GAS MILEAGE DOE CASE STUDY

The focus of this study is to discover which factors affect the gas mileage of a vehicle by selecting and adjusting 4 quantitative factors.

For the study we will use 3 different approaches.

1. A non-statistical approach for determining which factors and 2-way interactions should go into the prediction equation.

2. A statistical approach called ANOVA (ANalysis Of VAriances) to assist making the decision in 1)

3. Using Regression for the prediction

1.1 Full factorial Design

$$\text{fullfact}(4,2)=\begin{bmatrix} \text{"Run"} & \text{"Block"} & \text{"A"} & \text{"B"} & \text{"C"} & \text{"D"} \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 2 & 1 & 1 & -1 & -1 & -1 \\ 3 & 1 & -1 & 1 & -1 & -1 \\ 4 & 1 & 1 & 1 & -1 & -1 \\ 5 & 1 & -1 & -1 & 1 & -1 \\ 6 & 1 & 1 & -1 & 1 & -1 \\ 7 & 1 & -1 & 1 & 1 & -1 \\ 8 & 1 & 1 & 1 & 1 & -1 \\ 9 & 1 & -1 & -1 & -1 & 1 \\ 10 & 1 & 1 & -1 & -1 & 1 \\ 11 & 1 & -1 & 1 & -1 & 1 \\ 12 & 1 & 1 & 1 & -1 & 1 \\ 13 & 1 & -1 & -1 & 1 & 1 \\ 14 & 1 & 1 & -1 & 1 & 1 \\ 15 & 1 & -1 & 1 & 1 & 1 \\ 16 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

1. The non-statistical approach

We have seleceted these factors for the study:

- Tire Pressure
- Ignition timing
- Oil Type
- Gas Type

assuming that each factor has 2 levels:

A - Tire Pressure:	28 psi	35 psi
B - Ignition Timing:	Low	High
C - Oil Type:	1	2
D - Gas Type:	1	2

The investigation of all possible combinations of four factors at 2 levels will require $2^4=16$ runs.

In the fullfact-matrix -1 indicate a factor set at its low value and the +1 values represent a high factor setting.

Note that columns A, B, C and D are balanced vertically and that the columns are orthogonal.

The full factorial design allows estimating all possible factor combinations on the response:

Main effects: A, B, C, D

2-way interactions: AB, AC, AD BC, BD, CD

3-way interactions: ABC, ABD, ACD, BCD

4-way interaction: ABCD

A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1
1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1
-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1
1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1
-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1
1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1
-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1
1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1
1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1
-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1
-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1
1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1
-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

$AB = A \cdot B$
$AC = A \cdot C$

$ABC = A \cdot B \cdot C$

etc.

Seldom interactions beyond 2-ways are significant, so time and resources can be reduced by use of a fraction of the full factorial design.

1.2 Fractional factorial Design

We can sacrifice the knowledge of the 3-way interaction ABC in order to test a fourth factor by setting $D = ABC$

$D = ABC$ does not mean that the effects of D and ABC are the same - only that the effect of D and ABC can't be separated.

Multiplication by D on both sides gives the defining relation for the design (of **resolution IV**)

$ABCD = I$

The defining relation gives the **alias pattern**:

$A = BCD \quad AB = CD$
$B = ACD \quad AC = BD$
$C = ABD \quad AD = BC$

$$X := \text{fractfact}(4,1) = \begin{bmatrix} \text{“Run”} & \text{“Block”} & \text{“A”} & \text{“B”} & \text{“C”} & \text{“D=ABC”} \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 2 & 1 & 1 & -1 & -1 & 1 \\ 3 & 1 & -1 & 1 & -1 & 1 \\ 4 & 1 & 1 & 1 & -1 & -1 \\ 5 & 1 & -1 & -1 & 1 & 1 \\ 6 & 1 & 1 & -1 & 1 & -1 \\ 7 & 1 & -1 & 1 & 1 & -1 \\ 8 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

