STATISTICAL MODELLING

PRACTICAL III
SOLUTIONS

III.1 Show that $\mathbf{Q}_U$, $\mathbf{Q}_T$ and $\mathbf{Q}_{U_{Res}}$ are symmetric and idempotent where $\mathbf{Q}_U = \mathbf{M}_U - \mathbf{M}_G$, $\mathbf{Q}_T = \mathbf{M}_T - \mathbf{M}_G$ and $\mathbf{Q}_{U_{Res}} = \mathbf{M}_U - \mathbf{M}_T$. You are given that $\mathbf{M}_U = \mathbf{I}_n$, $\mathbf{M}_T$ and $\mathbf{M}_G$ are symmetric and idempotent and that $\mathbf{M}_T\mathbf{M}_G = \mathbf{M}_G\mathbf{M}_T = \mathbf{M}_G$.

$$\mathbf{Q}_U' = \mathbf{M}_U' - \mathbf{M}_G' = \mathbf{M}_U - \mathbf{M}_G = \mathbf{Q}_U \text{ and}$$
$$\begin{aligned}\mathbf{Q}_U^2 &= (\mathbf{M}_U - \mathbf{M}_G)(\mathbf{M}_U - \mathbf{M}_G)\\ &= \mathbf{M}_U^2 - \mathbf{M}_U\mathbf{M}_G - \mathbf{M}_G\mathbf{M}_U + \mathbf{M}_G^2\\ &= \mathbf{M}_U - \mathbf{M}_G - \mathbf{M}_G + \mathbf{M}_G \quad \text{as } \mathbf{M}_U = \mathbf{I}_n\\ &= \mathbf{M}_U - \mathbf{M}_G\\ &= \mathbf{Q}_U\end{aligned}$$

$$\mathbf{Q}_T' = \mathbf{M}_T' - \mathbf{M}_G' = \mathbf{M}_T - \mathbf{M}_G = \mathbf{Q}_T \text{ and}$$
$$\begin{aligned}\mathbf{Q}_T^2 &= (\mathbf{M}_T - \mathbf{M}_G)(\mathbf{M}_T - \mathbf{M}_G)\\ &= \mathbf{M}_T^2 - \mathbf{M}_T\mathbf{M}_G - \mathbf{M}_G\mathbf{M}_T + \mathbf{M}_G^2\\ &= \mathbf{M}_T - \mathbf{M}_G - \mathbf{M}_G + \mathbf{M}_G\\ &= \mathbf{M}_T - \mathbf{M}_G\\ &= \mathbf{Q}_T\end{aligned}$$

$$\mathbf{Q}_{U_{Res}}' = \mathbf{M}_U' - \mathbf{M}_T' = \mathbf{M}_U - \mathbf{M}_T = \mathbf{Q}_{U_{Res}} \text{ and}$$
$$\begin{aligned}\mathbf{Q}_{U_{Res}}^2 &= (\mathbf{M}_U - \mathbf{M}_T)(\mathbf{M}_U - \mathbf{M}_T)\\ &= \mathbf{M}_U^2 - \mathbf{M}_U\mathbf{M}_T - \mathbf{M}_T\mathbf{M}_U + \mathbf{M}_T^2\\ &= \mathbf{M}_U - \mathbf{M}_T - \mathbf{M}_T + \mathbf{M}_T \quad \text{as } \mathbf{M}_U = \mathbf{I}_n\\ &= \mathbf{M}_U - \mathbf{M}_T\\ &= \mathbf{Q}_{U_{Res}}\end{aligned}$$

III.2 Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} 2 & 4 \\ 1 & 6 \end{bmatrix}$$

Find $\mathbf{y}'\mathbf{A}\mathbf{y}$. Why do you think $\mathbf{y}'\mathbf{A}\mathbf{y}$ is called a quadratic form?

