# Statistics for Digital Formulation

## *Training for Scientists*

Dick Boddy


www.s4i.co.uk

# *Methods of Statistical Analysis*

- **Multiple Regression**
- **ANOVA with co-variates**
- **Discriminant Analysis**
- **Principal Component/Factor Analysis**
- **Partial Least Squares (PLS)**
- **Neural Networks**
- **Big Data Methods – Python, R…**

# *Multiple Regression*

**A method which has been widely applied with great success in model-building.**

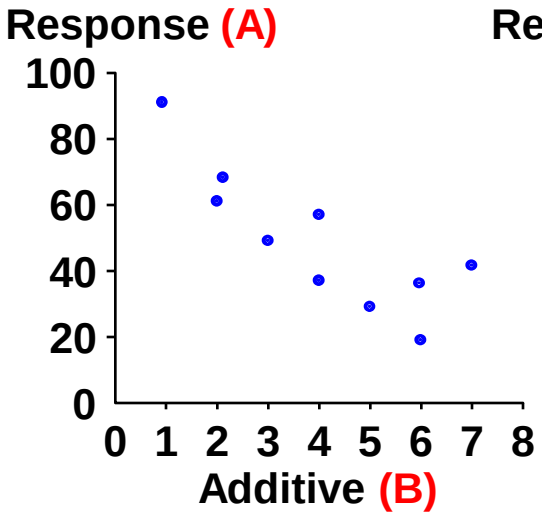
**Let us look at a Formulation example with a small data set for illustration purposes. In reality there could be many more variables and many more samples**

