of normality failure

Stem	Leaf	Count
1520	0	1
1518		
1516	34	2
1514	2	1
1512	6988	4
1510	112879	6
1508	123414	6
1506	00133611457	11
1504	024477790114555679	18
1502	00113334445578902278888	23
1500	033347880001223344445889	24
1498	00112222233780012234455688999	30
1496	12333556778892223344466899999	28
1494	00123344561123345667888889	25
1492	2233556691137	13
1490	34555	5
1488	575	3

Linkage extends to the Stem and leaf plot

Multiply Stem.Leaf by 10^-3

Figure 2c. Stem & Leaf Plot

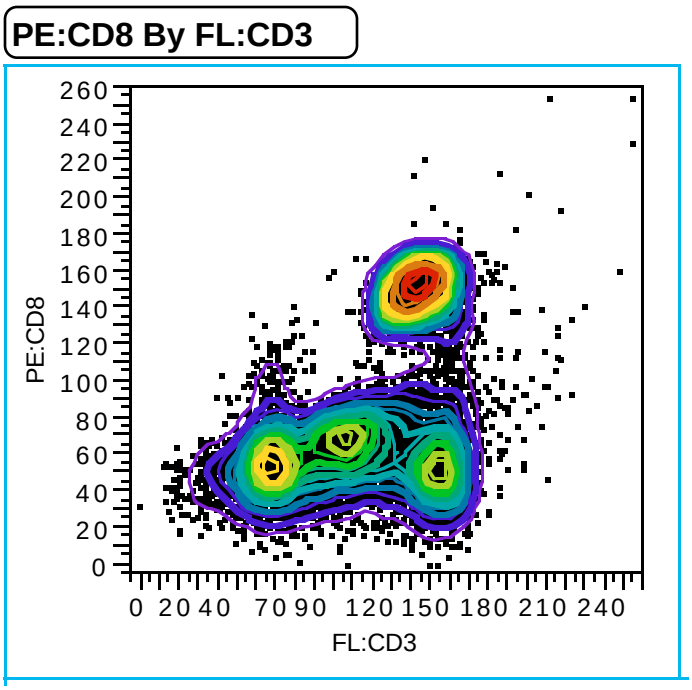

Figure 3a. Plot of CD8(helper) vs CD3(Killer) TCells with equal probability contours

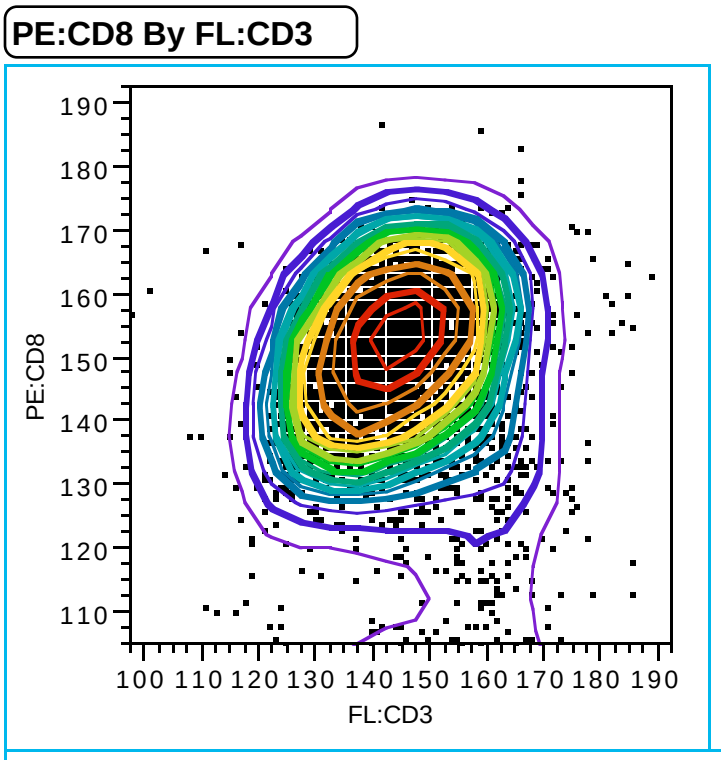

Figure 3b. Using the magnifier tool gives a more detailed view of the high CD3 and high Cd8 subset.

