

# Regression

---

Jay Benesch

- Brief history
- context of RF fault analysis
- regression methods via JMP and R
- summary

# History I

- Ordinary least squares regression invented by Gauss ~1795 but first publication Legendre ~1805
- Pearson published chi-squared test 1900
- Deming and Birge remark in 1934 RMP paper “On the Statistical Theory of Errors” that physicists by and large don’t use the chi-squared test. Rev. Mod. Phys. 6, 119–161 (1934)
- My first intro was likely via Phillip Bevington’s 1969 book Data Reduction and Error Analysis for the Physical Sciences (Amazon reviews of third edition suggest avoidance.)
- An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, John R. Taylor, 2<sup>nd</sup> edition 1996 is pretty good.

# Ordinary least squares

- for one variable,  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $S$  = sum of squares of deviations  $\varepsilon_i$
- differentiate with respect to  $\beta$ s, set derivatives equal to zero, solve for coefficients of line
- linear (functional relation) regression: expressions are linear in  $\beta$ s. Independent variables can enter with any functional dependence.

# History II

- In the 70's, as computers advanced, statisticians devised alternatives to ordinary least squares. Among these:
- Robust regression: M-estimator, MM-estimator, least trimmed squares
- Stepwise regression
- Principal component regression
- Devised new tests of normality. The Shapiro-Wilk test seems most widely applied for small samples. Good reference: J.P. Royston (1995) The W-test for normality Applied Statistics **44**, 547-551
- Kolmogorov-Smirnov-Lilliefors (Wikipedia calls it Lilliefors test, variation of K-S) test used for larger samples
- Classic texts: Applied Regression Analysis, 3<sup>rd</sup> edition, Norman R. Draper & Harry Smith; Statistics for Experimenters, George E.P. Box, William G. Hunter, J. Stuart Hunter

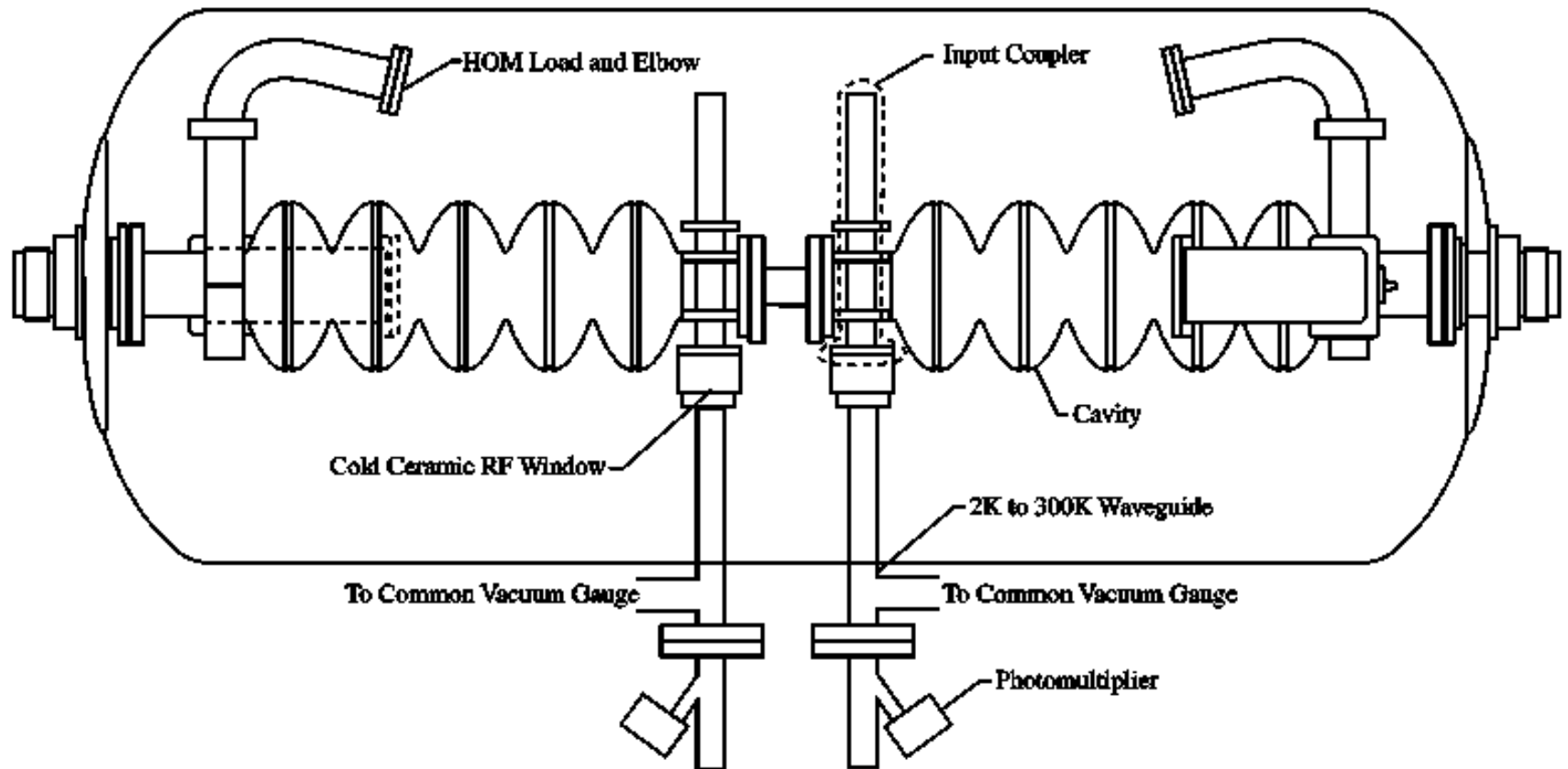
# Robust regression

- Attempts to mitigate the effect of outliers on the result without introducing undue bias for real tail events.
- Generally order residuals, sometimes into quantiles, reweight, and iterate on least squares.
- M-estimator is good at rejecting outliers only on dependent variable (y)
- MM-estimator is more aggressive and can reject outliers on both x and y, but all data still included
- Least trimmed squares orders all residuals by absolute value and throws away high stuff in both independent and dependent variable(s) iteratively until the  $R^2$  is reasonably stable. Most aggressive.
- Wikipedia does a good basic job on definitions.