$$\begin{aligned}\mathbf{y}'\mathbf{A}\mathbf{y} &= \begin{bmatrix} y_1 & y_2 \end{bmatrix}\begin{bmatrix} 2 & 4 \\ 1 & 6 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \\ &= \begin{bmatrix} y_1 & y_2 \end{bmatrix}\begin{bmatrix} 2y_1 + 4y_2 \\ 1y_1 + 6y_2 \end{bmatrix} \\ &= 2y_1^2 + 5y_1y_2 + 6y_2^2\end{aligned}$$

It is called a quadratic form because all the terms involve second-order terms in y, that is, either y_i^2 or y_iy_j

III.3 An investigation was conducted to examine differences between 3 brands of tyre in their braking distances (ft.) from a speed of 30 miles per hour. Altogether 9 tests were conducted with the particular brand tested at each test being chosen at random. The results are as follows:

Brand		
A	B	C
28	27	27
26	25	32
30	26	31

a) Give the matrix $\mathbf{X}_T$ in terms of both the individual elements and a partitioned matrix of **1** and **0** column vectors. Also give an expression for $\mathbf{M}_T$ in terms of **J** and **0** matrices.

b)

c) $$\mathbf{X}_T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_3 & \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} \\ \mathbf{0}_{3\times1} & \mathbf{1}_3 & \mathbf{0}_{3\times1} \\ \mathbf{0}_{3\times1} & \mathbf{0}_{3\times1} & \mathbf{1}_3 \end{bmatrix}$$

d)

e) $$\mathbf{M}_T = \begin{bmatrix} \frac{1}{3}\mathbf{J}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \frac{1}{3}\mathbf{J}_3 & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \frac{1}{3}\mathbf{J}_3 \end{bmatrix}$$

f)

g) Compute the column vectors $\mathbf{g} = \mathbf{M}_G\mathbf{y}$, $\mathbf{d}_G = \mathbf{Q}_G\mathbf{y}$, $\mathbf{t} = \mathbf{M}_T\mathbf{y}$, $\mathbf{t}_e = \mathbf{Q}_T\mathbf{y}$ and $\mathbf{d}_T = \mathbf{Q}_{U_{Res}}\mathbf{y}$. From these compute the sums of squares for the analysis of variance.

h)

i) The vectors for computing the sums of squares are given in the following table.

Brand	Stopping Distance **y**	Grand mean $\mathbf{g} = \mathbf{M}_G\mathbf{y}$	Total Test Deviations $\mathbf{d}_G = \mathbf{Q}_G\mathbf{y}$	Brand means $\mathbf{t} = \mathbf{M}_T\mathbf{y}$	Brand effects $\mathbf{t}_e = \mathbf{Q}_T\mathbf{y}$	Residual Test Deviations $\mathbf{d}_T = \mathbf{Q}_{U_{Res}}\mathbf{y}$
A	28	28	0	28	0	0
A	26	28	-2	28	0	-2
A	30	28	2	28	0	2
B	27	28	-1	26	-2	1
B	25	28	-3	26	-2	-1
B	26	28	-2	26	-2	0
C	27	28	-1	30	2	-3
C	32	28	4	30	2	2
C	31	28	3	30	2	1
SSq			48		24	24

a)

k) Construct the analysis of variance table. Use the R function `pf` to obtain the p-value — the call should be of the form `1 - pf(F,` ν_1`,` ν_2`)`.

Step 1: Set up hypotheses

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \mu$ *(or $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$)*

H_1: not all population Band means are equal *(or $\boldsymbol{\psi}_B = \mathbf{X}_B\boldsymbol{\alpha}$)*

Set $\alpha = 0.5$.

Step 2: Calculate test statistic

The analysis of variance table for the example is:

Source	df	SSq	MSq	F	p
Tests	8	48			
Brands	2	24	12	3.0	0.1250
Residual	6	24	4		

Step 3: Decide between hypotheses

The probability of exceeding an F of 3.0 with ν_1 = 2 and ν_2 = 6 is P(F ≥ 3.0) = 0.1250 > 0.05. Little evidence of a difference between the brands and the minimal expectation model $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$ appears to best describe the data.

