Statistics for Engineering, 4C3/6C3

Assignment 6

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 20 March 2014**

Question 1 [10]

Notes

- Use computer software for questions 1, 2 and 3, but make sure you can do the work by hand if this were a 3-factor system.
- Group hand-ins are allowed.

We are considering a bioreactor system, investigating four factors:

- **A** = feed rate: slow or medium
- **B** = initial inoculant size (300g or 700g)
- **C** = feed substrate concentration (40 g/L or 60 g/L)
- **D** = dissolved oxygen set-point (4mg/L or 6 mg/L)

The 16 experiments from a full factorial, 2^4, were randomly run, and the yields from the bioreactor, y, are reported here in standard order: y = [60, 59, 63, 61, 69, 61, 94, 93, 56, 63, 70, 65, 44, 45, 78, 77].

1. Calculate the 15 main effects and interactions and the intercept, using computer software.
2. Use a Pareto-plot to identify the significant effects. What would be your advice to your colleagues to improve the yield?
3. Refit the model using only the significant terms identified in the second question.
    - Explain why you don't actually have to recalculate the least squares model parameters (one way to answer this question is to fit the full model manually, then refit it with the terms dropped out; what do you notice while doing the calculations?).
    - Compute the standard error and confidence intervals and confirm that the effects are indeed significant at the 95% level.
4. Write down the exact settings for **A**, **B**, **C**, and **D** you would provide to the graduate student running a half-fraction in 8 runs for this system.
5. Before the half-fraction experiments are even run you can calculate which variables will be confounded (aliased) with each other. Report the confounding pattern for these main effects and for these two-factor interactions.
    - Generator =
    - Defining relationship =
    - Confounding pattern:
        - $\widehat{\beta}_{\mathbf{A}} \rightarrow$
        - $\widehat{\beta}_{\mathbf{B}} \rightarrow$
        - $\widehat{\beta}_{\mathbf{AB}} \rightarrow$
        - $\widehat{\beta}_{\mathbf{BC}} \rightarrow$
        - $\widehat{\beta}_{\mathbf{CD}} \rightarrow$

6. Now use the 8 yield values corresponding to your half fraction, and calculate as many parameters (intercept, main effects, interactions) as you can.
    - Report their numeric values.
    - Compare your parameters from this half-fraction (8 runs) to those from the full factorial (16 runs). Was much lost by running the half fraction?

Solution

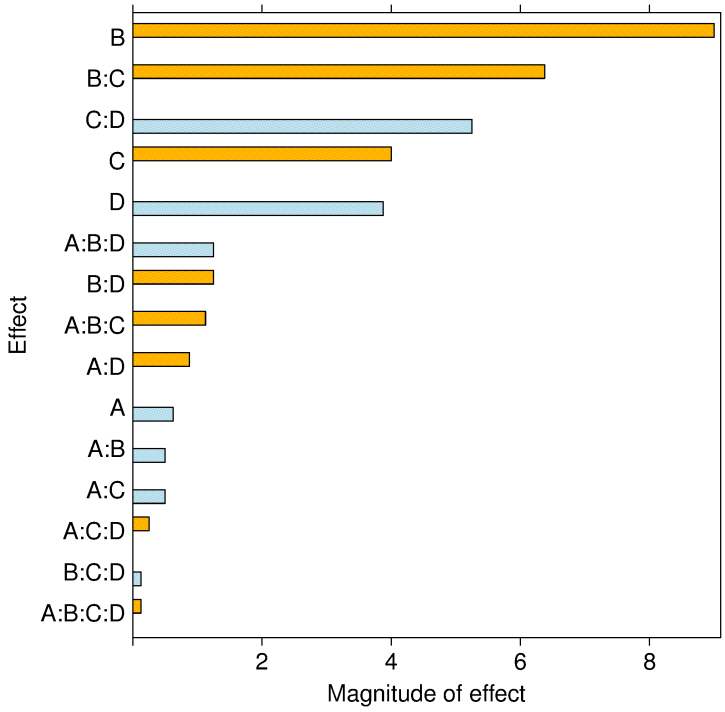
1. Using the computer code (at the end of the question), we found the complete model for all effects and interaction as:

$$\hat{y} = 66 - 0.6x_A + 9x_B + 4x_C - 3.9x_D - 0.5x_Ax_B - 0.5x_Ax_C + 0.9x_Ax_D + 6.4x_Bx_C + 1.3x_Bx_D - 5.3x_Cx_D + 1.1x_Ax_Bx_C - 1.2x_Ax_Bx_D + 0.3x_Ax_Cx_D - 0.1x_Bx_Cx_D + 0.1x_Ax_Bx_Cx_D$$

2. The Pareto plot shows the same important main effects: **B**, **C**, **D** and these two-factor interactions: **BC** and **CD**.