# Stepwise regression

- Covariance matrix is calculated for a list of independent variables supplied by user.
- Variables are introduced in order of significance (F test) and significance of fit recalculated. If a variable falls in significance after subsequent introductions, it is removed. User can set F values for entry and exit.

# Old style cavity pair



# RF Fault database

- “true arc” fault: simultaneous light and vacuum signals
- old fault logger data in my possession begins 1/30/95. 130447 true arc faults ending 12/20/2002.
- new fault logger data begins 1/1/2002. ~400K cavity faults since 1/1/2003 of which 294420 are “true arcs”.
- 424867 true arc faults available for analysis. This is the entire population, not a sample from a population.
- Two analysis tools: JMP, [www.jmp.com](http://www.jmp.com), an exploratory data analysis package from SAS, and
- R [www.r-project.org](http://www.r-project.org) a comprehensive environment for doing statistics. Started as an open source clone of S, developed at Bell Labs at the same time as Unix and C, because AT&T wouldn't license S.



# JMP

- John's Macintosh Project: the result of a sabbatical by John Saul to see if the Mac interface allowed a new type of exploratory data analysis. Limited in data set size to about a third of your available RAM to reduce impact on SAS's original product. Inexpensive for students; \$1320 for first year and about half that annually thereafter for others.
- Great for trying things out and figuring out how best to deal with repetitive analyses. Has a comprehensive scripting language so stuff can be automated after methods are tested.
- The RF fault data analysis methods were developed in spreadsheets and JMP from 1995-2002 and then turned into code using R by Michele Joyce to my requirements.

# R

- What every statistician uses
- FOSS
- Many other fields use it to some extent. Finance uses it and there once were lots of jobs for R-mongers there. I haven't subscribed to stat journals for a while, so now??
- It's a language and a development environment, not just a collection of modules.
- At least half a dozen texts available FOSS on line
- One connection to ROOT I've found: Adam L. Lyon  
“Analysis of Experimental Particle Physics Data in R with the RootTreeToR Package”  
<http://user2007.org/program/presentations/lyon.pdf>

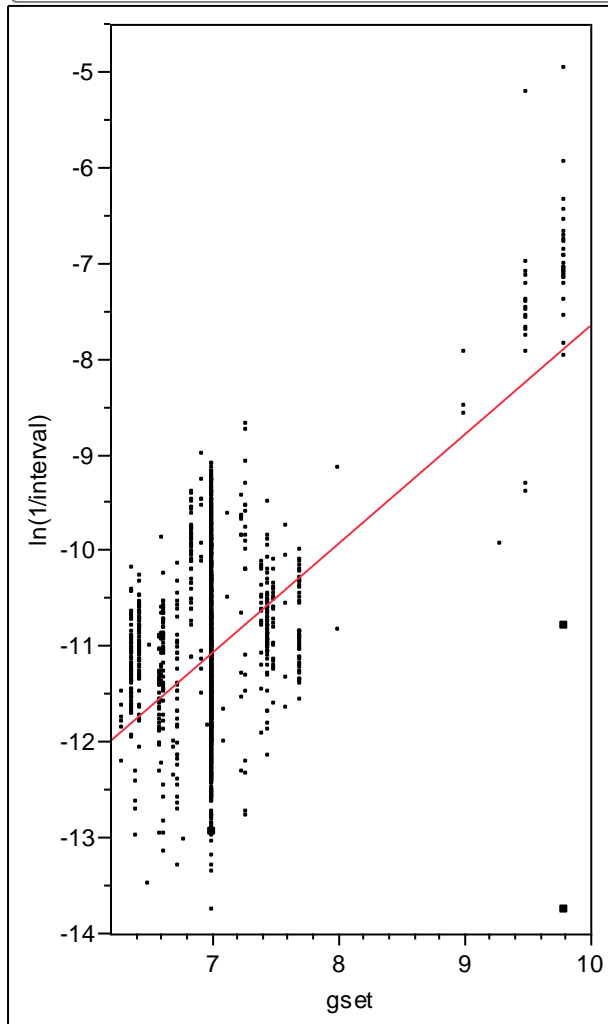
# for even more verbose discussions

---

- TN 95-059
  - TN 98-045
  - TN 01-020
  - TN 05-057
  - TN 10-008
- 
- and if you want to know why Ops needs to improve machine matching radically in the 12 GeV era: TN 05-074.
  - <https://jlabdoc.jlab.org/docushare/dsweb/View/Collection-11306>

# Cavity 0L03-1 2003-2012

## Bivariate Fit of $\ln(1/\text{interval})$ By gset



— Linear Fit

### Linear Fit

$$\ln(1/\text{interval}) = -19.03937 + 1.1404755 \cdot \text{gset}$$

### Summary of Fit

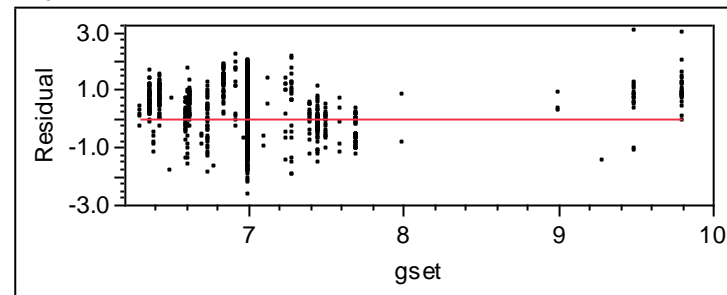
RSquare	0.308641
RSquare Adj	0.30831
Root Mean Square Error	0.802138
Mean of Response	-11.0144
Observations (or Sum Wgts)	2092

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	600.3353	600.335	933.0293
Error	2090	1344.7603	0.643	<b>Prob &gt; F</b>
C. Total	2091	1945.0956		<.0001*