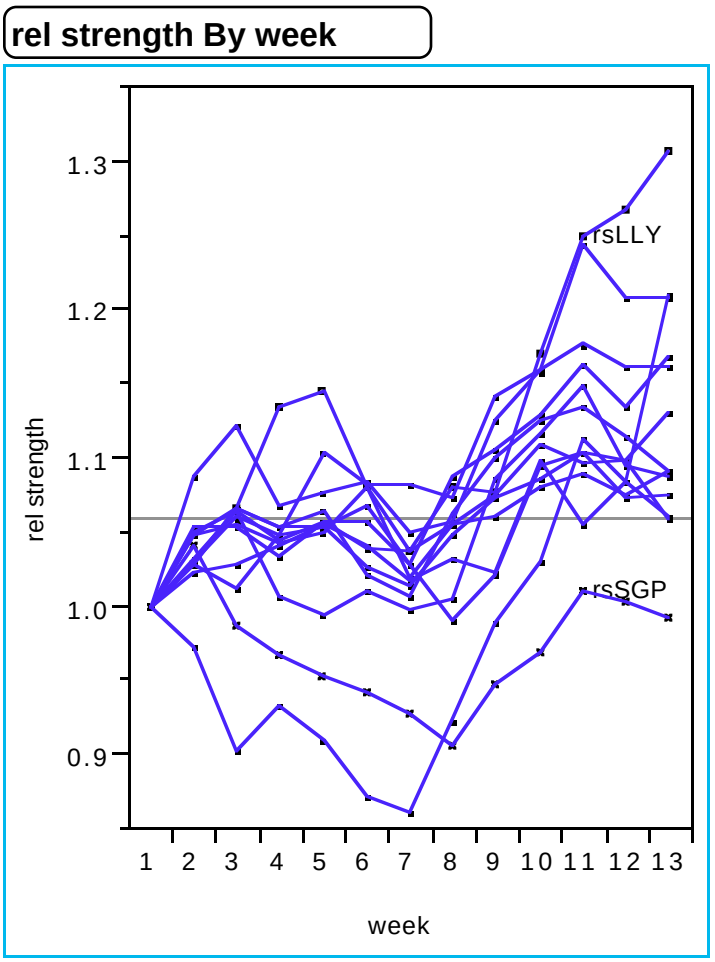

Figure 4a. Porfolio pruning. Prune any stock (like SGP) whose relative strength is less than the mean relative strength for the quarter and buy more strong stocks (like LLY).

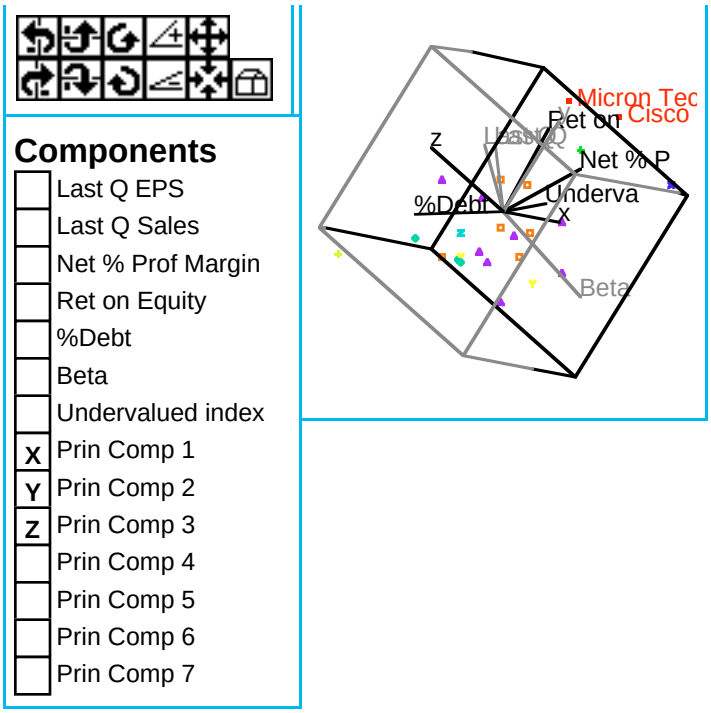

Figure 4b. Cisco and Micron are attractive. They are large in the Ret on Equity & Net % Profit Margin Principal Components.

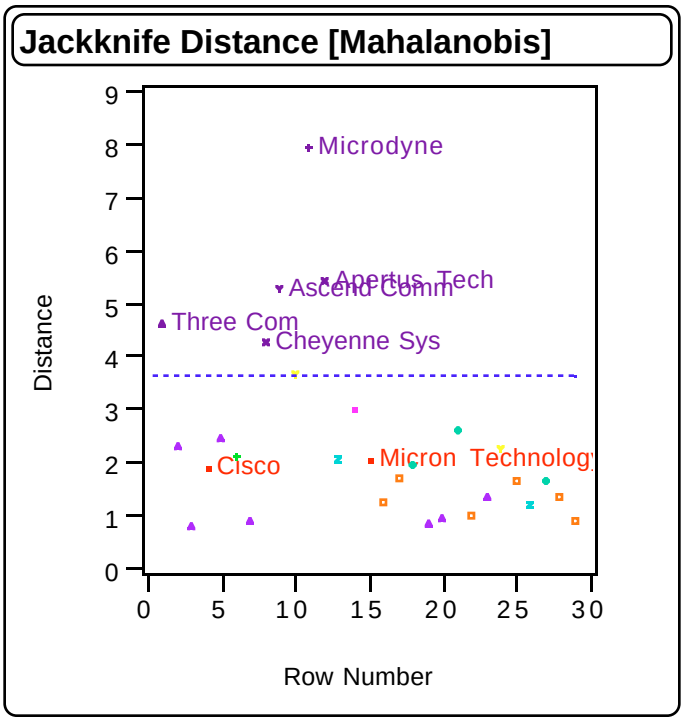

Figure 4c: Outlier Detection Plot reveals more stocks.

