Key for Long Problem Set 6

(1). The file Acetone.RData contains results from a study to determine the amount of acetone in aqueous solutions of cellulose acetate. The analytical method requires a chemical disintegration of the cellulose acetate followed by an analysis for acetone. The effect of the disintegration step on the amount of acetone was investigated using the following 2^3 full-factorial design

factor/level	–1	+1
A: pH of solvent	acidic	basic
B: solvent (%water)	100	0
C: disintegration time (min)	3.00	6.00

where the solvent is a mixture of water and methanol. Analyze the data by first finding the full-factorial model, including all possible main effects and interactions. Use a qqnorm plot to evaluate the model's parameters and identify those that are significant. Reanalyze the data using this simpler model and comment on your results.

Answer. A complete model for a 2^3 full-factorial design includes eight terms: an intercept, three first-order effects in the factors, three binary interactions between the factors, and a ternary interaction between the factors

$$y = \beta_0 + \beta_a A + \beta_b B + \beta_C C + \beta_{ab} AB + \beta_{ac} AC + \beta_{bc} BC + \beta_{abc} ABC$$

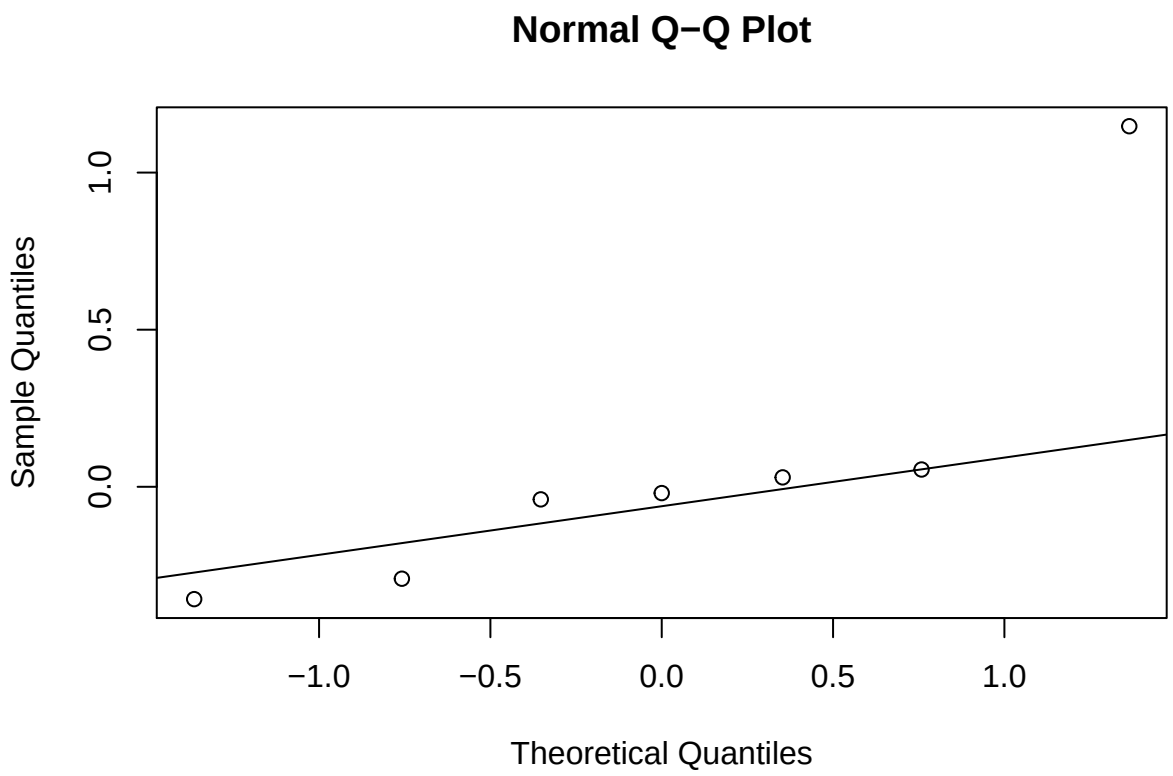
where y is the concentration of acetone, A is pH, B is the solvent, and C is time. Evaluating the model, we find that

```
load("Acetone.RData")
lm.acetone = lm(percent ~ pH * solvent * time, data = acetone)
summary(lm.acetone)
```

```
##
## Call:
## lm(formula = percent ~ pH * solvent * time, data = acetone)
##
## Residuals:
## ALL 8 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.2675         NA      NA       NA
## pH                1.1475         NA      NA       NA
## solvent          -0.2925         NA      NA       NA
## time              0.0550         NA      NA       NA
## pH:solvent       -0.3575         NA      NA       NA
## pH:time           0.0300         NA      NA       NA
## solvent:time     -0.0200         NA      NA       NA
## pH:solvent:time  -0.0400         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

Because the number of experiments equals the number of parameters in our model, there are no degrees of freedom for evaluating the significance of these coefficients; thus, we must rely on a qqnorm plot to identify those coefficients that likely differ significantly from zero.