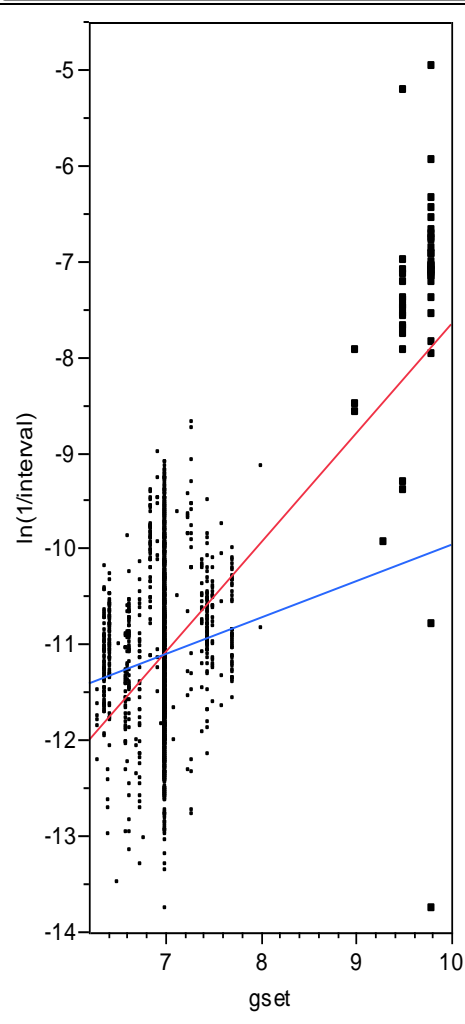
### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-19.03937	0.263305	-72.31	0.0000*
gset	1.1404755	0.037337	30.55	<.0001*



# Cavity 0L03-1 less high gradient

## Bivariate Fit of ln(1/Interval) By gset



### Linear Fit

$$\ln(1/\text{interval}) = -19.03937 + 1.1404755 \cdot g_{\text{set}}$$

### Summary of Fit

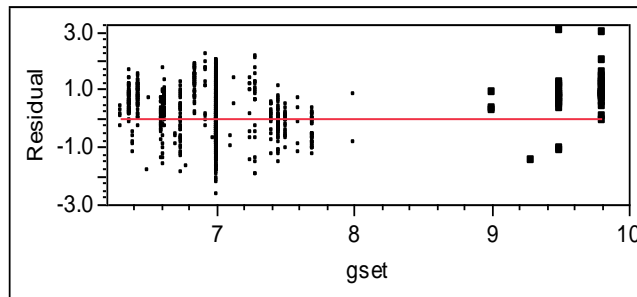
RSquare	0.308641
RSquare Adj	0.30831
Root Mean Square Error	0.802138
Mean of Response	-11.0144
Observations (or Sum Wgts)	2092

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	600.3353	600.335	933.0293
Error	2090	1344.7603	0.643	<b>Prob &gt; F</b>
C. Total	2091	1945.0956		<.0001*

### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-19.03937	0.263305	-72.31	0.0000*
gset	1.1404755	0.037337	30.55	<.0001*



### Linear Fit

$$\ln(1/\text{interval}) = -13.73613 + 0.3776977 \cdot g_{\text{set}}$$

### Summary of Fit

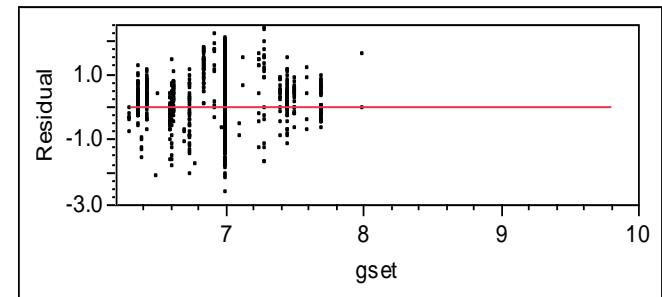
RSquare	0.014804
RSquare Adj	0.014321
Root Mean Square Error	0.771251
Mean of Response	-11.1015
Observations (or Sum Wgts)	2044

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	18.2513	18.2513	30.6833
Error	2042	1214.6403	0.5948	<b>Prob &gt; F</b>
C. Total	2043	1232.8916		<.0001*

### Parameter Estimates

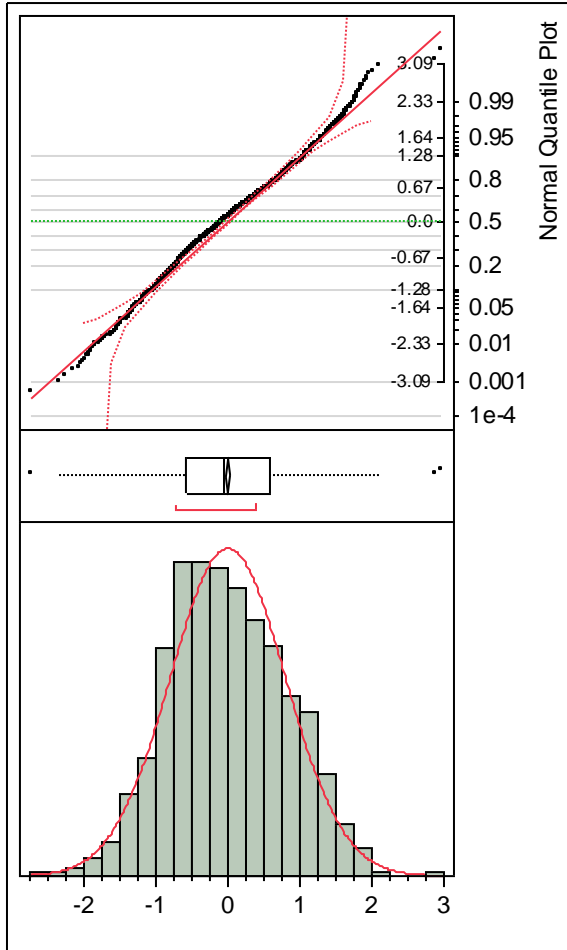
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-13.73613	0.475936	-28.86	<.0001*
gset	0.3776977	0.068186	5.54	<.0001*