Measurement Unit Analysis		% Tolerance
Repeatability (EV)	0.3905181	39.0518
Reproducibility (AV)	0.0751552	7.5155
Operator * Part Variation (IV)	0.4260078	42.6008
Gage R&R (RR)	0.5827824	58.2782
Part Variation (PV)	0.8161442	81.6144
Total Variation (TV)	1.0028592	
Sigma Multiple	5.1500000	
Tolerance	1.0000000	

Table 5: Gage R&R Report showing part# Variation.

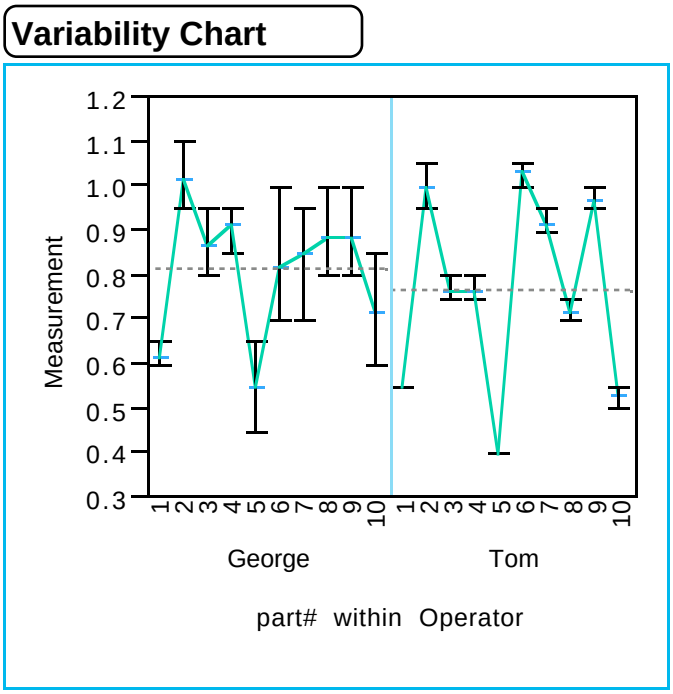

Figure 5a: Variability plot, Operator profiles similar.

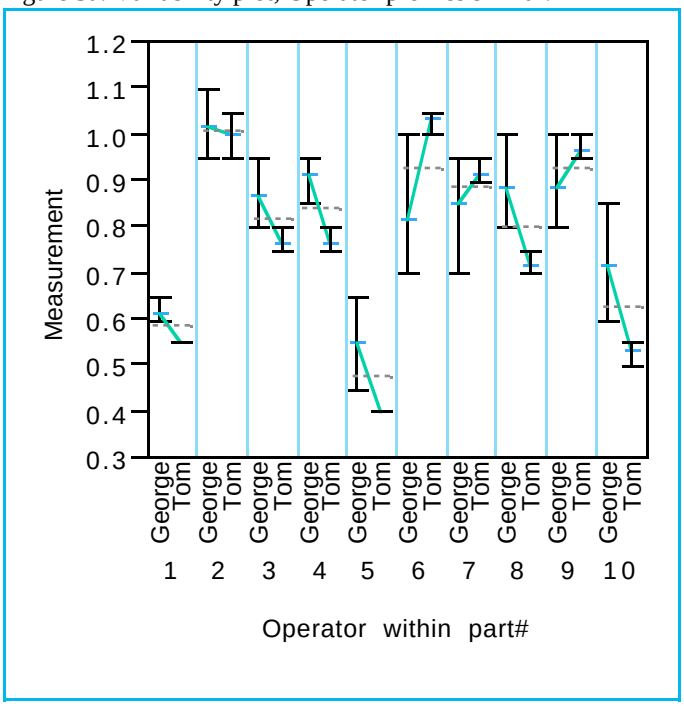

Figure 5b: Variability Plot, parts 1, 5, & 10 are low.