1.3 Performing the Experiments

The different combinations of the settings were replicated 3 times

$$Y := \begin{bmatrix} 22.27 & 21.12 & 21.37 \\ 17.35 & 18.60 & 17.97 \\ 22.49 & 23.15 & 22.08 \\ 18.36 & 17.63 & 17.04 \\ 14.22 & 15.40 & 10.46 \\ 27.08 & 24.54 & 24.57 \\ 9.96 & 13.80 & 11.92 \\ 22.78 & 26.97 & 27.14 \end{bmatrix}$$

$i := 0..7 \qquad M_i := \text{mean}\left(Y^{\langle i \rangle}\right) \qquad S_i := \text{Stdev}\left(Y^{\langle i \rangle}\right)$

The average and the standard deviation of the responses for each run are calculated

$$M = \begin{bmatrix} 21.587 \\ 17.973 \\ 22.573 \\ 17.677 \\ 13.36 \\ 25.397 \\ 11.893 \\ 25.63 \end{bmatrix} \qquad S = \begin{bmatrix} 0.605 \\ 0.625 \\ 0.54 \\ 0.661 \\ 2.58 \\ 1.458 \\ 1.92 \\ 2.47 \end{bmatrix}$$

1.4 Complete experimental matrix with response values, averages and standard deviations

Run	A	B	C	D	AB	AC	AD	Mileage data [mpg]			Avg [mpg]	Std Dev
								Trial 1	Trial 2	Trial 3		
1	-1	-1	-1	-1	1	1	1	22.27	21.12	21.37	21.59	0.60
2	1	-1	-1	1	-1	-1	1	17.35	18.60	17.97	17.97	0.63
3	-1	1	-1	1	-1	1	-1	22.49	23.15	22.08	22.57	0.54
4	1	1	-1	-1	1	-1	-1	18.36	17.63	17.04	17.68	0.66
5	-1	-1	1	1	1	-1	-1	14.22	15.40	10.46	13.36	2.58
6	1	-1	1	-1	-1	1	-1	27.08	24.54	24.57	25.40	1.46
7	-1	1	1	-1	-1	-1	1	9.96	13.80	11.92	11.89	1.92
8	1	1	1	1	1	1	1	22.78	26.97	27.14	25.63	2.47

1.5 Factorscreening and graphical analysis

$gen := \text{"A,B,C,D,AB,AC,AD"}$

Necessary in order to have 2-way interactions calculated.

$QM := \text{quickscreen}(X, Y, gen)$

$QM =$

"Factor"	"(−) Avg"	"(+) Avg"	"(+) − (−)"
"A"	17.353	21.669	4.316
"B"	19.579	19.443	−0.136
"C"	19.953	19.07	−0.883
"D"	19.138	19.884	0.746
"AB"	19.459	19.563	0.104
"AC"	15.226	23.797	8.571
"AD"	19.752	19.271	−0.481

A , (-)Avg = 17.353
is the average of all the responses with A at low level (-1)

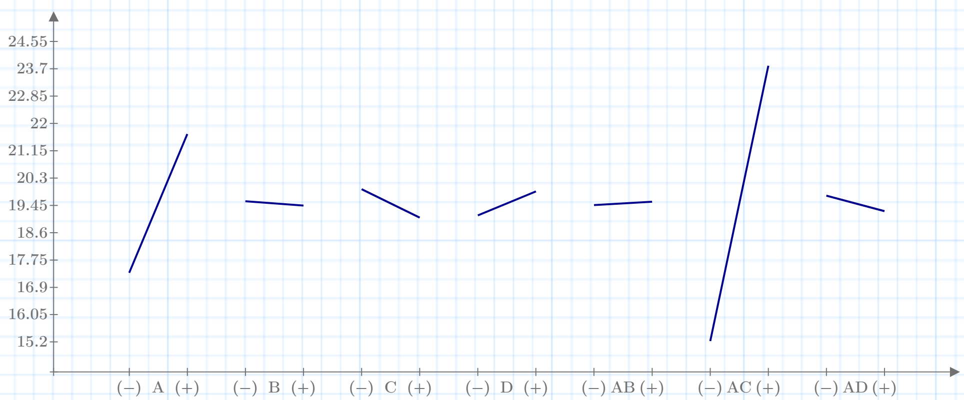
A , (+)Avg = 21.669
is the average of all the responses with A at high level (+1)