# test for normality 0L031 residuals

## Distributions

### Residuals ln(1/interval)



#### Quantiles

100.0%	maximum	2.9845
99.5%		1.86795
97.5%		1.57099
90.0%		1.09443
75.0%	quartile	0.5913
50.0%	median	-0.0502
25.0%	quartile	-0.594
10.0%		-0.9824
2.5%		-1.4739
0.5%		-1.9121
0.0%	minimum	-2.7208

#### Fitted Normal

##### Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	$\mu$	2.727e-14	-0.034385	0.0343846
Dispersion	$s$	0.8019465	0.7783627	0.8270149

-2log(Likelihood) = 5012.3742825258

##### Goodness-of-Fit Test

KSL Test

D	Prob>D
0.035025	< 0.0100*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Shapiro-Wilk test used by JMP for less than 2000 samples and KSL test for greater numbers. This distribution is rejected at the P=0.01 level.

KSL: Kolmogorov-Smirnov-Lilliefors (Wikipedia calls it Lillefors test, variation of K-S test)

— Normal(2.7e-14,0.80195)

# 0L031 – what else is there?

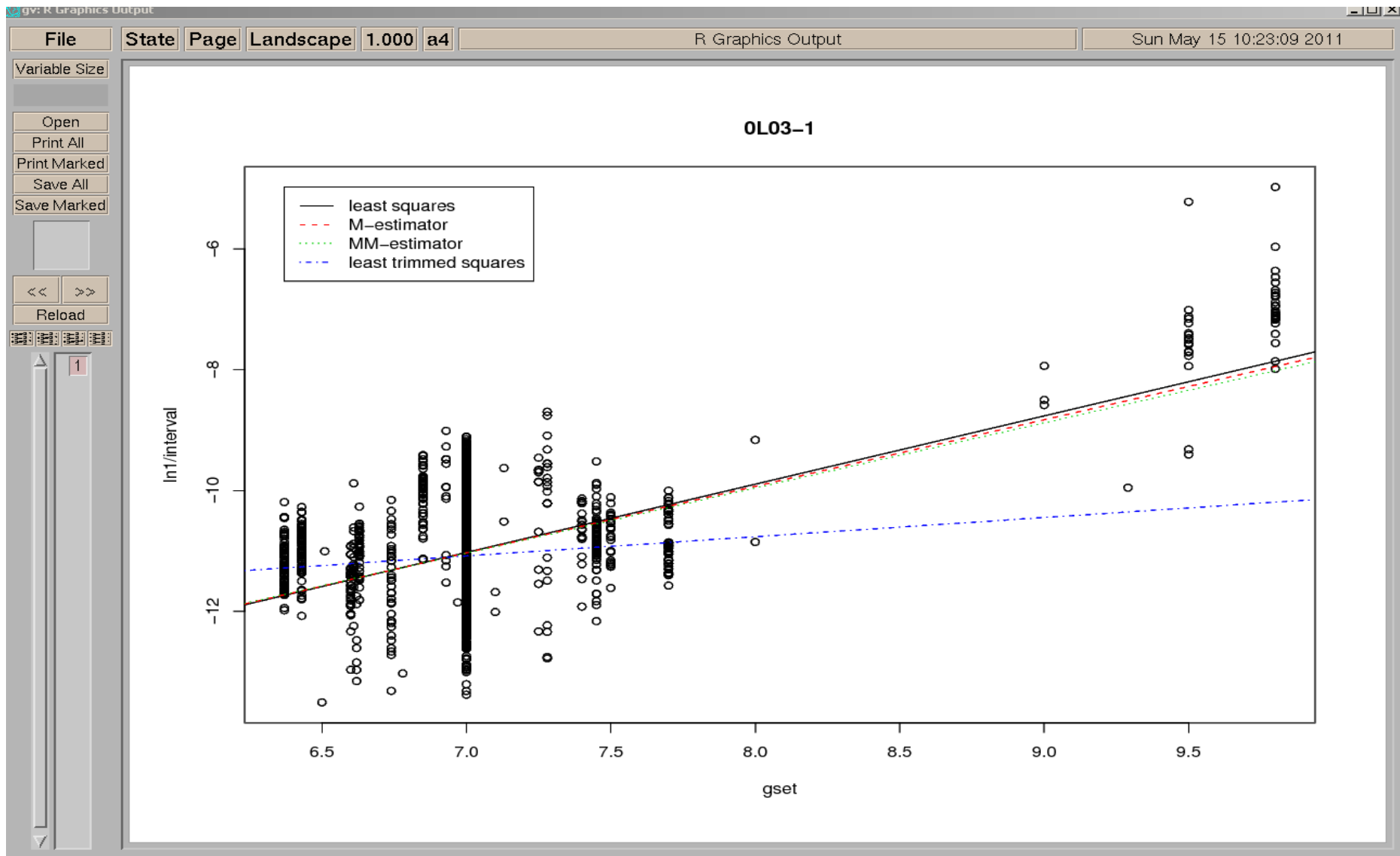
- “fratricide” discovered early on: cavity cold window charged by field emitted electrons from other cavities.
- 2002 fault logger stores gsets from all cavities in zone for each fault
- allows stepwise regression to apportion blame, or in one case to find a wiring fault
- real time demo of stepwise regression with JMP on my desktop PC