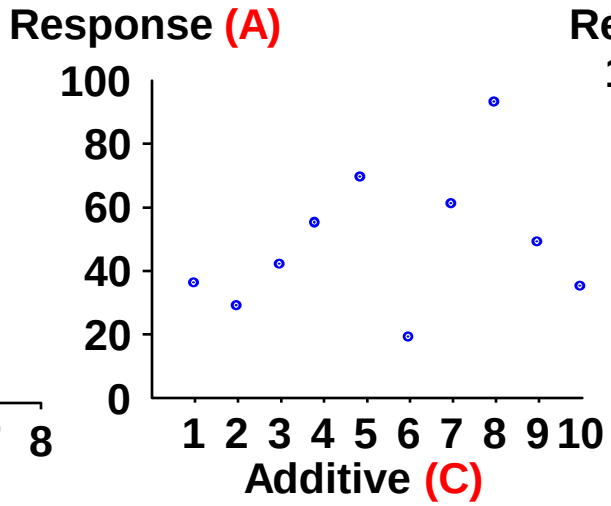
**Aim: to minimise the response.**

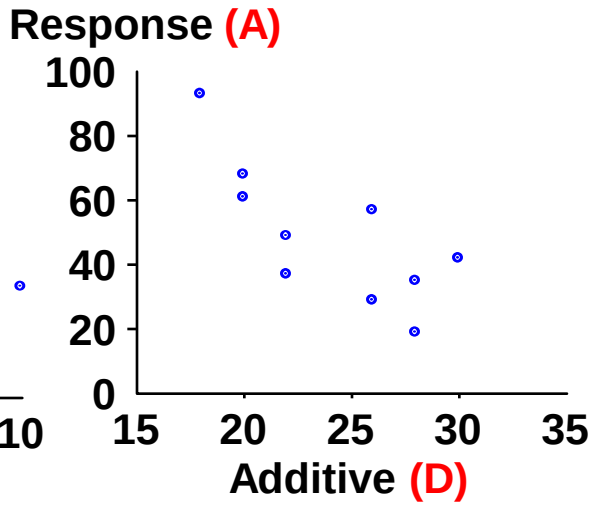
# Results of a Formulation Trial

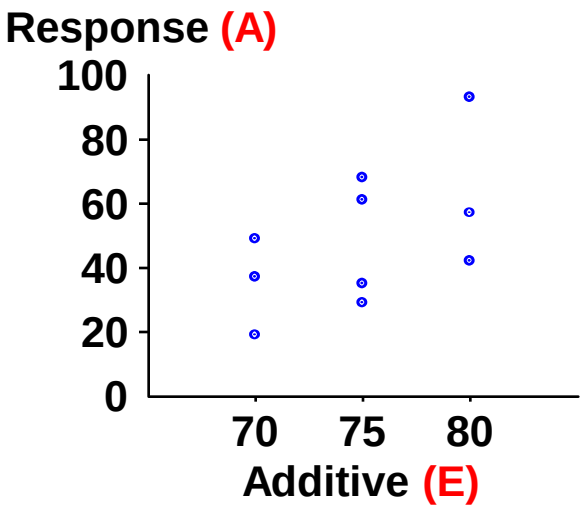
| Sample | Response A | Additive B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 37 | 4 | 1 | 22 | 70 | 300 |
| 2 | 29 | 5 | 2 | 26 | 75 | 300 |
| 3 | 42 | 7 | 3 | 30 | 80 | 300 |
| 4 | 57 | 4 | 4 | 26 | 80 | 200 |
| 5 | 68 | 2 | 5 | 20 | 75 | 200 |
| 6 | 19 | 6 | 6 | 28 | 70 | 200 |
| 7 | 61 | 2 | 7 | 20 | 75 | 200 |
| 8 | 93 | 1 | 8 | 18 | 80 | 400 |
| 9 | 49 | 3 | 9 | 22 | 70 | 400 |
| 10 | 35 | 6 | 10 | 28 | 75 | 400 |
| Mean | 49.0 | | | | | |
| SD | 21.6 | | | | | |

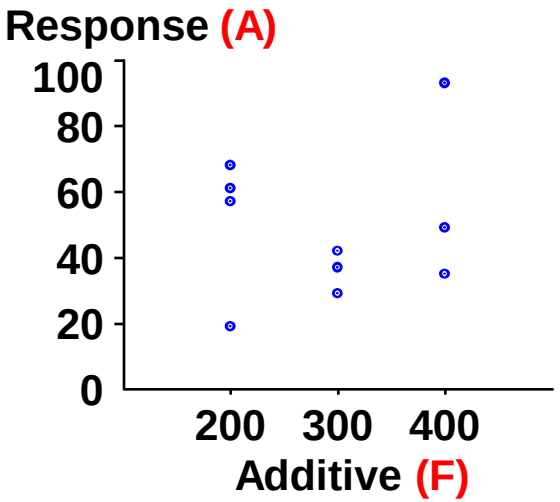
# Scatter Diagrams



# *Relationships*

**From the plots we can see two strong relationships with the response:**

**Additive B and Additive D**

# *Statistics at the First Step*

**Constant:** 49.00 **%fit:** 0.0 **d.f.:** 9 **RSD:** 21.59

**Variables in the equation**

| Variable | Coeff. | Decrease in %fit | Test val. to del. |
|---|---|---|---|

**Variables available to add**

| Variable | Increase in %fit | Test val. to incl. |
|---|---|---|
| B | 69.10 | 4.23** |
| C | 8.35 | 0.85 |
| D | 57.61 | 3.30* |
| E | 30.08 | 1.86~ |
| F | 1.52 | 0.35 |

# *Entry of First Variable into the model*

**Constant:** 84.89 **%fit:** 69.10 **d.f.:** 8 **RSD:** 12.73

**Variables in the equation**

| Variable | Coeff. | Decrease in %fit | Test val. to del. |
|---|---|---|---|
| B | -8.9722 | 69.10 | 4.23** |

