

Correcting for multiple testing

Old tricks, state-of-the-art solutions
and common sense

Ivan Langhans
CQ Consultancy



Third European DOE User Meeting, Luzern 2010

CQ Consultancy

*Expertise in Chemometrics;
the interaction between Chemistry and Statistics*

CQ Consultancy

- **Chemometrics & Qualimetrics**
 - Applied Statistics
 - Design of Experiments
 - Statistical Process Control
 - **Multivariate Analysis** (analytical, process, -omics, ...)
- **Training, Consulting, Contracting**
- **Chemistry, Food industries, Pharmaceutical industries**
- **www.cq.be / www.cqplus.ch**



Third European DOE User Meeting, Luzern 2010

Menu

- I. The Case
- II. What's the problem?
 1. In Pursuit of (the) True Variance
 2. What are the odds?
 3. Choosing is loosing
- III. The solutions (actually more problems)
 1. The many versions of the truth
 2. The False Discovery Channel
 3. The not so freedom of choice
 4. Forget about statistics (for a moment)
- IV. Cracking the Case
- V. Conclusions
- VI. Apologies



Third European DOE User Meeting, Luzern 2010

I. The Case

2^{5-1} design, no replicates



Third European DOE User Meeting, Luzern 2010

Response 1

Response 1

ANOVA for Response Surface Reduced 2FI Model
Analysis of variance table [Partial sum of squares - Type III]

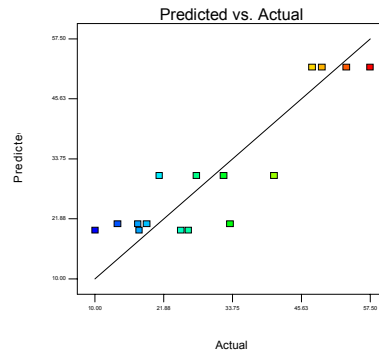
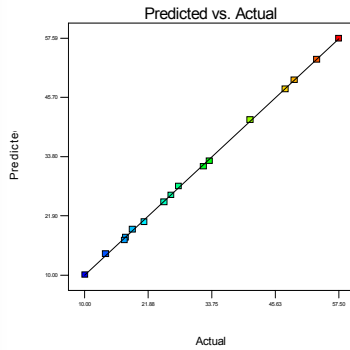
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	3326.88	12	277.24	809.95	< 0.0001
A-A	516.43	1	516.43	1508.73	< 0.0001
B-B	1749.33	1	1749.33	5110.64	< 0.0001
C-C	27.30	1	27.30	79.76	0.0030
D-D	168.35	1	168.35	491.83	0.0002
E-E	36.91	1	36.91	107.82	0.0019
AB	409.05	1	409.05	1195.04	< 0.0001
AD	13.88	1	13.88	40.54	0.0078
AE	19.58	1	19.58	57.20	0.0048
BD	3.71	1	3.71	10.83	0.0461
CD	160.66	1	160.66	469.35	0.0002
CE	94.58	1	94.58	276.30	0.0005
DE	127.13	1	127.13	371.40	0.0003
Residual	1.03	3	0.34		
Cor Total	3327.91	15			

Std. Dev.	0.59	R-Squared	0.9997
Mean	30.71	Adj R-Squared	0.9985
C.V. %	1.91	Pred R-Squared	0.9912
PRESS	29.21	Adeq Precision	90.071



Third European DOE User Meeting, Luzern 2010

It's Beuuuuutiful!



Third European DOE User Meeting, Luzern 2010

The Other Case

6 factor CCF with MRV cube, 6 centerpoints



Third European DOE User Meeting, Luzern 2010

ANOVA for Response Surface Reduced Quadratic Model
Analysis of variance table [Partial sum of squares - Type III]

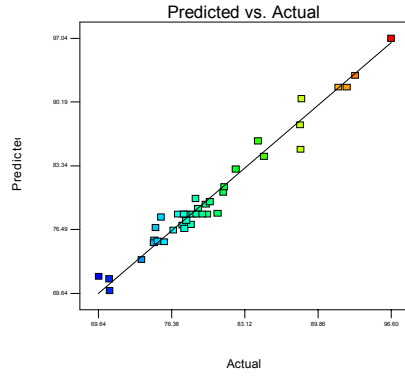
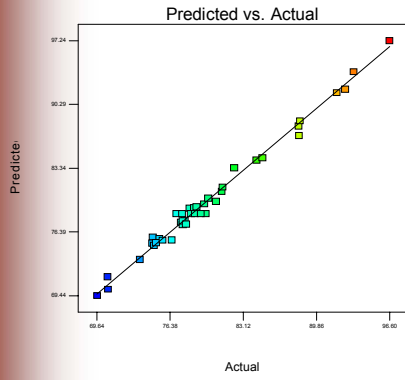
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	1527.58	15	101.84	181.48	< 0.0001
A-A	530.59	1	530.59	945.50	< 0.0001
B-B	309.80	1	309.80	552.06	< 0.0001
C-C	164.59	1	164.59	293.29	< 0.0001
D-D	65.83	1	65.83	117.30	< 0.0001
E-E	11.99	1	11.99	21.36	0.0001
F-F	0.25	1	0.25	0.44	0.5116
AB	74.59	1	74.59	132.93	< 0.0001
AE	4.06	1	4.06	7.24	0.0128
AF	10.53	1	10.53	18.77	0.0002
CD	13.87	1	13.87	24.71	< 0.0001
DE	2.27	1	2.27	4.05	0.0554
EF	8.53	1	8.53	15.20	0.0007
B2	59.43	1	59.43	105.91	< 0.0001
D2	3.61	1	3.61	6.42	0.0182
F2	2.76	1	2.76	4.92	0.0363
Residual	13.47	24	0.56		
Lack of Fit	8.20	19	0.43	0.41	0.9278
Pure Error	5.26	5	1.05		
Cor Total	1541.04	39			