   The advice to improve yield would be to:
    - **A**: use either the slow or medium feedrate, whichever has the better process economics
    - **B**: operate with the larger inoculant size: 700g
    - **C**: use a higher feed concentration 60 g/L
    - **D**: use the lower dissolved oxygen set point of 4 mg/L
    - **BC**: in this case the **BC** interaction works in our favour (high $\times$ high)
    - **CD**: the **CD** interaction also works in our favour, since -5.3 $\times$ (+1) $\times$ (-1) leads to an increased yield.

   At these conditions the expected yield is in the region of 93 to 94% (runs 7 and 8 from the standard order).



3. The model does not have to be refitted because the columns in matrix $\mathbf{X}$ are orthogonal, meaning that the coefficient estimates do not depend on the levels of any other variables.

By dropping out the insignificant coefficients and keeping only the 5 parameters from the Pareto plus the intercept, we have 6 parameters, 16 data points, so 10 degrees of freedom. The residual vector is found from $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \underbrace{\mathbf{X}_{\text{sub}}}_{16\times 6}\underbrace{\mathbf{b}_{\text{sub}}}_{6\times 1}$.

The subset matrix of $\mathbf{X}_{\text{sub}}$ is found by sub-sampling from the full 16×16 matrix; similarly for the coefficient vector $\mathbf{b}$. From this we can calculate:

- The standard error is $S_E = 3.1$, which is pretty tight, considering the ranges of y-values in the data set
- The critical t-value for the 95% confidence level = 2.23
- The standard error for the parameters in the model is given by $\left(\mathbf{X}^T\mathbf{X}\right)^{-1} S_E^2$. We can use this form because apart from the intercept column, each column is centered around zero. So $S_E(b_i) = \sqrt{\frac{3.1^2}{16}}$ = 0.78.
- The confidence intervals for each of the significant effects are:

$$\begin{aligned} 7.3 \leq \quad & \beta_B \quad && \leq 10.7 \\ 2.3 \leq \quad & \beta_C \quad && \leq 5.7 \\ -5.6 \leq \quad & \beta_D \quad && \leq -2.1 \\ 4.6 \leq \quad & \beta_{BC} \quad && \leq 8.1 \\ -7.0 \leq \quad & \beta_{CD} \quad && \leq -3.5 \end{aligned}$$

4. A half-fraction of a 2^4 factorial has 8 experiments. We can generate the levels for 3 of the factors, **A**, **B** and **C** from a full factorial in these 8 runs. The generating term for the fourth factor **D** is best set to the highest level of confounding, the **ABC** term.

   Using that concept, we would ask the graduate student to run these 8 experiments in *random order*:

Experiment	Feed rate	Inoculant size	Feed concentration	DO set point
1	Slow	300g	40 g/L	4 mg/L
2	Medium	300g	40 g/L	6 mg/L
3	Slow	700g	40 g/L	6 mg/L
4	Medium	700g	40 g/L	4 mg/L
5	Slow	300g	60 g/L	6 mg/L
6	Medium	300g	60 g/L	4 mg/L
7	Slow	700g	60 g/L	4 mg/L
8	Medium	700g	60 g/L	6 mg/L

5.
   - Generator = **D = ABC**
   - Defining relationship = **I = ABCD**
   - Confounding pattern:
     - $\widehat{\beta}_{\mathbf{A}} \rightarrow$ **A + BCD**
     - $\widehat{\beta}_{\mathbf{B}} \rightarrow$ **B + ACD**
     - $\widehat{\beta}_{\mathbf{C}} \rightarrow$ **C + ABD**
     - $\widehat{\beta}_{\mathbf{D}} \rightarrow$ **D + ABC**
     - $\widehat{\beta}_{\mathbf{AB}} \rightarrow$ **AB + CD**
     - $\widehat{\beta}_{\mathbf{AC}} \rightarrow$ **AC + BD**
     - $\widehat{\beta}_{\mathbf{AD}} \rightarrow$ **AD + BC**
     - $\widehat{\beta}_{\mathbf{BC}} \rightarrow$ **BC + AD**
     - $\widehat{\beta}_{\mathbf{BD}} \rightarrow$ **BD + AC**

- $\widehat{\beta}_{\mathbf{CD}} \rightarrow$ **CD + AB**

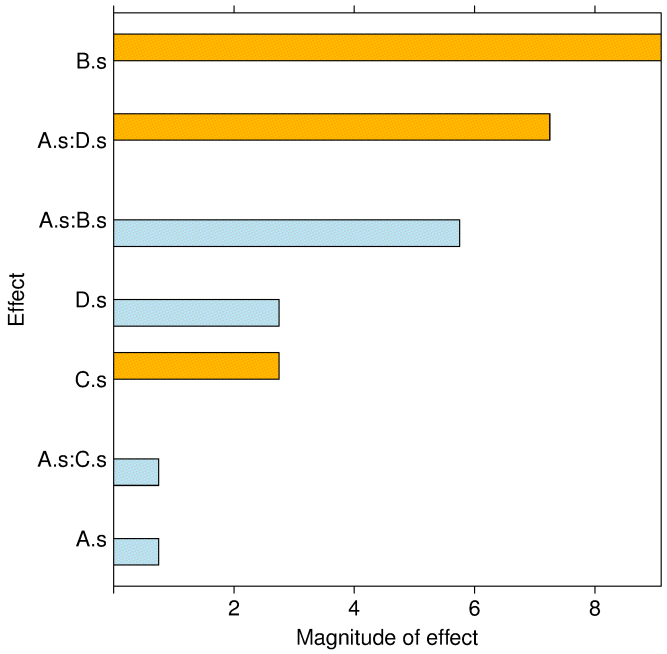
6. Selecting the rows from the full factorial design which correspond to the 8 runs from the half factorial we get $y = [60, 63, 70, 61, 44, 61, 94, 77]$ corresponding to the table order in question 5.