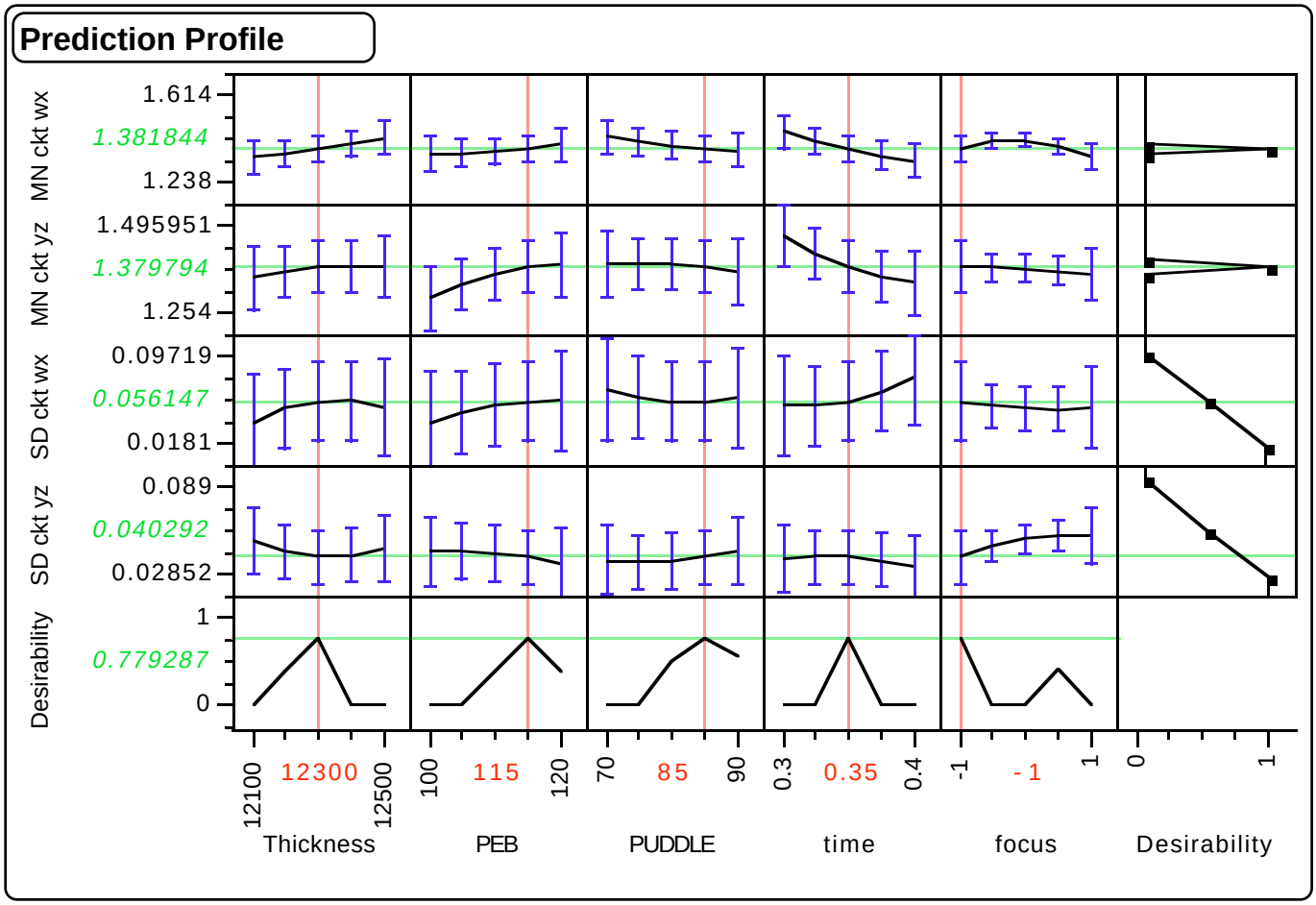

Figure 6a. JMP's Prediction Profiler with desirability settings. Here we see that by setting a tight "most desirable" setting at 1.38 for the means, and a "smaller is better" setting for the standard deviations, the last row in the profiler displays the best combination of factor settings to best accomodate these desirability constraints.