```
qqnorm(lm.acetone$coeff[-1])
qqline(lm.acetone$coeff[-1])
```



Examining the qqnorm plot we see that four values fall, more or less, along the straight-line that represents results consistent with a normal distribution centered at zero. The remaining three coefficients—for pH (1.1475), for solvent (–0.3575), and for pH:solvent (–0.2925)—deviate from the qqline and likely are significant coefficients; a simpler model, therefore, is

$$y = \beta_0 + \beta_a A + \beta_b B + \beta_{ab} AB$$

where A is pH and B is the solvent. Evaluating this model

```
lm.acetone = lm(percent ~ pH * solvent, data = acetone)
summary(lm.acetone)
```

```
##
## Call:
## lm(formula = percent ~ pH * solvent, data = acetone)
##
## Residuals:
##      1      2      3      4      5      6      7      8
## -0.005 -0.145 -0.045 -0.025  0.005  0.145  0.045  0.025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.26750    0.03849 136.864 1.71e-08 ***
## pH           1.14750    0.03849  29.815 7.54e-06 ***
```

```
## solvent     -0.29250    0.03849  -7.600 0.001608 **
## pH:solvent  -0.35750    0.03849  -9.289 0.000747 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1089 on 4 degrees of freedom
## Multiple R-squared:  0.9961, Adjusted R-squared:  0.9932
## F-statistic: 344.3 on 3 and 4 DF,  p-value: 2.786e-05
```

suggests that all three coefficients and the intercept are significant ($p << 0.001$). The positive coefficient for pH suggests that more a basic pH level favors the formation of acetone. The negative coefficient for the solvent suggests that more acetone forms when the water:methanol mixture includes more water than methanol. Finally, the negative coefficient for the interaction between pH and the solvent is consistent with the independent effects of pH and solvent: more basic solutions (+1) for pH and more water than methanol (–1) favor a higher concentration of acetone. Note that the pH:solvent term also favors more acidic solutions (–1) with more methanol than water (–1), but the main effects for pH and for solvent make this a less desirable option.

(2). An alternative approach to determining significant factors is to estimate the standard deviation for factor effects (s_{FE}) by making duplicate runs at each set of factor levels. The variance for the difference between the duplicate runs is

$$s^2 = \frac{\sum d_i^2}{2n}$$

where d_i is the difference between results for a given set of factor levels and n is the number of different factor levels (that is, $n = 2^k$). The standard deviation for factor effects is

$$s_{FE} = \sqrt{\frac{2s^2}{n}}$$

A coefficient that falls outside a confidence interval of $0 \pm t(\alpha, \nu)s_{FE}$ is considered significant. The degrees of freedom is the number of unique factor level; that is, $\nu = n$. Use this approach and the objects "trial.one" and "trial.two" in the file Acetone.RData to reanalyze the data from the previous problem and compare your conclusions to those determined earlier.

Answer. To calculate s_{FE} we use the following code (there are lots of ways to arrive at the standard deviation for factor effects; however, regardless of your code, you should arrive at the same result):

```
s.sqr = sum((trial.one - trial.two)^2)/(2 * length(trial.one))
s.sqr
```

```
## [1] 0.00538125
```

```
s.fe = sqrt((2*s.sqr)/length(trial.one))
s.fe
```

```
## [1] 0.0366785
```

To establish the confidence interval, we note that the value of t for $\alpha = 0.05$ and for eight degrees of freedom is $t(0.05, 8) = 2.31$, which gives the confidence interval as

$$0 \pm t(0.05, 8)s_{FE} = 0 \pm (2.31)(0.03668) = 0 \pm 0.0847$$

For a coefficient that falls within this confidence interval, there is reason to believe that it is explained by random error and that it does not differ significantly from zero; this is true for all but the intercept, pH, solvent, and pH:solvent, in agreement with our results in the previous problem.

(3). The file tRNA.RData contains results from the study of the esterification of tRNA by arginine. The parameters are the pH, the amount of enzyme used to catalyze the reaction, and the amount of arginine

used. The reaction was monitored by using ^{14}C-labeled arginine and measuring the amount of radioactivity as counts. The following central-composite design was used

factor/level	–1.7	–1.0	0	+1.0	+1.7
A: enzyme (mg)	3.2	6.0	10.0	14.0	16.8
B: arginine (pmol)	860	1000	1200	1400	1540
C: pH	6.6	7.0	7.5	8.0	8.4

Develop a suitable model that predicts the counts as a function of the available factors, retaining terms where p is less than 0.10. Does your model predict successfully the intercept? Create a perspective plot that displays counts on the z-axis as a function of the two most important factors; if your model includes a third factor, then set its value to a level of zero. What is the expected count for an experiment that uses 7.0 mg of enzyme, 1300 pmol of arginine, and a pH of 7.5?

Answer. With a central-composite design, we have sufficient information to explore both first-order and second-order effects for each factor. A good starting point for a model is to include an intercept, the three first-order effects, the three second-order effects, and three binary, first-order interactions; other higher order interactions are unlikely. Our model, therefore, is

$$y = \beta_0 + \beta_a A + \beta_b B + \beta_c C + \beta_{aa} A^2 + \beta_{bb} B^2 + \beta_{cc} C^2 + \beta_{ab} AB + \beta_{ac} AC + \beta_{bc} BC$$

where y is the counts, A is the mg of enzyme, B is the pmol of arginine, and C is the pH. Evaluating this model