III.4 Use R to produce a randomized layout for an experiment involving 21 runs to which 7 treatments are to be allocated so that each is replicated 3 times. Use the seed 911 in producing the layout. The effect of this is that everyone should get the same answer — in practice, one should randomly choose the seed.

Follow the procedure described in section B.1, Completely Randomized Design of appendix B, Obtaining randomized layouts with R. The functions and their output is as follows:

```
> #
> # Obtaining randomized layout for a CRD
> #
> n <- 21
> CRDRand.unit <- list(Run = n)
> Treatment <- factor(rep(1:7, times = 3))
> CRDRand.lay <- fac.layout(unrandomized=CRDRand.unit,
+                           randomized=Treatment, seed=911)
> CRDRand.lay
   Units Permutation Run Treatment
1      1           7   1         3
2      2           4   2         6
3      3          20   3         7
4      4          15   4         2
5      5          11   5         6
6      6          18   6         1
7      7          19   7         1
8      8           9   8         2
9      9          12   9         1
10    10          10  10         3
11    11          13  11         5
12    12          14  12         2
13    13           2  13         4
14    14           3  14         5
15    15           6  15         4
16    16           8  16         7
17    17           1  17         4
18    18          17  18         6
19    19          21  19         7
20    20           5  20         3
21    21          16  21         5
> #remove Treatment object in workspace to avoid using it by mistake
> remove(Treatment)
```

III.5 An advertising agency conducts an experiment to assess the effectiveness of various formats of TV commercials. Forty regular viewers are randomly assigned to one of four formats of a TV commercial for a wine. Each viewer is interviewed and a score measuring impact recorded. The results are as follows:

FORMAT			
A	B	C	D
20	28	33	33
23	27	34	29
21	22	25	31
23	28	26	29
26	23	27	27
24	29	33	25
26	27	25	26
23	25	32	26
20	28	25	33
24	21	34	32

The response variable Impact has been saved in *CRDForm.dat.rda* and is available from the Statistical Modelling resources web site. It has been entered from the above table by columns. Add the 40-level factor Viewer and the factor Format with levels A–D to this data frame. This can be done by assigning factors to `CRDForm.dat$Viewer` and `CRDForm.dat$Format`, using the `rep` function for Format to save on typing. However, you will have to be careful about your use of `each` and `times` to ensure that the values that you generate line up with the values of Impact.

Obtain boxplots for each format as an initial exploration of the data. (Note: Appendix C, *Analysis of designed experiments in R*, contains a summary of the instructions for analysing a completely randomized design.)

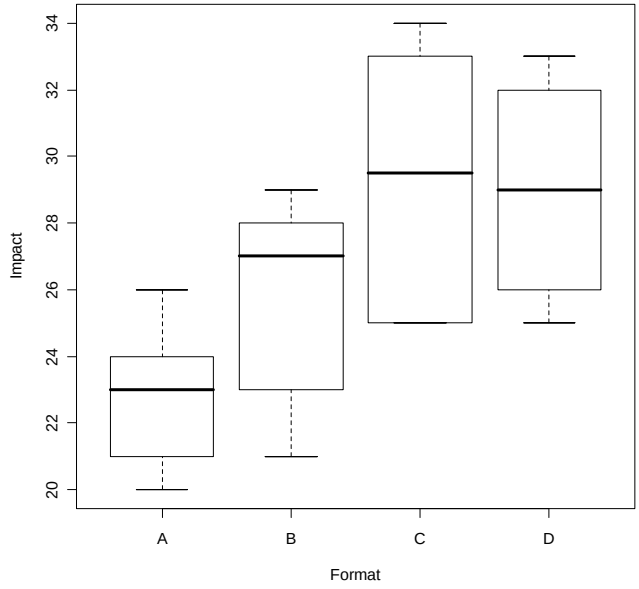
Use the R function `aov` to perform an analysis of variance on this data to determine if there is any measurable effect of format on the impact score.