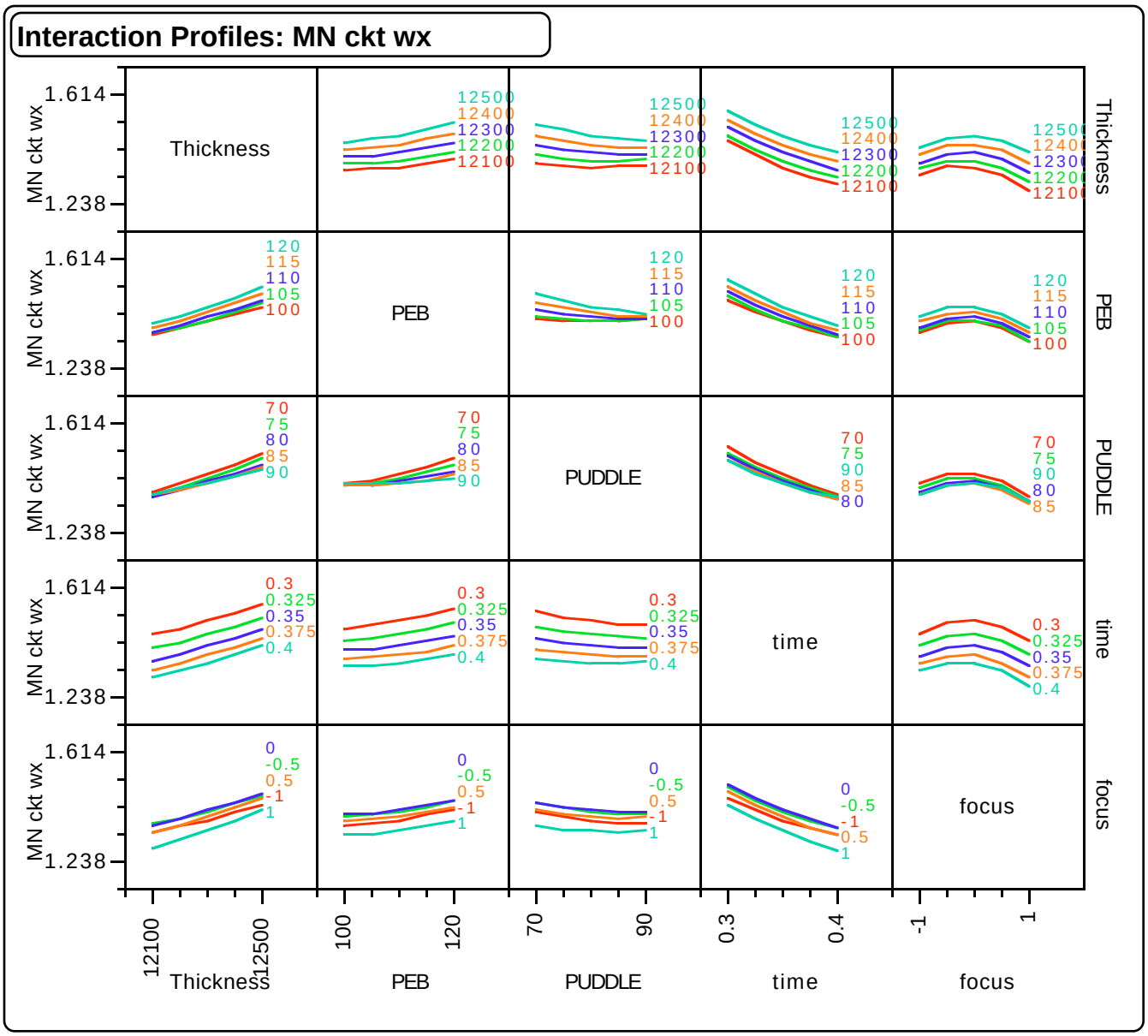

Figure 6b. JMP's two-way interaction profiler. For good, robust designs we want to "design on flats" when at all possible. This will make a factor relatively insensitive to other factor settings and thus minimize variability. In this plot we see that a factor setting of PUDDLE=70 would lead to excess variability with a small change in setting of Thickness, PEB, or time but that PUDDLE=85 designs more on flats and is therefore is preferable as a setting for robust engineering design.