# Fault Analysis in R

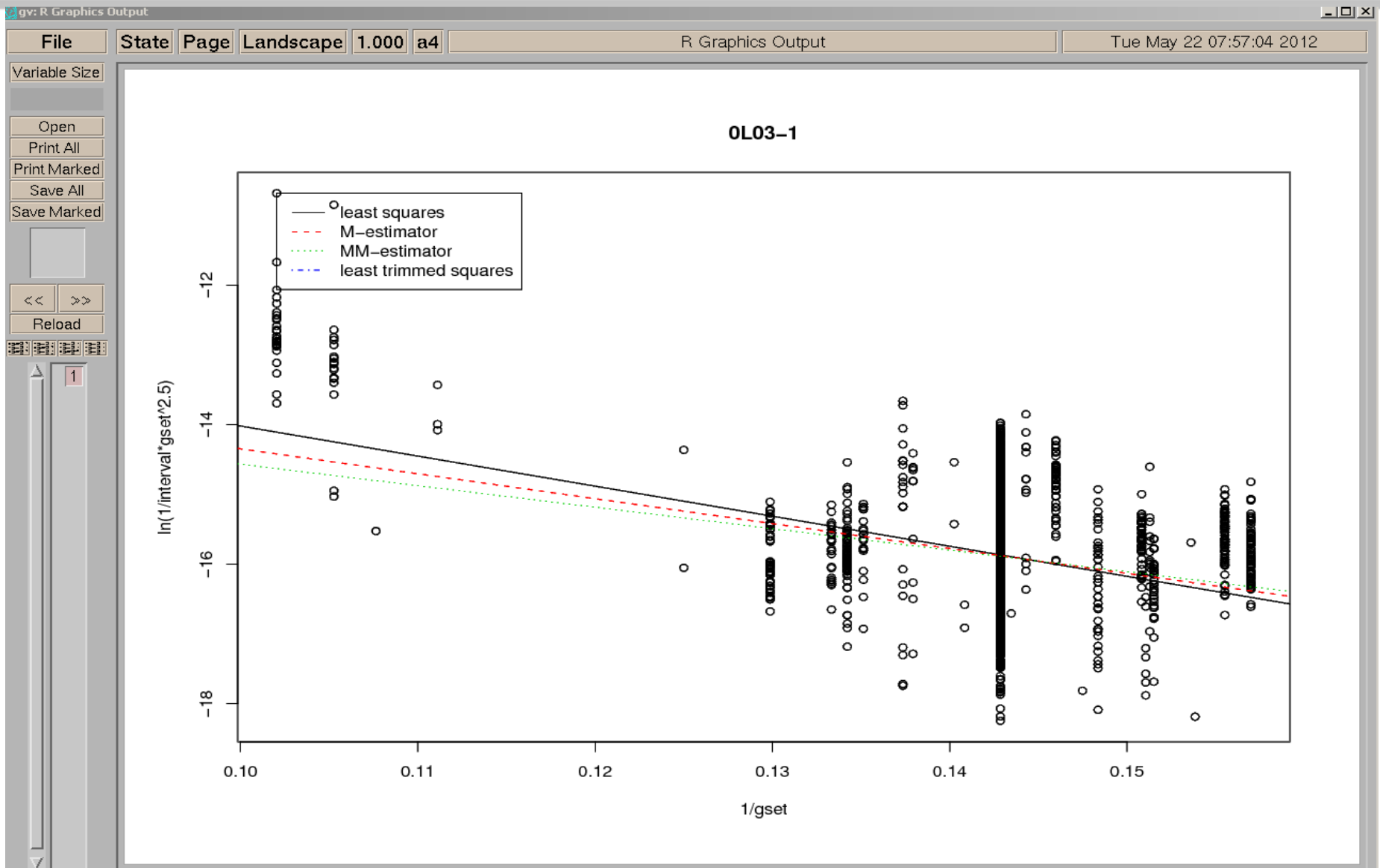
- Coded in ? and R with wrapper. Includes JMP-derived cuts on gradient change and interval duration.
- Fault logger logs RF faults keeping data on ten types. Ops and EES use FaultViewer on few day time scale.
- FaultCompiler compiles fault info, generally “since last event”. Checks via archiver whether a sub-threshold vacuum excursion occurred with an arc-only fault. If so, converts it to “true arc” fault in the database.
- FaultAnalyzer produces graphs and a text file with results of (ordinary) least squares, M-estimator, MM-estimator and least trimmed squares.
- slideShow displays the graphs. Text file formatted for easy spreadsheet viewing.