   Then forming the $\mathbf{X}$ matrix from the table in question 5 we solve for the parameters as follows:

   - $\widehat{b}_0 = 66.25 \rightarrow$ **I + ABCD**
   - $\widehat{b}_{\mathbf{A}} = -0.75 \rightarrow$ **A + BCD** (previous estimate for **A** was -0.625)
   - $\widehat{b}_{\mathbf{B}} = 9.25 \rightarrow$ **B + ACD** (previous estimate for **B** was 9.9)
   - $\widehat{b}_{\mathbf{C}} = 2.75 \rightarrow$ **C + ABD** (previous estimate for **C** was 4.0)
   - $\widehat{b}_{\mathbf{D}} = -2.75 \rightarrow$ **D + ABC** (previous estimate for **A** was -3.9)
   - $\widehat{b}_{\mathbf{AB}} = -5.75 \rightarrow$ **AB + CD** (previous estimate for **AB** was insignificant, while **CD** was -5.25)
   - $\widehat{b}_{\mathbf{AC}} = 0.75 \rightarrow$ **AC + BD** (previous estimates for both **AC** and **BD** were insignificant)
   - $\widehat{b}_{\mathbf{AD}} = 7.25 \rightarrow$ **AD + BC** (previous estimate for **AD** was insignificant, while **BC** was 6.4)

   You can verify for yourself that each coefficient from the half fraction is just the sum of the effects estimated from the full factorial. For example, $\widehat{b}_{\mathbf{AD}} = 7.25 \rightarrow$ **AD + BC** = 0.875 + 6.375 = 7.25.

   So these estimates from the half-fraction are comparable to the estimates from the full fraction and are shown below:



R code for this question

```
# Generate the design matrix
A <- B <- C <- D <- c(-1, 1)
f <- expand.grid(A=A, B=B, C=C, D=D)
A <- f$A
B <- f$B
C <- f$C
D <- f$D
# Set the response values in standard order, and solve thefull factorial model
```

```
y <- c(60, 59, 63, 61, 69, 61, 94, 93, 56, 63, 70, 65, 44, 45, 78, 77)

mod.full <- lm(y ~ (A+B+C+D)^4)
b <- coef(mod.full)

# Pareto plot
coeff.full <- coef(mod.full)[2:length(coef(mod.full))]
library(lattice)



bitmap('bioreactor-pareto-plot.png', type="png256", width=8,
        height=8, res=300, pointsize=14)
library(lattice)
coeff <- sort(abs(coeff.full), index.return=TRUE)
barchart(coeff$x,
         xlim=c(0, max(abs(coeff.full))+0.1),
         xlab=list("Magnitude of effect", cex=1.5),
         ylab = list("Effect", cex=1.5),
         groups=(coeff.full>0)[coeff$ix], col=c("lightblue", "orange"),
         scales=list(cex=1.5)
)
dev.off()

# Refit the model with only: B, C, D, BC, CD and intercept
mod.partial <- lm(y ~ B + C + D + B*C + C*D)
summary(mod.partial)
# and check confidence intervals
confint(mod.partial)

# Half-fraction generated from D = A*B*C defining relationship I = ABCD

# Create a logical vector, indicating which subset of the full runs to use:
subset <- D == A*B*C
A.s <- A[subset]
B.s <- B[subset]
C.s <- C[subset]
D.s <- D[subset]

y[subset]

mod.frac <- lm(y[subset] ~ A.s + B.s + C.s + D.s + A.s*B.s + A.s*C.s + A.s*D.s)
summary(mod.frac)
coeff.frac <- coef(mod.frac)[2:length(coef(mod.frac))]

bitmap('bioreactor-pareto-plot-half-fraction.png', type="png256", width=8,
       height=8, res=300, pointsize=14)

coeff <- sort(abs(coeff.frac), index.return=TRUE)
barchart(coeff$x,
         xlim=c(0, max(abs(coeff.full))+0.1),
         xlab=list("Magnitude of effect", cex=1.5),
         ylab = list("Effect", cex=1.5),
         groups=(coeff.full>0)[coeff$ix], col=c("lightblue", "orange"),
         scales=list(cex=1.5)
)
dev.off()
```