Use R to obtain a residuals-versus-fitted-values plot and a normal probability plot as a check on the assumptions underlying the analysis. (Note that you will need to use the functions `resid.error` and `fitted.error` from the `dae` package. For instructions on the installation and use of `dae` see the Statistical Modelling resources web site.

Also, use Tukey's HSD procedure to determine exactly which Formats differed.

```
> load("CRDForm.dat.rda")
> #add factors Students and Format to data frame
> CRDForm.dat <- data.frame(Viewer = factor(1:40),
+              Format = factor(rep(1:4, each=10), labels=c("A","B","C","D")),
+                                          CRDForm.dat)
> attach(CRDForm.dat)
>#
# initial analyses
#
> boxplot(split(Impact, Format), xlab = "Format", ylab = "Impact")
> TV.aov <- aov(Impact ~ Format + Error(Viewer), CRDForm.dat)
> summary(TV.aov)

Error: Viewer
          Df Sum Sq Mean Sq F value    Pr(>F)
Format     3 274.88   91.62   9.454 9.616e-05
Residuals 36 348.90    9.69
```

```
> #
# plots for diagnostic checking
#
> res <- resid.errors(TV.aov)
> fit <- fitted.errors(TV.aov)
> plot(fit, res, pch = 16)
> qqnorm(res, pch = 16)
> qqline(res)
```



There would appear to be both differences in the medians and differences in spread between the formats.

Step 1: Set up hypotheses

H_0: $\alpha_A = \alpha_B = \alpha_C = \alpha_D = \mu$ (or $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$)

H_1: not all population Format means are equal (or $\boldsymbol{\psi}_F = \mathbf{X}_F\boldsymbol{\alpha}$)

Set $\alpha = 0.5$.

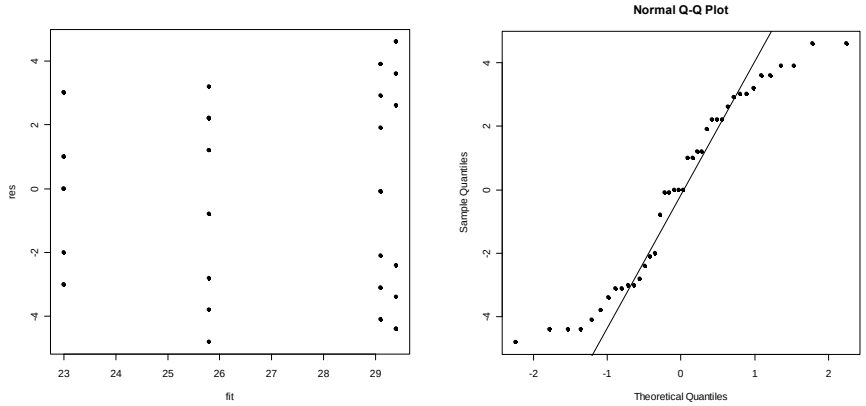
Step 2: Calculate test statistic

The analysis of variance table for the example is:

Source	df	SSQ	MSQ	F	Prob
Viewers	39	623.78			
Formats	3	274.90	91.63	9.45	<0.001
Residual	36	348.90	9.69		

Step 3: Decide between hypotheses

The probability of exceeding an F of 9.45 with $\nu_1 = 3$ and $\nu_2 = 36$ is < 0.001 < 0.05. Strong evidence of a difference between the formats and the maximal expectation model $\boldsymbol{\psi}_F = \mathbf{X}_F\boldsymbol{\alpha}$ best describes the data.



The homogeneity of variance assumption appears to be met as the size of the bands in the residual-versus-fitted-values plot are about equal. The normality assumption appears not be met as there is distinct curvature in the normal probability plot. However, the normality assumption is not crucial so the analysis can be used.