The slope for A is calculated as

$$\alpha_A = \frac{21.669 - 17.353}{1-(-1)} = \frac{4.316}{2} = 2.158$$

The slope of A is called the half-effect of factor A

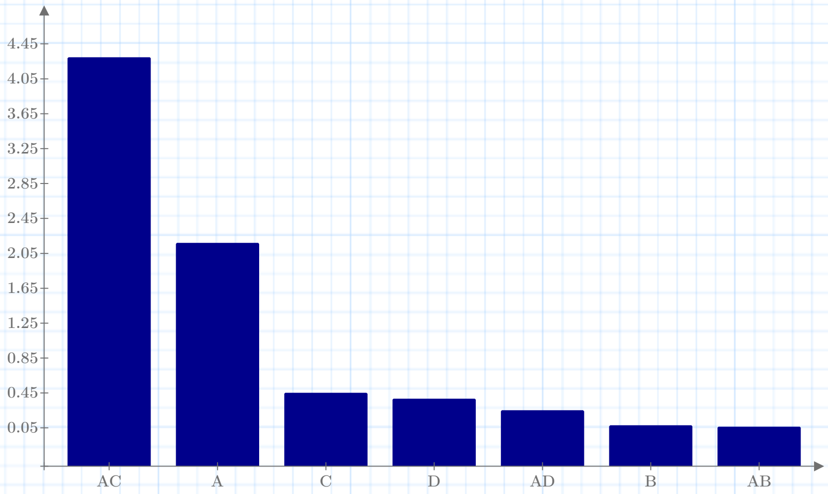
$EM := \text{effectsgraph}(QM)$



The steepness of the slope reflects the importance of the effect. It is apparent that the most important effects are factor A and the AC interaction.

Since AC is aliased with BD we must go back to the factors to determine which interaction is responsible for the steep slope.
It is unlikely to have a 'tire pressure x ignition timing' (AD) interaction so steep slope must be due to the 'ignition timing x type of oil' (BC) interaction.

$PM := Pareto(QM)$



The Pareto plot shows the half-effects in an ordered histogram

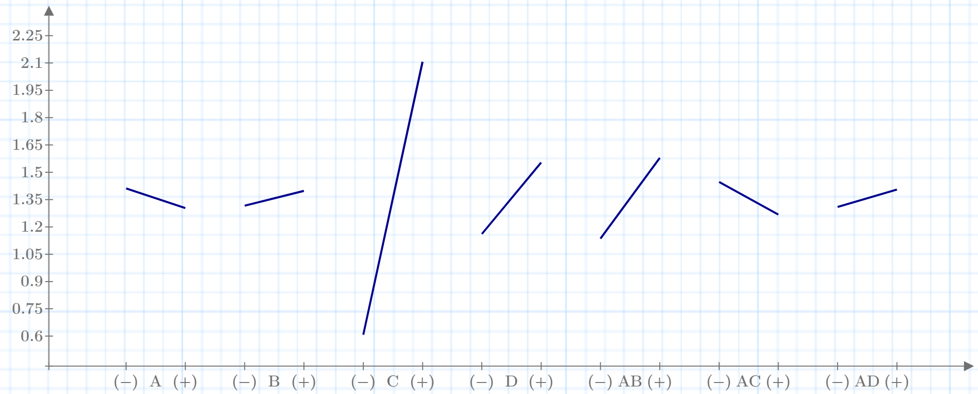
$QS := \text{quickscreen}(X, S, gen)$

Using the standard deviation instead of the mean values quickscreen can be used to determine if any of the factors has a big effect on the variability.

$QS =$

"Factor"	"(−) Avg"	"(+) Avg"	"(+) − (−)"
"A"	1.411	1.303	−0.108
"B"	1.317	1.398	0.081
"C"	0.608	2.107	1.499
"D"	1.161	1.554	0.393
"AB"	1.136	1.579	0.443
"AC"	1.447	1.268	−0.179
"AD"	1.31	1.405	0.095

$ES := \text{effectsgraph}(QS)$



Again we can draw the effects plot.