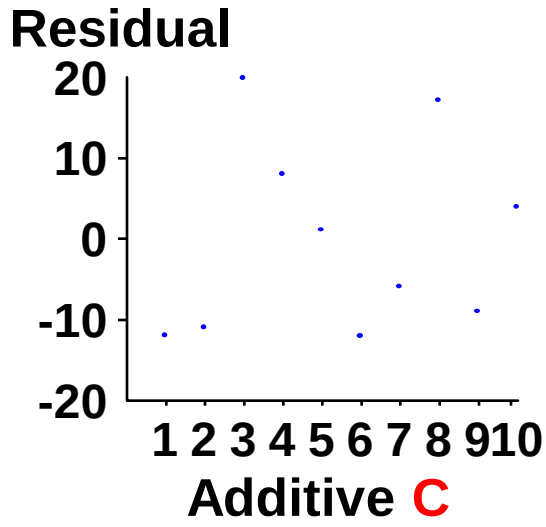
**Variables available to add**

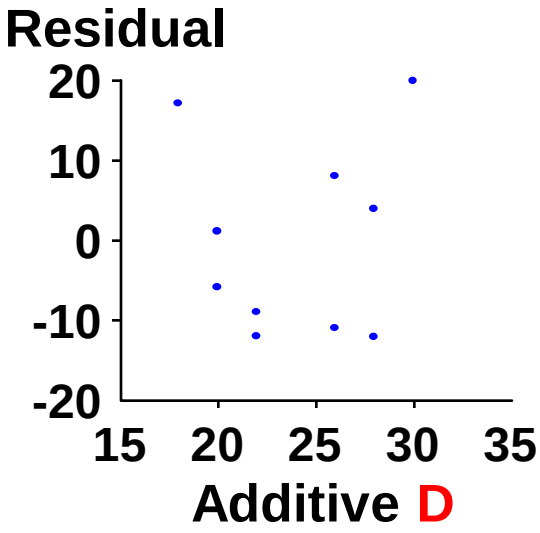
| Variable | Increase in %fit | Test val. to incl. |
|---|---|---|
| C | 0.61 | 0.38 |
| D | 4.77 | 1.13 |
| E | 24.31 | 5.08** |
| F | 1.52 | 0.60 |

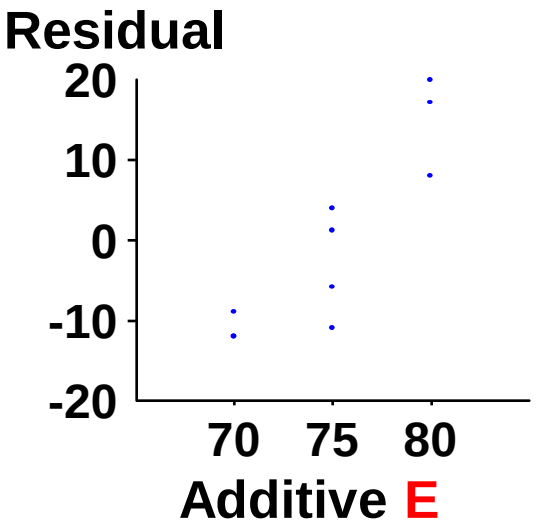
# *Residuals*

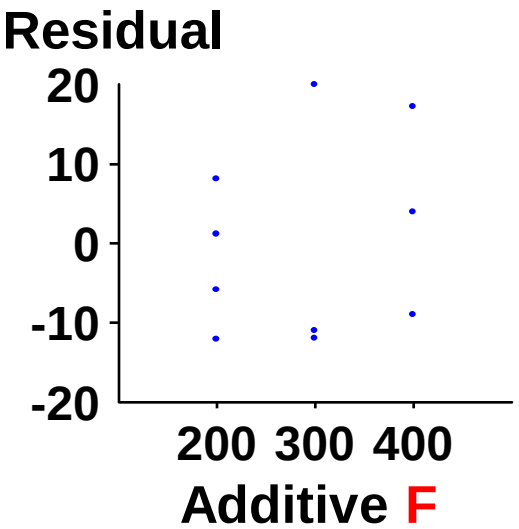
| Sample | Additive B | Actual Response | Predicted Response | Residual |
|---|---|---|---|---|
| 1 | 4 | 37 | 49.0 | -12.0 |
| 2 | 5 | 29 | 40.0 | -11.0 |
| 3 | 7 | 42 | 22.1 | 19.9 |
| 4 | 4 | 57 | 49.0 | 8.0 |
| 5 | 2 | 68 | 66.9 | 1.1 |
| 6 | 6 | 19 | 31.1 | -12.1 |
| 7 | 2 | 61 | 66.9 | -5.9 |
| 8 | 1 | 93 | 75.9 | 17.1 |
| 9 | 3 | 49 | 58.0 | -9.0 |
| 10 | 6 | 35 | 31.1 | 3.9 |

# Residuals v. Other Variables



# *Entry of Second Variable*

**Constant: -112.54** **%fit: 93.41** **d.f.: 7** **RSD: 6.29**

**Variables in the equation**

| Variable | Coeff. | Decrease in %fit | Test val. to del. |
|---|---|---|---|
| B | -8.6093 | 63.33 | 8.20** |
| E | 2.6130 | 24.31 | 5.08** |

**Variables available to add**

| Variable | Increase in %fit | Test val. to incl. |
|---|---|---|
| C | 1.22 | 1.16 |
| D | 0.67 | 0.83 |
| F | 1.52 | 1.34 |