Std. Dev.	0.75	R-Squared	0.9913
Mean	79.95	Adj R-Squared	0.9858
C.V. %	0.94	Pred R-Squared	0.9771
PRESS	35.32	Adeq Precision	58.142



Third European DOE User Meeting, Luzern 2010

Yup. It's plastic.



Third European DOE User Meeting, Luzern 2010

II. What's the Problem??



Third European DOE User Meeting, Luzern 2010

II.1 In Search of (the) True Variance

- Hypothesis tests need an estimate of the experimental error, it's their yardstick
- The true variance σ^2 will never be known
- Obvious estimator: pure error variance
- Popular estimator: MS_{residual}
- But MS_{residual} is only a good estimator if you fitted the correct model.
- Hmm . . .

- Too few terms in the model: MS_{residual} inflated
- Too many terms: MS_{residual} too optimistic making it more likely that terms get unrealistically low p-values



Third European DOE User Meeting, Luzern 2010

II.2 What are the odds?

If you test each hypothesis at the 5% significance level (5% false positive probability) and you do that for a lot of terms, the "overall" α can get pretty big and there may be several false positives in the model you report!



Third European DOE User Meeting, Luzern 2010

II.3 Choosing is losing!

- If you selected 10 terms from a larger set of candidate terms, the degrees of freedom "eaten" by the model is not **NOT 10** and the df left for estimating the variance is not equal to $n - 10$.

"... a quiet scandal that despite this awareness most textbooks completely ignore this problem and advocate the blind use of for instance C_p based stepwise or all subsets regression, followed by an estimate of the prediction error based on the C_p value of the "best model." (Breiman '92)

- The true or "**generalized degrees of freedom**" (Ye '98) can be considerably higher
- Post-hoc in multiple comparisons



Third European DOE User Meeting, Luzern 2010

Put II.1 and II.2 together, things get worse

And it doesn't get better if you add II.3!

Weisberg: $R^2 > 90\%$ with just noise



Third European DOE User Meeting, Luzern 2010

III. The “Solutions”



Third European DOE User Meeting, Luzern 2010

III.1 The many versions of the truth (estimating σ^2)

- Pure error
 - Pro: OK
 - Con: - you may have few replicates (\Rightarrow low power)
 - you may have no replicates
- Apply a correction/penalty to the MS_{residual}
 - Pro: better than not correcting
 - Con: - mainly developed for factorial designs (Lenth, Dong, ...)
 - “Fixed-X” estimation (Breiman)
- Use an external estimate of σ^2
- Use the next best thing to having an external estimate of σ^2 :
(double-loop) cross-validation
 - Pro: fundamentally correct
 - Con: - leaving out points of sparse design not appreciated by MLR
 - design points are mainly high-leverage points
 - \Rightarrow not representative



Third European DOE User Meeting, Luzern 2010

III.2 The False Discovery Channel

- Bonferroni correction (or variant): perform each test at the α/k sig. level
 - Con: decreases power (should anticipate this in planning your design size . . .)
- Control the **False Discovery Rate (FDR)** instead:
FDR: the proportion of false positives in the fraction that turned out to be statistically significant (Steinberg)
 - Pro: sensible thing to do
 - Con: - still reduces “actual α ” and consequently power
 - comes in many flavors



Third European DOE User Meeting, Luzern 2010

III.3 The not so freedom of choice

- Calculation of the cost of model selection:
Generalized Degrees of Freedom (Ye '98)

- Principle:

$$p = \text{tr}(H) = \sum_i h_{ii} \approx \sum_i \frac{\partial \hat{\mu}_i}{\partial y_i}$$

⇒ estimate p by **perturbation analysis** on y

- gdf depends on: n , p , p_{select} , β , S/N , method
- gdf can be **MUCH** higher than p (so df_{resid} should be much lower)



Third European DOE User Meeting, Luzern 2010

III.4 Forget about statistics (for a moment)

- Typically 1 – 10 statistically significant effects
- You can't have 10 important effects
- Think about practical relevance: what difference are you interested in?
You could opt for statistical equivalence testing:
if CI included in "acceptance" interval, ignore (remove) that term
- Use your knowledge/hunch about σ^2 (with $\sigma^2_{\text{measurement}}$ as a lower limit)



Third European DOE User Meeting, Luzern 2010

IV. Cracking The Case



Third European DOE User Meeting, Luzern 2010

The Case (2^{5-1})

- For the 2^{5-1} I couldn't check using the pure error, a 2 FI model would saturate the design so no FDR either.
 - ⇒ only gdf calculation: cost = 15 instead of 12
 - ⇒ so df_{residual} not 3 but only 1!