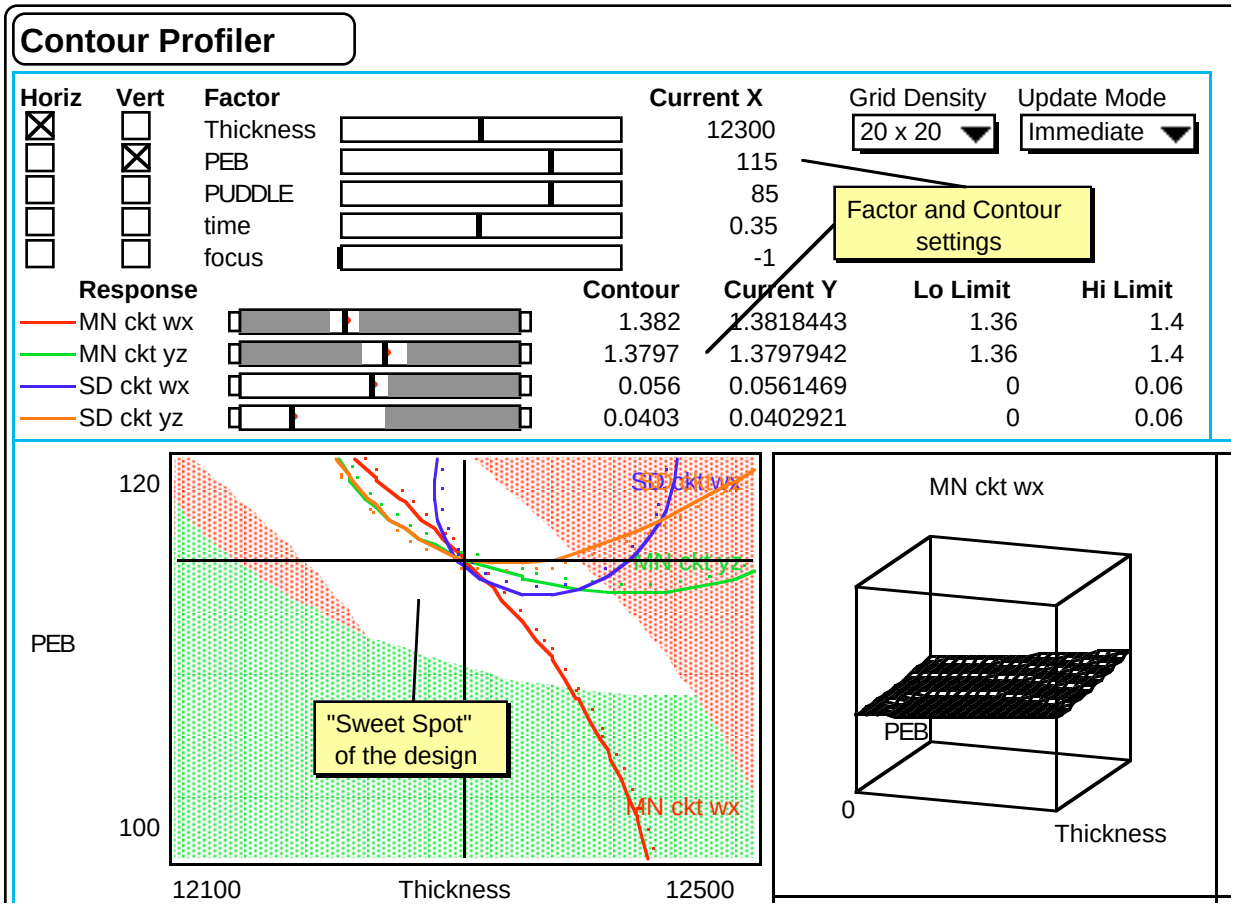

Figure 6c. JMP's overlay contour profiler shows overlayed contours for each response variable for any set of independent variables taken two at a time. The white space ("sweet spot") is the region of robust designs that satisfy the constraints. If minimizinng cost is a consideration moving the crosshair tool to one of the corners of the sweet spot and designing at that corner may be appropriate.

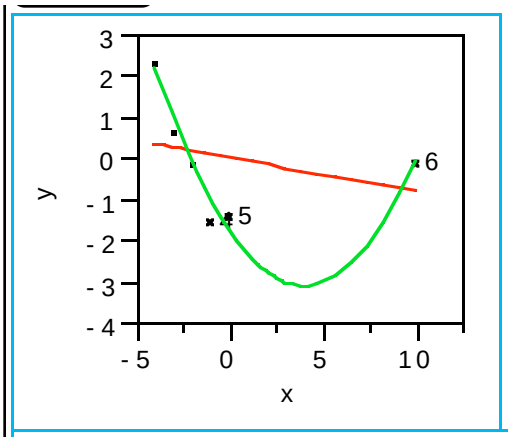

Figure 7a: Linear and quadratic fits to Huber's data

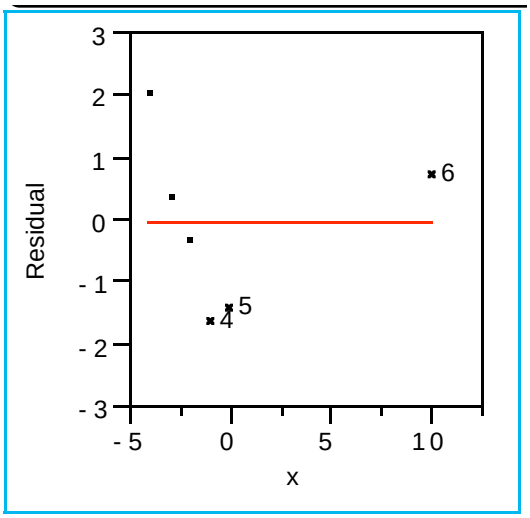

Figure 7b: Residuals form the linear fit showing obsservation 4, not 6, as the largest residual.

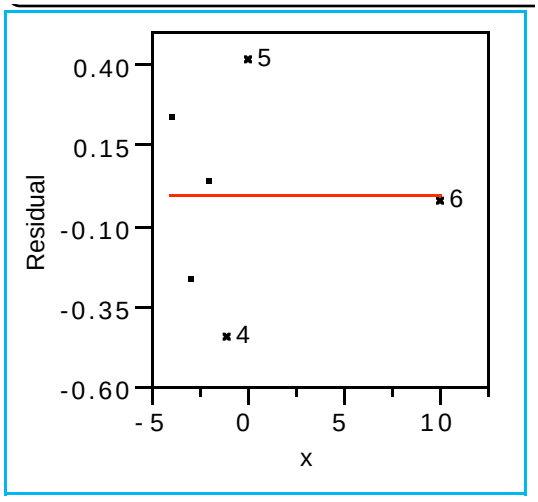

Figure 7c: Residual plot for y vs a quadratic model in x. Note that the highly leveraged point, 6, has a very small residual.