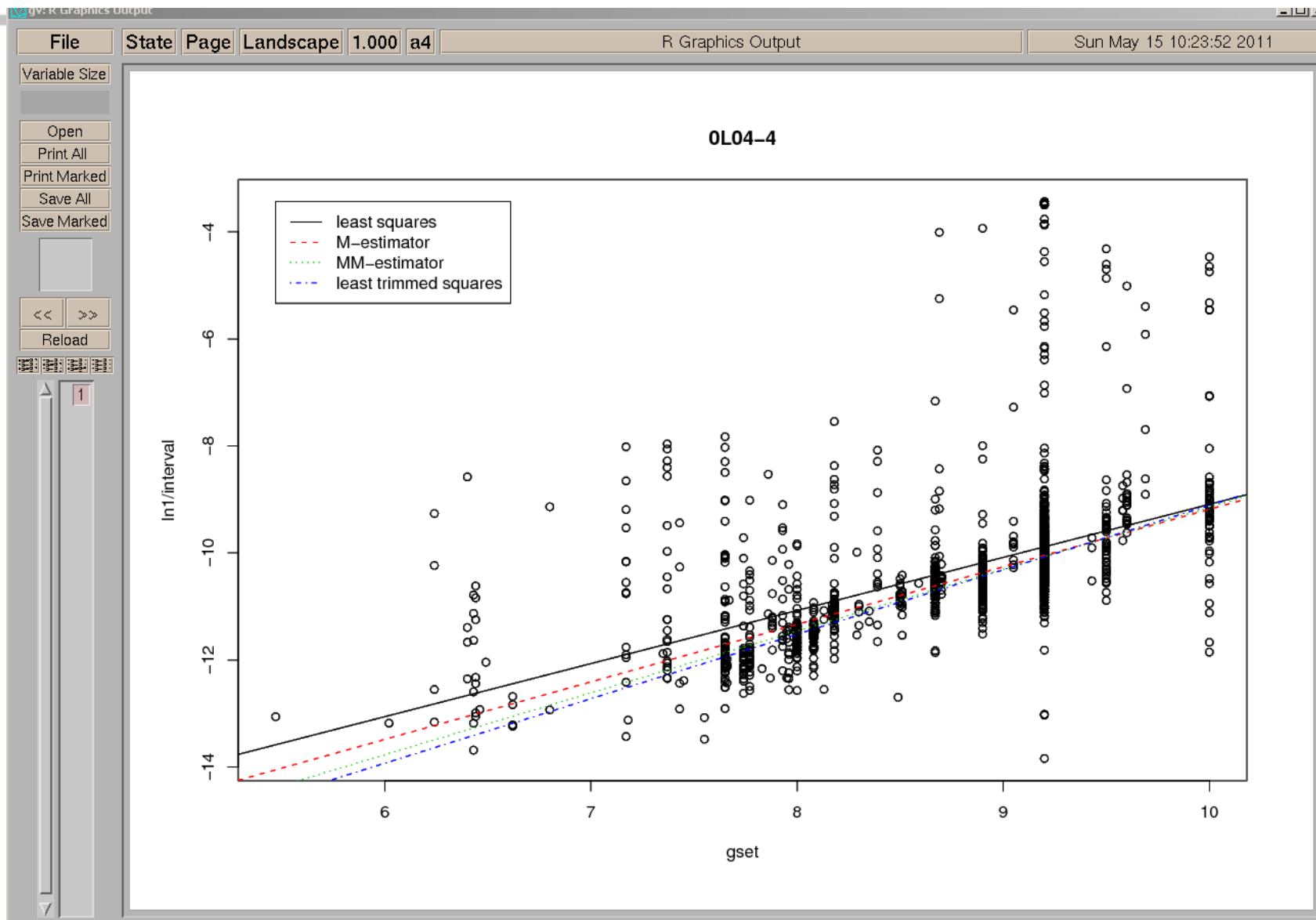
# slideShow 0L03-1



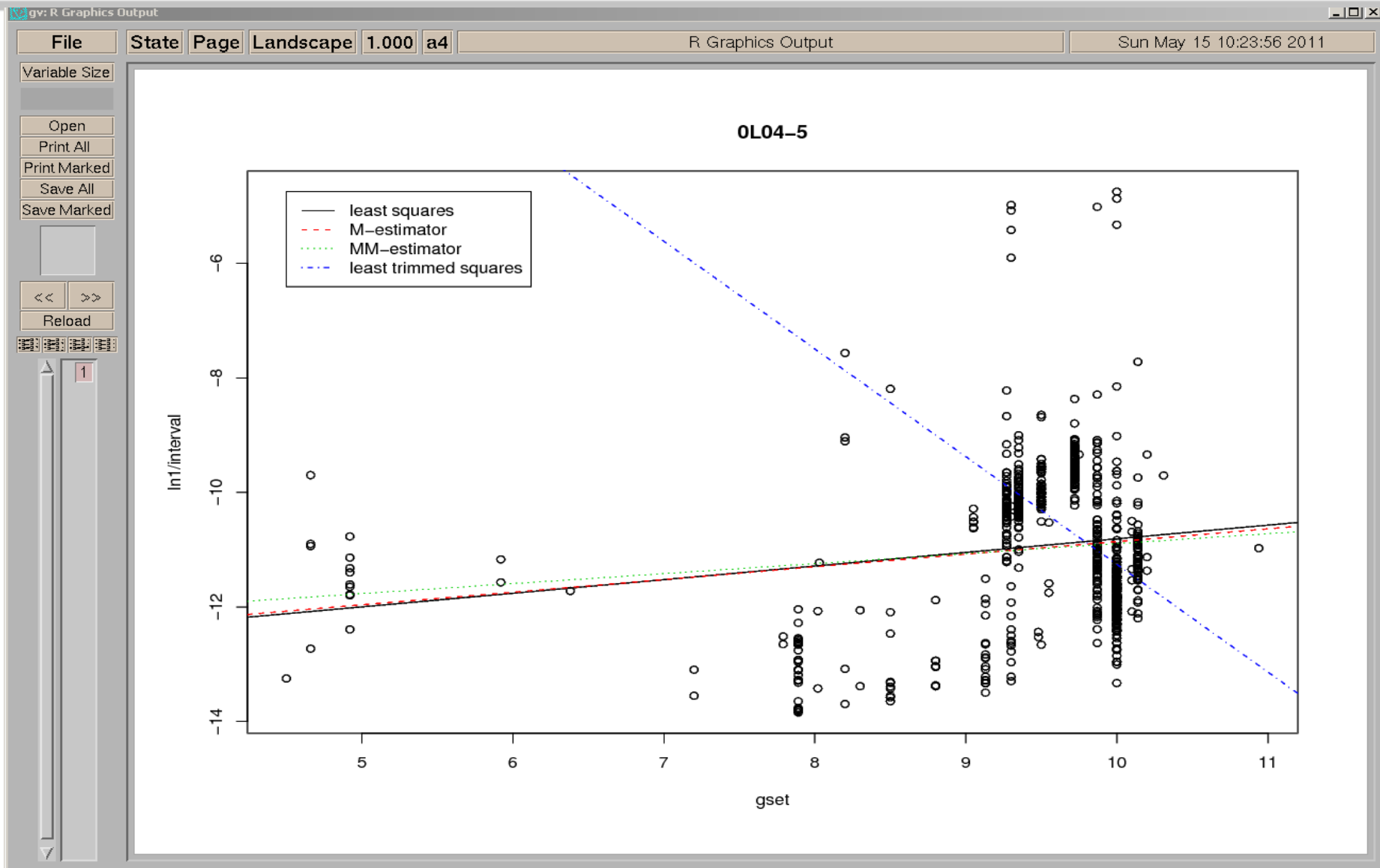
# slideShow 0L03-1 Fowler-Nordheim



# slideShow 0L04-4



# slideShow 0L04-5

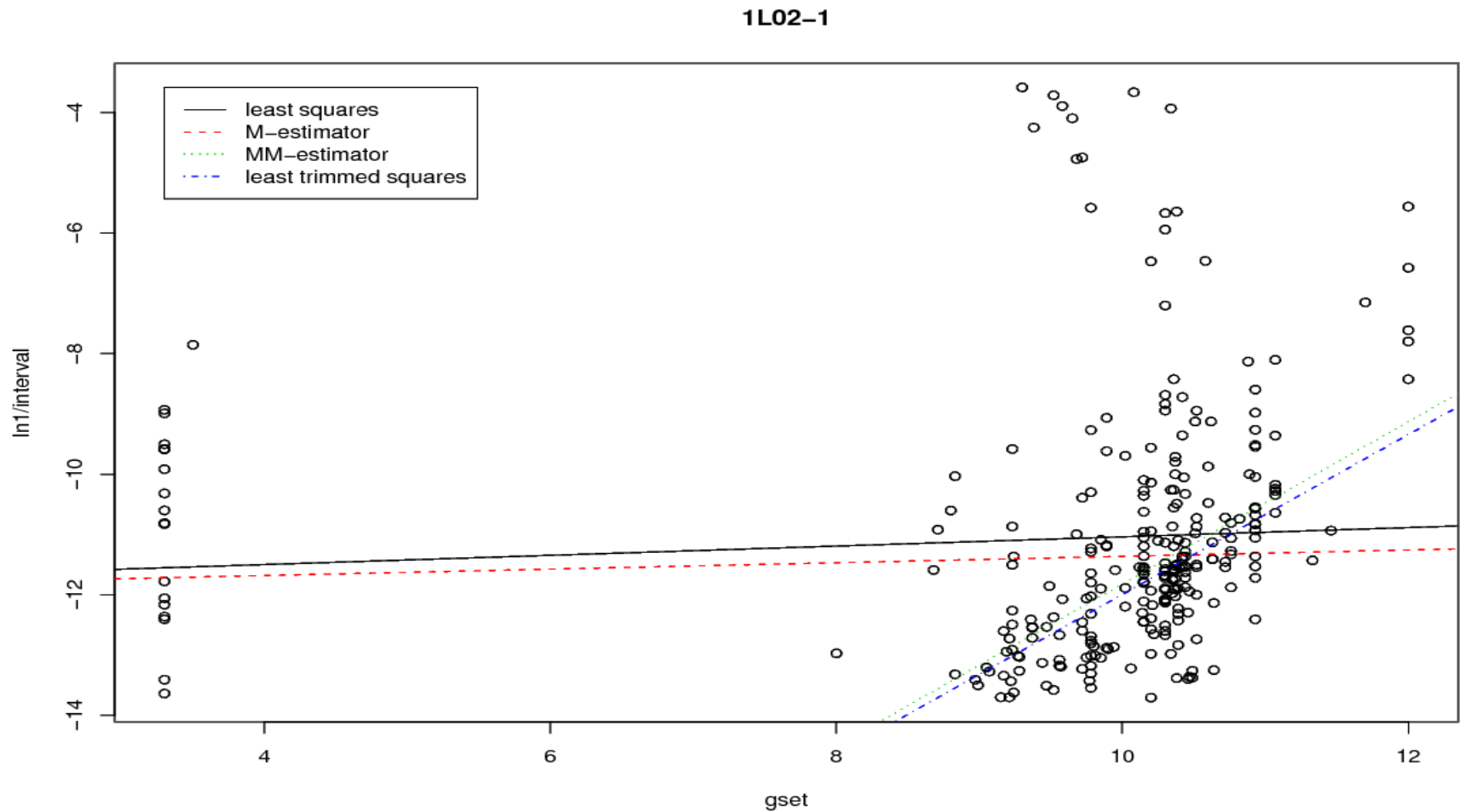


# slideShow 1L02-1

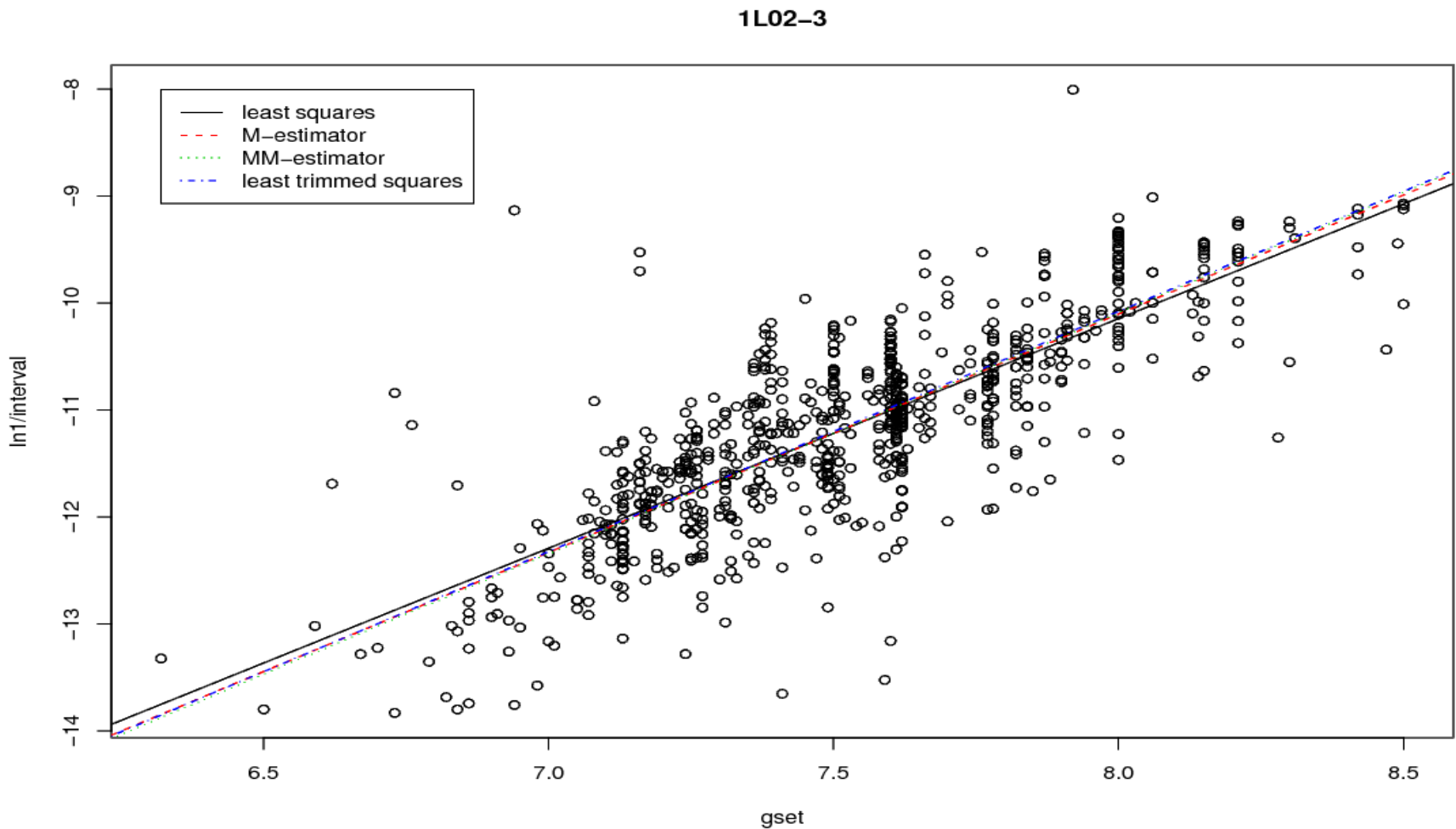
State Page Landscape 1.000 a4

R Graphics Output

Sun May 15 10:04:06 2011



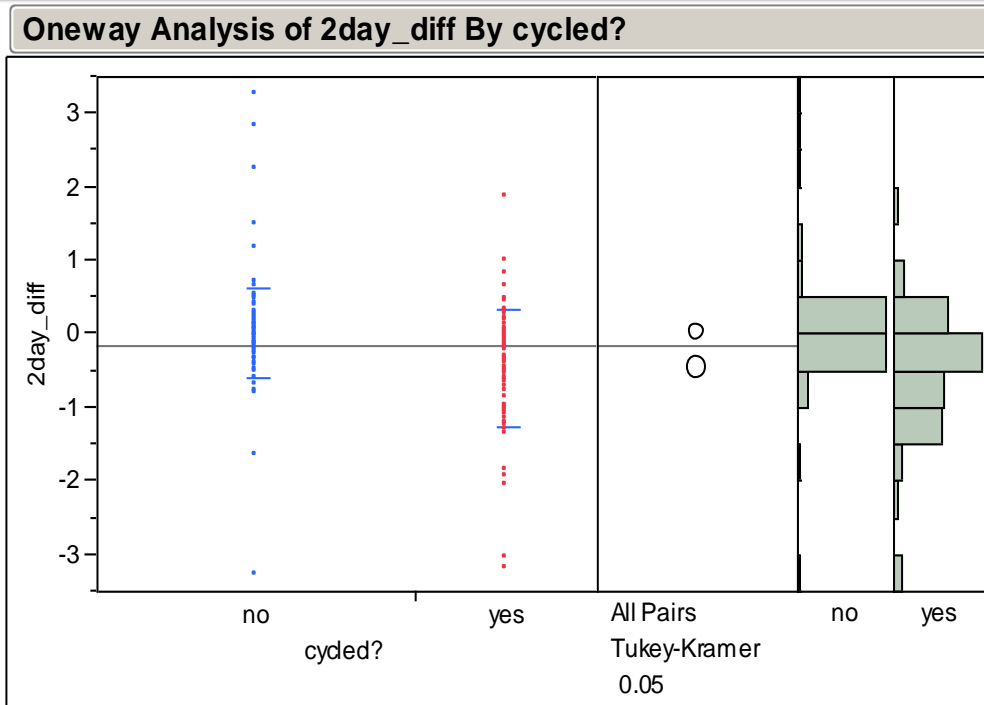
# slideShow 1L02-3



# What happens when CMs go to 300K

- Isabelle, uncontrolled warmup, 10% hit in gradient at 2 day fault interval. TN 05-057
- Spring 2009. CHL 4K down. SBR refrigerator couldn't keep everything cold. Ten modules warmed to 300K. TN 10-008 discusses this on pages 7-12. 7% hit with controlled cycle.