```
load("tRNA.RData")
lm.trna = lm(counts ~ enzyme * arginine * pH + I(enzyme^2)
             + I(arginine^2) + I(pH^2) - enzyme:arginine:pH, data = tRNA)
summary(lm.trna)
```

```
##
## Call:
## lm(formula = counts ~ enzyme * arginine * pH + I(enzyme^2) +
##     I(arginine^2) + I(pH^2) - enzyme:arginine:pH, data = tRNA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.984  -44.604   -8.723   53.167  141.539
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5013.51      42.75 117.275  < 2e-16 ***
## enzyme            143.39      28.22   5.081 0.000477 ***
## arginine          -28.11      28.22  -0.996 0.342762
## pH                 42.40      28.22   1.503 0.163860
## I(enzyme^2)       -71.73      27.15  -2.642 0.024668 *
## I(arginine^2)    -129.51      27.15  -4.770 0.000757 ***
## I(pH^2)           -63.08      27.15  -2.323 0.042557 *
## enzyme:arginine   -71.75      37.04  -1.937 0.081453 .
## enzyme:pH          14.25      37.04   0.385 0.708481
## arginine:pH       -22.00      37.04  -0.594 0.565709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 104.8 on 10 degrees of freedom
## Multiple R-squared:  0.8634, Adjusted R-squared:  0.7405
## F-statistic: 7.025 on 9 and 10 DF,  p-value: 0.002665
```

suggests that the following coefficients are significant: the intercept (5013.51), the enzyme's first-order (143.39) effect and second-order effect (–71.73), arginine's second-order effect (–129.51), the pH level's second-order effect (–63.08), and the interaction between arginine and the enzyme (–71.75), or

$$y = beta_0 + \beta_a A + \beta_{aa} A^2 + \beta_{bb} B^2 + \beta_{cc} C^2 + \beta_{ab} AB$$

where y is the counts, A is the mg of enzyme, B is the pmol of arginine, and C is the pH. To test the model, we use the six replicate trials at the center of the central composite design (trials 15–20). A t-test of the mean for these results against the model's intercept

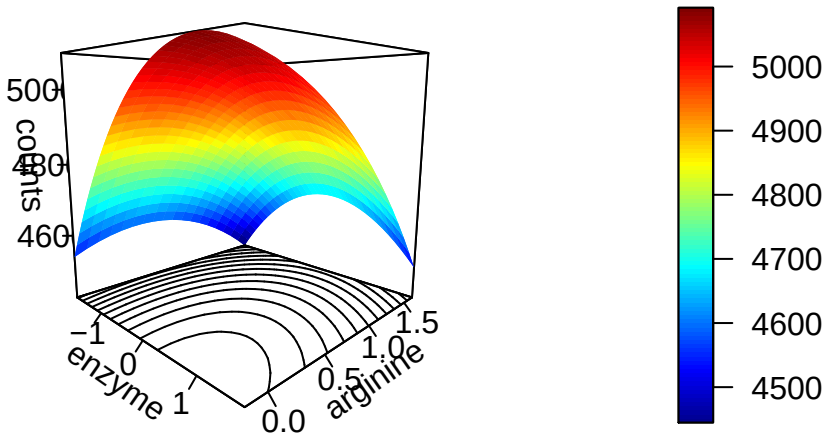
```
t.test(tRNA$counts[15:20], mu = lm.trna$coeff[1], conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  tRNA$counts[15:20]
## t = 0.010766, df = 5, p-value = 0.9918
## alternative hypothesis: true mean is not equal to 5013.513
## 95 percent confidence interval:
##  4937.464 5090.203
## sample estimates:
## mean of x
##  5013.833
```

suggests that random error in the measurements is more than sufficient to explain the difference between the mean of 5013.83 for the replicates to the model's intercept of 5013.51; thus, we have good confidence in our model.

The following code creates a surface plot; note that the term for pH^2 is omitted as its factor level is set to zero.

```
enz = seq(-1.7, 1.7, 0.1)
arg = seq(-.17, 1.7, 0.1)
model = function(enz, arg)
  {5013.51 + 143.39 * enz - 129.51 * arg^2 - 71.73 * enz^2 - 71.75 * enz * arg}
z = outer(enz, arg, model)
library(plot3D)
persp3D(enz, arg, z, xlab = "enzyme", ylab = "arginine", zlab = "counts",
        phi = 10, theta = 45, ticktype = "detailed", contour = TRUE)
```



To determine the counts for 7 mg of enzyme, 1300 pmoles of arginine, and a pH of 7.5, we need to convert each into its corresponding coded value; these are –0.75 for the enzyme, +0.5 for arginine, and 0 for the pH. Substituting these back into or model gives the counts as 4860.