# Whatever happened to Additive D?

## *An Alternative Model*

**Constant:** 49.00 **%fit:** 0.0 **d.f.:** 9 **RSD:** 21.59

**Variables in the equation**

| Variable | Coeff. | Decrease in %fit | Test val. to del. |
|---|---|---|---|

**Variables available to add**

| Variable | Increase in %fit | Test val. to incl. |
|---|---|---|
| B | 69.10 | 4.23** |
| C | 8.35 | 0.85 |
| D | 57.61 | 3.30* |
| E | 30.08 | 1.86~ |
| F | 1.52 | 0.35 |

# An Alternative Model

**Constant:** 144.68 **%fit:** 57.61 **d.f.:** 8 **RSD:** 14.91

**Variables in the equation**

| Variable | Coeff. | Decrease in %fit | Test val. to del. |
|---|---|---|---|
| D | -3.9868 | 57.61 | 3.30* |

**Variables available to add**

| Variable | Increase in %fit | Test val. to incl. |
|---|---|---|
| B | 16.26 | 2.09~ |
| C | 2.04 | 0.59 |
| E | 36.00 | 6.28** |
| F | 0.59 | 0.31 |

# An Alternative Model

**Constant:** -88.78 **%fit:** 93.61 **d.f.:** 7 **RSD:** 6.19

**Variables in the equation**

| Variable | Coeff. | Decrease in %fit | Test val. to del. |
|---|---|---|---|
| D | -4.1960 | 63.53 | 8.34** |
| E | 3.1797 | 36.00 | 6.28** |

**Variables available to add**

| Variable | Increase in %fit | Test val. to incl. |
|---|---|---|
| B | 0.47 | 0.69 |
| C | 2.64 | 2.05~ |
| F | 0.55 | 0.75 |

# *Limitations to the Analysis*

**Two possible equations:-**

***(i) Response = -112.5 - 8.61 Ad. B + 2.61 Ad. E***

**Increase Additive B and reduce Additive E.**

**% fit = 93.4**

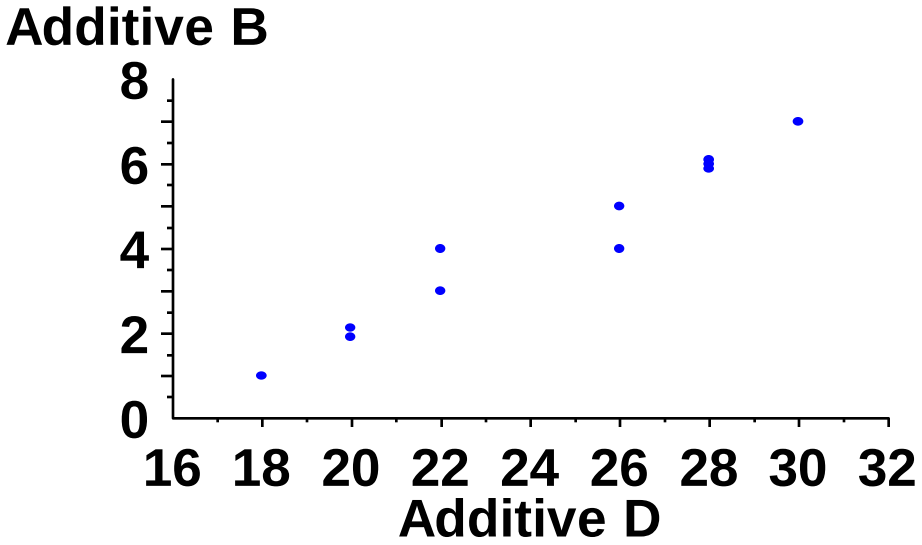
***(ii) Response = -88.8 - 4.20 Ad. D + 3.18 Ad. E***

**Increase Additive D and reduce Additive E.**

**% fit = 93.6**

# *Limitations to the Analysis*

**The conflicting advice is due to *bad design*.**




**There is a *correlation* between *Additives B* and *D.***

**We are unable to say which variable is affecting the response.**

# *Inter-correlation Matrix*

| | B | C | D | E | F |
|---|---|---|---|---|---|
| B | * | -0.26 | 0.97 | -0.07 | 0.00 |
| C | | * | -0.20 | -0.04 | 0.44 |
| D | | | * | 0.07 | -0.06 |
| E | | | | * | 0.00 |
| F | | | | | * |