- Key TN 10-008 graph regenerated on next slide.

# 2 day interval gradient change for cycled vs uncycled



2day\_diff is the difference in predicted gradient for 2 day fault interval in Dec. 2008 less that in Dec. 2009. 75 cycled, 130 uncycled with valid models both times.

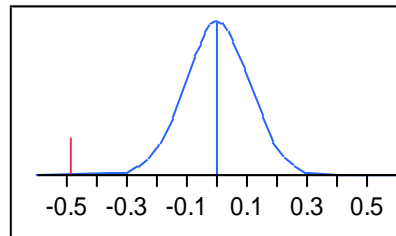
Excluded Rows 111

## t Test

yes-no

Assuming unequal variances

Difference	-0.48527	t Ratio	-4.55586
Std Err Dif	0.10651	DF	124.0112
Upper CL Dif	-0.27444	Prob >  t	<.0001*
Lower CL Dif	-0.69609	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*





# TMVA

- Toolkit for MultiVariate Analysis for CERN's ROOT.
- <http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf>
- machine learning emphasis
- from what I've read of the Users Guide in the week since I learned about TMVA, it wouldn't be helpful at JLAB.  
Read and decide for yourselves.

# Conclusions

- Pay attention to estimators of significance,  $F$  or  $t$ , when you're doing regression.
- Check Shapiro-Wilk or KSL test for normality in addition to chi-squared (if you can afford the CPU cycles). Make Q-Q residual plots and look for deviations from straight.
- Try stepwise regression on a subset of data if you're not sure which variables are best to regress against.
- Try one or more methods of robust regression. You may be able to reduce the number of data cuts you make, increasing the statistics and the significance of the physical result.