Differences between all pairs of Format means

```
> #
# multiple comparisons
#
> model.tables(TV.aov, type="means")
Tables of means
Grand mean

26.825

 Format
Format
   A    B    C    D
23.0 25.8 29.4 29.1
> q <- qtukey(0.95, 4, 36)
> q
[1] 3.808798
```

$$w(5\%) = \frac{3.809}{\sqrt{2}} \times \sqrt{\frac{9.69 \times 2}{10}} = \frac{3.809}{\sqrt{2}} \times 1.392 = 3.75$$

Differences between all pairs of Dose means

Format		A	B	D	C
	Mean	23.0	25.8	29.1	29.4
A	23.0				
B	25.8	2.8			
D	29.1	6.1	3.3		
C	29.4	6.4	3.6	0.3	
	w(5%)		3.75		

Only C and D are significantly different form A: B is not significantly different to any other format.

III.6 An experiment was conducted to test the effect of adding a small percentage of coal dust to the sand used for making concrete. Twenty batches of concrete were mixed under as identical conditions as possible and one of the five chosen additions of coal selected at random for each batch so that each addition occurred four times. From each batch a cylinder of concrete was made and tested for the breaking strength in pounds per square inch (lb/in^2).

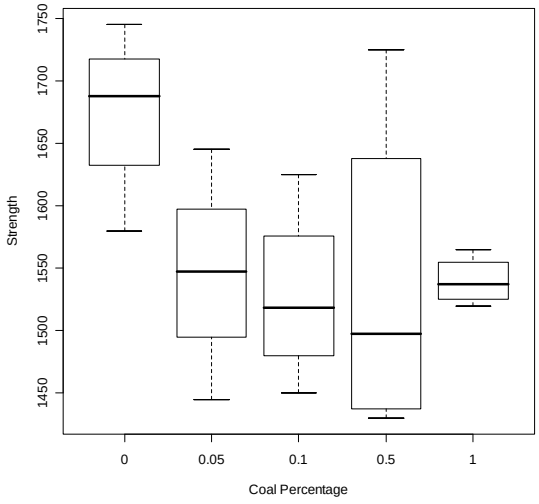
PERCENTAGE COAL				
0	0.05	0.1	0.5	1.0
1690	1550	1625	1725	1530
1580	1445	1450	1550	1545
1745	1645	1510	1430	1565
1685	1545	1527	1445	1520

This response variable Strength has been saved in *CRDConcr.dat.rda* and is available from the Statistical Modelling resources web site. As in exercise III.5, you will need to add the factor Batch and Cola.Percent (or some similar names). Obtain boxplots for each format as an initial exploration of the data.

Use R to perform an analysis of variance on this data, including the checking of assumptions. Investigate the fitting of a quadratic response to the treatment means and obtain a plot of the means together with the fitted quadratic. (Hint: the `ylim` argument of `plot` will be useful in changing the range on the y-axis.)