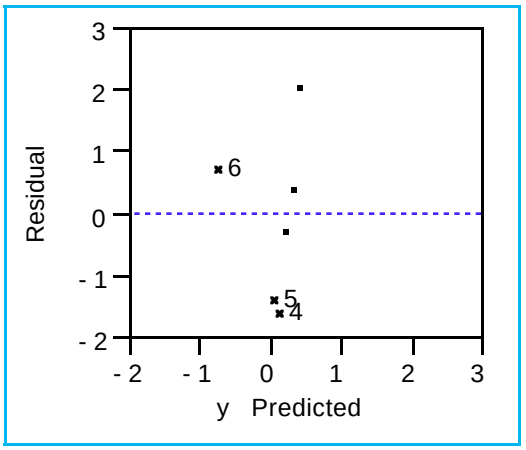

Figure 7d: Residual plot of y vs predicted y also masks the true nature of point 6 for the linear fit of y vs. x.

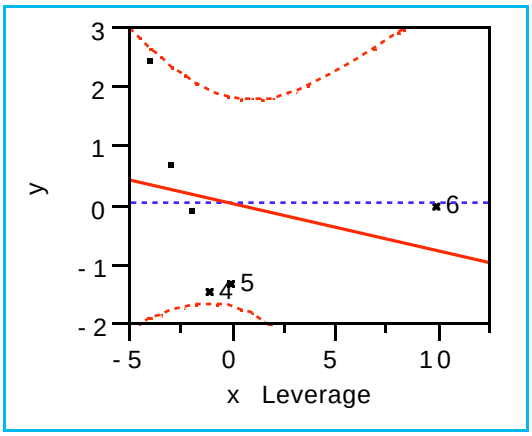

Figure 7e. Point 6's true nature is revealed in a leverage plot. Since it is far out along the x axis it is having a very big effect on the parameter estimate for x and it may cause us to consider an alternative model. In this case the quadratic model was sufficient giving an R^2=.952 and a much more reasonable residual plot (compare figures 7b and 7c).

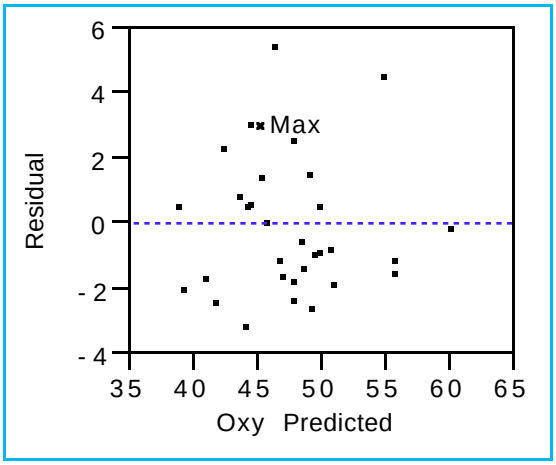

Figure 8a: Max is not an outlier.

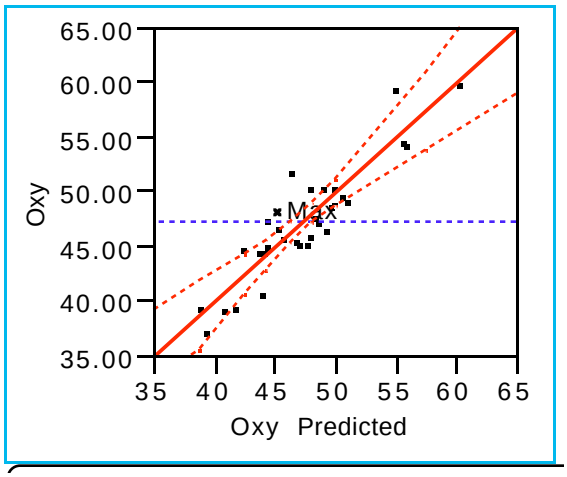

Figure 8b: Max is not offering leverage for the whole model.