**The ideal inter-correlation is *zero.***

**With a good design we can obtain *zero* or, if not, *low values*.**

# *Large poorly-designed Data Sets*

**Say *10* responses**

**50 variables** **additives**

**treat (conc.) levels 1%**

***1000*** **measured samples**

**Giving 500,000 data points**

# *Variables*

**Many inter-correlations.**

**Use Principal Component Analysis (PCA) or its cousin, Factor Analysis.**

**Group variables with high inter-correlations into Factors.**

**Factors have *zero* inter-correlations with each other.**

**Probably *3-5* factors explain *99%* of the input variables**

# *Responses*

**Minor Responses could be highly correlated.**

**Use PCA or Factor Analysis to reduce number of responses.**

**Probably *2* factors and *3* major variables.**

# *Data collected over time*

**Cusum Plot – finds step changes. Important information occurs in the steps**

**Autocorrelation plot – adjacent samples are highly correlated giving negligible information**

# *Possible responses*

***10*** **responses** → ***5*** **responses**

***50*** **variables** → ***3*** **factors**

***1000*** **samples** → ***40*** **steps**

***500,000*** **data points** → ***600 points***

**We have removed the trivia from the data and can now concentrate the statistical analysis on the important features of the data.**

# *A Formulation Design with no interactions*

***7 components to be investigated in 12 formulations.***

| Form. | A | B | C | D | E | F | G | Consump. |
|---|---|---|---|---|---|---|---|---|
| 1 | $A_2$ | $B_1$ | $C_2$ | $D_1$ | $E_1$ | $F_1$ | $G_2$ | 11.76 |
| 2 | $A_2$ | $B_2$ | $C_1$ | $D_2$ | $E_1$ | $F_1$ | $G_1$ | 11.37 |
| 3 | $A_1$ | $B_2$ | $C_2$ | $D_1$ | $E_2$ | $F_1$ | $G_1$ | 11.91 |
| 4 | $A_2$ | $B_1$ | $C_2$ | $D_2$ | $E_1$ | $F_2$ | $G_1$ | 12.11 |
| 5 | $A_2$ | $B_2$ | $C_1$ | $D_2$ | $E_2$ | $F_1$ | $G_2$ | 11.32 |
| 6 | $A_2$ | $B_2$ | $C_2$ | $D_1$ | $E_2$ | $F_2$ | $G_1$ | 12.02 |
| 7 | $A_1$ | $B_2$ | $C_2$ | $D_2$ | $E_1$ | $F_2$ | $G_2$ | 12.62 |
| 8 | $A_1$ | $B_1$ | $C_2$ | $D_2$ | $E_2$ | $F_1$ | $G_2$ | 12.11 |
| 9 | $A_1$ | $B_1$ | $C_1$ | $D_2$ | $E_2$ | $F_2$ | $G_1$ | 12.40 |
| 10 | $A_2$ | $B_1$ | $C_1$ | $D_1$ | $E_2$ | $F_2$ | $G_2$ | 11.98 |
| 11 | $A_1$ | $B_2$ | $C_1$ | $D_1$ | $E_1$ | $F_2$ | $G_2$ | 12.24 |
| 12 | $A_1$ | $B_1$ | $C_1$ | $D_1$ | $E_1$ | $F_1$ | $G_1$ | 11.74 |

**All inter-correlations are *zero.***

# *Case Study: Better oils for marine engines - Less wear*

**Test was run on a slightly scaled down marine engine.**

***3*** **cylinders**

**Each run takes *one* week.**

***4*** **runs in each phase (*12* results)**

**Could be run-to-run differences.**

# *Investigate seven additives*

**Type and treat-rate of dispersant**

**Type and treat-rate of detergent**

**Anti-wear packages and treat-rate**

**Etc**

**Etc**

**Look at several base-oils to ensure robust conclusions.**

# *Advice from on high*

**1st two phases successful**

**Research Director demanded his favourite additive be included.**

**Phase 3: Abandon favourite additive – was detrimental.**

**If it had been included in a large design, half of the results would have been useless.**

# *Phase 4 & Conclusion*

**Phase 4**

**Optimised treat-rates for less wear and lower cost.**

**Conclusion**

***70%* less wear**

**Took *40%* of world market**

**Still successful *25* years later.**