Note that C has a big effect on the variability.

If we want our factor settings for maximizing the gas mileage to be robust, we should set C at low level to reduce variability.

Conclusion:

A positive

B and D positive such that BD will be positive

C negative to reduce variability

Factor	+/- Setting	True Settting
A	+	35 psi
B	+	High
C	-	Type 1 Oil
D	+	Type 2 Gas

$$halfeffects(QM)=\begin{bmatrix} \text{"A"} & 2.158 \\ \text{"B"} & -0.068 \\ \text{"C"} & -0.441 \\ \text{"D"} & 0.373 \\ \text{"AB"} & 0.052 \\ \text{"AC"} & 4.285 \\ \text{"AD"} & -0.24 \end{bmatrix}$$

$$yield(A,B,C,D):=\text{mean}(Y)+2.158\cdot A-0.068\cdot B-0.441\ C+0.373\ D+0.052\ A\cdot B+4.285\ A\cdot C-0.24\ A\cdot D$$

$$yield(1,1,-1,1)=17.942$$

2. Statistical Methods - Anova

For each factor and 2-way interaction we test the hypothesis

$H_0: \quad \mu_{A\_minus} = \mu_{A\_plus}$

against the alternative

$H_1: \quad \mu_{A\_minus} \neq \mu_{A\_plus}$

using the F-distribution with parameters df_A and df_E

anova(X, Y, gen) =

"Source"	"SSE"	"DF"	"MSE"	"F"	"P"
"A"	111.759	1	111.759	44.587	$5.313 \cdot 10^{-6}$
"B"	0.111	1	0.111	0.044	0.836
"C"	4.673	1	4.673	1.864	0.191
"D"	3.338	1	3.338	1.332	0.265
"AB"	0.065	1	0.065	0.026	0.874
"AC"	440.755	1	440.755	175.843	$4.778 \cdot 10^{-10}$
"AD"	1.387	1	1.387	0.553	0.468
"Error"	40.104	16	2.507	NaN	NaN
"Total"	602.191	23	NaN	NaN	NaN

$\alpha := 0.05 \qquad df\_A := 1 \qquad df\_E := 16$

$F\_crit := \mathrm{qF}(1-\alpha, df\_A, df\_E) = 4.494$

If the F-value (column 5) is greater than the critical value, we accept the hypothesis H_0 otherwise we reject H_0 and accept the alternative H_1

The only significant effects are A and AC = BD

$y_{pred} = mean(Y) + 2.158 \cdot A + 4.285 \cdot A \cdot C$

The prediction equation

The coefficients are the half effects

halfeffects(QM) =

"A"	2.158
"B"	−0.068
"C"	−0.441
"D"	0.373
"AB"	0.052
"AC"	4.285
"AD"	−0.24

3. Regression

$C := \text{polyfitc}(X, Y, \text{"A B C AB AC AD"})$

$C =$

"Term"	"Coefficient"	"Std Error"	"95% CI Low"	"95% CI High"	"VIF"	"T"	"P"
"Intercept"	19.511	0.373	14.773	24.25	*NaN*	52.321	0.006
"A"	2.158	0.373	−2.58	6.896	1	5.787	0.054
"B"	−0.068	0.373	−4.806	4.67	1	−0.182	0.443
"C"	−0.441	0.373	−5.18	4.297	1	−1.183	0.223
"AB"	0.052	0.373	−4.686	4.79	1	0.14	0.456
"AC"	4.285	0.373	−0.453	9.024	1	11.492	0.028
"AD"	−0.24	0.373	−4.979	4.498	1	−0.645	0.318

If the p-value in the last column is smaller than $\alpha = 0.05$ the coefficient is significant - otherwise not. Based on regression the prediction equation is

$$y_{pred} = mean(Y) + 2.158 \cdot A + 4.285 \cdot A \cdot C$$

which supports the result from the ANOVA analysis