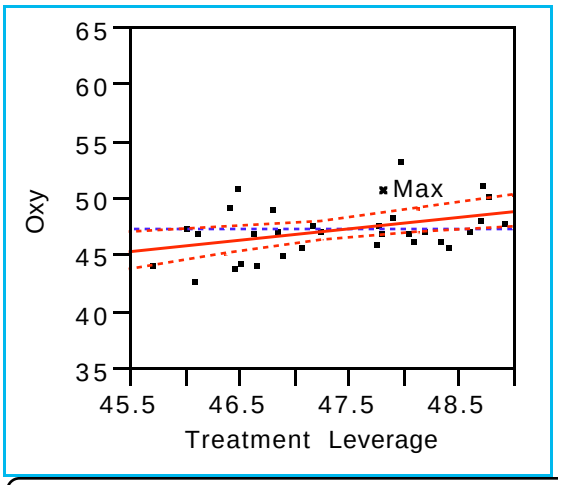

Figure 8c: Max is not offering much leverage for Treatment.

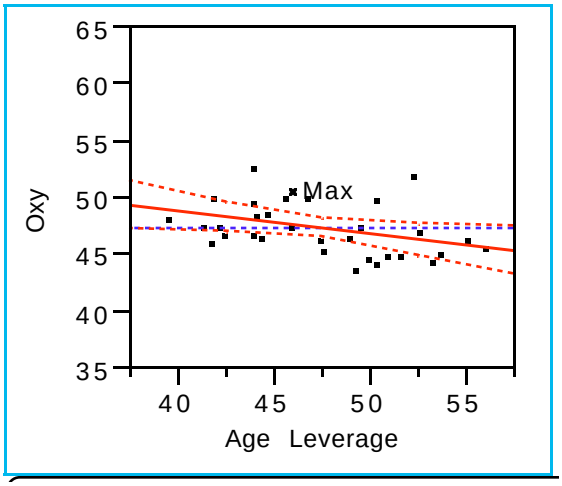

Figure 8d: Max is not offering much leverage for Age.

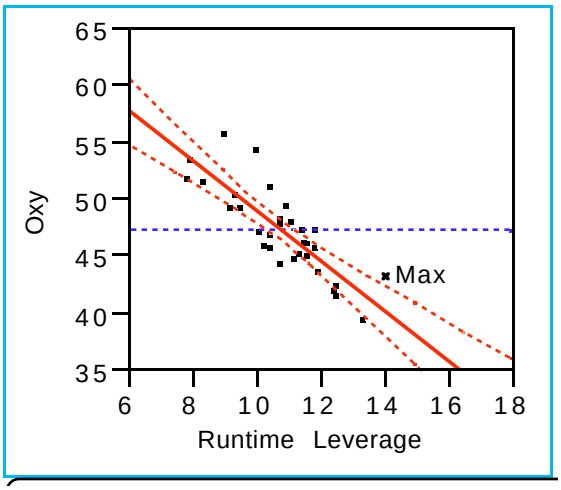

Figure 8e: Max is offering considerable leverage for Runtime.

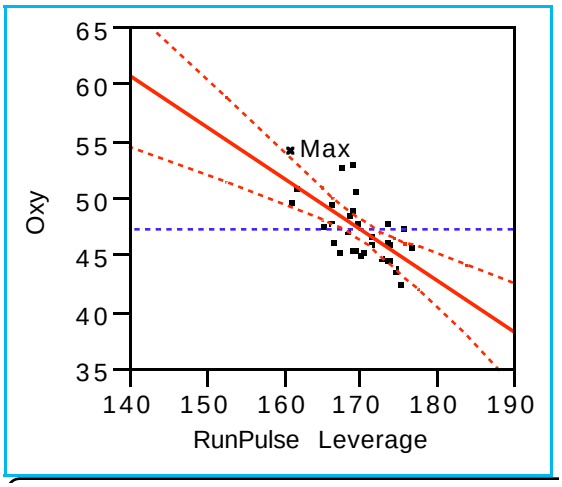

Figure 8e: Max is offering considerable leverage for Runpulse and is causing Runpulse to be even more significant.

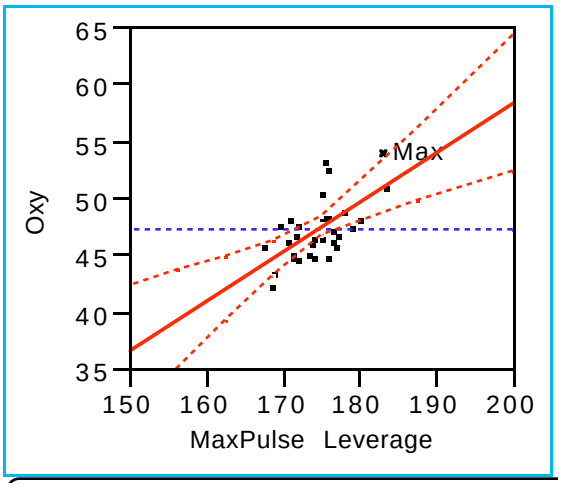

Figure 8e: Max is offering considerable leverage for Maxpulse and is causing MaxPulse to be even more significant. MaxPulse and RunPulse are collinear effects since one is a mirror reflection of the other.