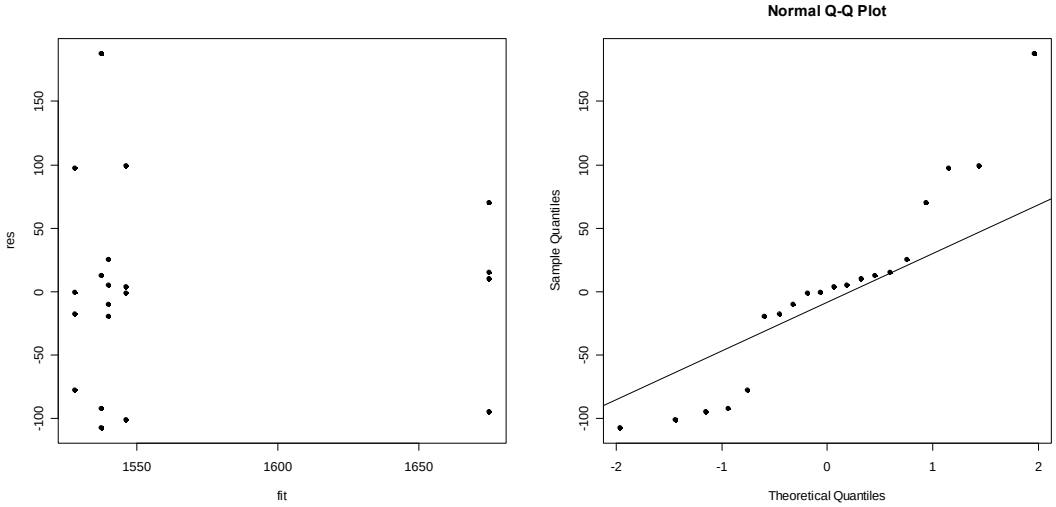
```
> #add factors Batch and Coal.Percent to data frame
> load("CRDConcr.dat.rda")
> CRDConcr.dat$Batch <- factor(1:20)
> CRDConcr.dat$Coal.Percent <- factor(rep(c(0,0.05,0.1,0.5,1), times=4))
> attach(CRDConcr.dat)
> boxplot(split(Strength, Coal.Percent), style.bxp = "old", medchar = T,
                                                          medpch = 8)
```

The initial boxplot is as follows:



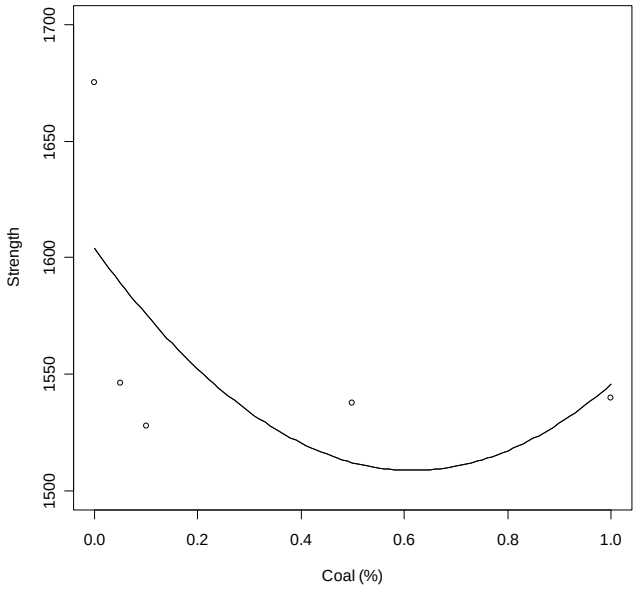
It certainly seems that the Strength when no coal is added is greater than any of the treatments where coal is added. There is evidence of variance heterogeity.

```
> CRDConcr.aov <- aov(Strength ~ Coal.Percent + Error(Batch), CRDConcr.dat)
> summary(CRDConcr.aov)

Error: Batch
             Df Sum Sq Mean Sq F value Pr(>F)
Coal.Percent  4  60805   15201  2.1378 0.1263
Residuals    15 106662    7111
> #
> # plots for diagnostic checking
> #
> res <- resid.errors(CRDConcr.aov)
> fit <- fitted.errors(CRDConcr.aov)
> plot(fit, res, pch = 16)
> qqnorm(res, pch = 16)
> qqline(res)
```




```
> #
> # fit polynomials
> #
> t <- 5
> Coal.Percent.lev <- c(0,0.05,0.1,0.5,1)
> CRDConcr.dat$Coal.Percent <- ordered(CRDConcr.dat$Coal.Percent,
+                                      levels=Coal.Percent.lev)
> contrasts(CRDConcr.dat$Coal.Percent) <- contr.poly(t,
+                                      scores=Coal.Percent.lev)
> contrasts(CRDConcr.dat$Coal.Percent)
             .L          .Q          .C           ^4
0    -0.3894500  0.35604141 -0.63448472  0.344952668
0.05 -0.3304424  0.12910207  0.15179535 -0.806906827
0.1  -0.2714349 -0.07364748  0.70097180  0.479100928
0.5   0.2006258 -0.82481130 -0.28119097 -0.019164037
1     0.7907015  0.41331530  0.06290853  0.002017267
> CRDConcr.aov <- aov(Strength ~ Coal.Percent + Error(Batch), CRDConcr.dat)
> summary(CRDConcr.aov, split = list(Coal.Percent = list(L = 1, Q = 2,
+                                                        Dev = 3:4)))

Error: Batch
                   Df Sum Sq Mean Sq F value  Pr(>F)
Coal.Percent        4  60805   15201  2.1378 0.12633
  Coal.Percent: L   1  10768   10768  1.5144 0.23742
  Coal.Percent: Q   1  10741   10741  1.5105 0.23800
```

```
   Coal.Percent: Dev  2  39296   19648  2.7631 0.09514
 Residuals           15 106662    7111
> #
> #get equation of fitted quadratic
> #
> CP <- as.vector(Coal.Percent)
> CP <- as.numeric(CP)
> CP2 <- CP*CP
> CRDConcr.lm <- lm(Strength ~ CP + CP2)
> coef(CRDConcr.lm)
(Intercept)          CP         CP2
  1604.0062   -308.9594    250.6947
> #
> # plot means and fitted line
> #
> CRDConcr.tab <- model.tables(CRDConcr.aov, type="means")
> Coal.Percent.Mean <- CRDConcr.tab$tables$Coal.Percent
> plot(x=Coal.Percent.lev, y=Coal.Percent.Mean, ylim=c(1500,1700),
+      xlab="Coal (%)", ylab="Strength")
> CRDConcr.coef <- coef(CRDConcr.lm)
> coalx <- seq(0, 1, 0.01)
> Coal.Percent.Fit <- CRDConcr.coef[[1]] + CRDConcr.coef[[2]]*coalx +
+                                          CRDConcr.coef[[3]]*coalx*coalx
> lines(x=coalx, y=Coal.Percent.Fit, type="l")
```