Third European DOE User Meeting, Luzern 2010

The Case (2^{5-1})

	SS	b	df	pDX	GDF corrected
A-A	516.43	5.68	1	< 0.0001	0.0164
B-B	1749.33	10.46	1	< 0.0001	0.0089
C-C	27.3	1.31	1	0.0030	0.0710
D-D	168.35	-3.24	1	0.0002	0.0287
E-E	36.91	-1.52	1	0.0019	0.0611
AB	409.05	5.06	1	< 0.0001	0.0184
AD	13.88	0.93	1	0.0078	0.0992
AE	19.58	1.11	1	0.0048	0.0837
BD	3.71	0.48	1	0.0461	0.1878
CD	160.66	-3.17	1	0.0002	0.0294
CE	94.58	2.43	1	0.0005	0.0383
DE	127.13	-2.82	1	0.0003	0.0330
Residual	1.03		3		



Third European DOE User Meeting, Luzern 2010

The Other Case (CCF)

1. Applied conservative FDR calculation based on full quadratic model
2. Calculated gdf for "conservative" forward selection model \Rightarrow gdf \approx 23 (instead of 15)
3. Used pure error estimate for σ^2



Third European DOE User Meeting, Luzern 2010

The Other Case (CCF)

Model	SS	MS	F	pDX	pGDF	pFDR (cut=0.0082)	p(PE, df=5)	
A-A	530.59	1	530.59	945.5	8.64E-21	2.42E-16	8.64E-21	3.24E-06
B-B	309.8	1	309.8	552.06	4.50E-18	2.12E-14	4.50E-18	1.22E-05
C-C	164.59	1	164.59	293.29	5.86E-15	3.72E-12	5.86E-15	5.77E-05
D-D	65.83	1	65.83	117.3	1.01E-10	4.75E-09	1.01E-10	5.17E-04
E-E	11.99	1	11.99	21.36	1.09E-04	2.44E-04	1.09E-04	1.97E-02
F-F	0.25	1	0.25	0.44	5.13E-01	5.16E-01	5.13E-01	6.46E-01
AB	74.59	1	74.59	132.93	2.85E-11	1.85E-09	2.85E-11	3.86E-04
AE	4.06	1	4.06	7.24	1.28E-02	1.55E-02	1.28E-02	1.06E-01
AF	10.53	1	10.53	18.77	2.27E-04	4.52E-04	2.27E-04	2.49E-02
CD	13.87	1	13.87	24.71	4.47E-05	1.16E-04	4.47E-05	1.50E-02
DE	2.27	1	2.27	4.05	5.55E-02	6.03E-02	5.55E-02	2.01E-01
EF	8.53	1	8.53	15.2	6.80E-04	1.15E-03	6.80E-04	3.58E-02
B^2	59.43	1	59.43	105.91	2.80E-10	1.01E-08	2.80E-10	6.57E-04
D^2	3.61	1	3.61	6.42	1.82E-02	2.14E-02	1.82E-02	1.23E-01
F^2	2.76	1	2.76	4.92	3.63E-02	4.05E-02	3.63E-02	1.66E-01
Residual	13.47	24	0.56					
Lack of Fit	8.2	19	0.43	0.41	0.9278			
Pure Error	5.26	5	1.05					



Third European DOE User Meeting, Luzern 2010

V. Conclusions

- There's obviously a problem but the solution is less obvious
- GDF while valuable in providing insight, won't save the day if you have a reasonable n/p
- FDR might be a nice to have
- Using PE as an estimator is a simple but not very efficient solution
- Things will often clear up if you throw out small sized effects (irrespective of statistical significance)
- Keep things in the back of your head when sizing your design



Third European DOE User Meeting, Luzern 2010

References

Weisberg, S. (2005). *Applied Linear Regression*, John Wiley

On FDR in DOE

M. Tripolski Kimel, Y. Benjamini, D.M. Steinberg. The false discovery rate for multiple testing in factorial experiments. *Technometrics*, 50, (2008), 32-39.

On Generalized degrees of freedom

Jianming Ye, On Measuring and Correcting the Effects of Data Mining and Model Selection, *JASA*, Vol. 93, No. 441 (Mar., 1998), pp. 120-131

On model uncertainty, model selection, optimism and the meaning of life

BREIMAN, L. and SPECTOR, P. 1992. Submodel selection and evaluation in regression. The random X case. *Internat. Statist. Rev.* 60 291 319.

David Draper. Assessment and propagation of model uncertainty (with discussion). *J Roy Stat Soc B*, 57:45-97, 1995.

Chris Chatfield, Model Uncertainty, Data Mining and Statistical Inference, *J.Roy. Stat. Soc. B*, Vol. 158, No. 3 (1995), pp. 419-466



Third European DOE User Meeting, Luzern 2010