Step 1: Set up hypotheses

a) H_0: $\alpha_k - \mu - \gamma_1 x_k - \gamma_2 x_k^2 = 0$ for all k
   H_1: $\alpha_k - \mu - \gamma_1 x_k - \gamma_2 x_k^2 \neq 0$ for all k

b) H_0: $\gamma_2 = 0$
   H_1: $\gamma_2 \neq 0$

c) H_0: $\gamma_1 = 0$
   H_1: $\gamma_1 \neq 0$

Set $\alpha = 0.5$.

Step 2: Calculate test statistics

Source	df	SSQ	MSQ	F	Prob
Batches	19	167467.			
Coal	4	60805.	15201.	2.14	0.126
Linear	1	10768.	10768.	1.51	0.237
Quadratic	1	10741.	10741.	1.51	0.238
Deviations	2	39296.	19648.	2.76	0.095
Residual	15	106662.	7111.		

Step 3: Decide between hypotheses

As none of the F values are significant, it would appear that there is no difference between the five coal additions and the minimal expectation model $\boldsymbol{\psi}_G = \mathbf{X}_G\mu$ *appears to best describe the data.*

The plot of the fitted quadratic also confirms that it does not provide a good description of the trend in the strength means. From this plot it would appear that perhaps there is a difference between the treatment with no added coal and those for which coal has been added and that there is little difference between any of the treatments with added coal. A nested-factorial model reflecting this could be fitted and tested — such models are discussed in chapter VI, Factorial experiments.

Note that the residual-versus-fitted-values plot looks satisfactory, except for a single rather large residual, so that the homogeneity of variance assumption appears to be met. Also, the only problem with the normal probability plot appears to be a single large outlier. Hence, the normality assumption cannot be rejected. However, the reason for the large outlier warrants further investigation to see if a reason for it can be established.