Question #1.

Solution 1. *For part (a):*

The model here is in the formof :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ X_{13} & X_{23} \\ X_{14} & X_{24} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

and $E[\epsilon_i] = 0$ *for* $i = 1, 2, 3, 4$.

To make this a normal theory Gauss Markov model, or in other words, to make the ordinary least square estimate become the best linear unbiased estimator(BLUE), we need to impose the following assumptions:

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

which implies normal, independence and equal variance.

For part (b):

If conditioned on $X_{1i} = X_{1j}$ *for* $i \neq j$, *assume* $X_{2i} = 1$ *and* $X_{2j} = -1$, *then we have:*

$$E[y_i] = \beta_1 X_{1i} + \beta_2$$
$$E[y_j] = \beta_1 X_{1j} - \beta_2$$

So

$$E[y_i] - E[y_j] = 2\beta_2 \Longrightarrow \beta_2 = \frac{1}{2}\Big(E[y_i] - E[y_j]\Big)$$

So β_2 *means on average half amount of difference in escaped vapor between using or not using the devie, while conditioned on the same temperature.*

For part (c):

If the condition in part (a) is satisfied, then our model is a Gauss-Markov model, so $\mathbf{b}$ *is the best linear unbiased estimator of* $\boldsymbol{\beta}$*(BLUE), or in other words,* $\mathbf{b}$ *has the minimum variance among all linear estimators of* $\boldsymbol{\beta}$ *that are unbiased.*

For part (d):

We have:

$$\mathbf{c}^T\mathbf{y} = (y_3 + y_4 - y_1 - y_2) \sim N(\mathbf{c}^T\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{c}^T\mathbf{c})$$

with

$$\begin{aligned}
\mathbf{c}^T\mathbf{X}\beta &= (-1,-1,1,1)\begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ X_{13} & X_{23} \\ X_{14} & X_{24} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \\
&= (-1,-1,1,1)\begin{bmatrix} 0 & -1 \\ 30 & -1 \\ 0 & 1 \\ 30 & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \\
&= [0,4]\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \\
&= 4\beta_2
\end{aligned}$$

and

$$\mathbf{c}^T\mathbf{c} = 4$$

So we have:

$$\hat{\beta}_2 = \frac{1}{4}(Y_3 + Y_4 - Y_1 - Y_2)$$

Meanwhile the estimates for σ^2 *is:*

$$\hat{\sigma}^2 = \frac{MSE}{\sum(X_{2i} - \overline{X}_2)^2} = \frac{SSE/(4-2)}{\sum(X_{2i} - 0)^2} = \frac{SSE/2}{4} = \frac{SSE}{8}$$

So we have

$$\frac{\hat{\beta}_2 - 0}{\hat{\sigma}} = \frac{\frac{1}{4}(Y_3 + Y_4 - Y_1 - Y_2)}{\sqrt{SSE/8}} = \frac{Y_3 + Y_4 - Y_1 - Y_2}{\sqrt{2 * SSE}} \sim t(2)$$

under the null hypothesis

$$H_0 : \beta_2 = 0$$

which implies

$$\frac{(\hat{\beta}_2 - 0)^2}{\hat{\sigma}^2} = \frac{(Y_3 + Y_4 - Y_1 - Y_2)^2}{2 * SSE} \sim F(1,2)$$

so the numerator degree of freedom is 1*, denominator degree of freedom is* 2*.*

For part (e)*:*

Given the F *test in part* (d)*, we have:*

$$H_0 : \beta_2 = 0 \text{ versus } H_1 : \beta_2 \neq 0$$

For part (f)*:*

Suppose we add te interaction effect, then our model becomes

$$Y_i = \beta_1(X_{1i}) + \beta_2(X_{2i}) + \beta_3(X_{1i} \cdot X_{2i}) + \epsilon_i$$

then our design matrix becomes

$$\mathbf{X} = \begin{bmatrix} 0 & -1 & 0 \\ 30 & -1 & -30 \\ 0 & 1 & 0 \\ 30 & 1 & 30 \end{bmatrix}$$

we can see that the rank of $\mathbf{X}$ *is* 3 *since apparently the* 3 *columns of* $\mathbf{X}$ *are linearly independent. Since the design matrix is still full rank, so we can get an estimate for* β_3 *which is the coefficient for interaction.*

Our initial SSE has degree of freedom of $4 - 2 = 2$*, so if we add the interaction term, it can still afford to lose one more degree of freedom.*

Question 2.

Solution 2. *For part* (a)*:*

The "unconstrained" cell means model is given as :

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

in our case we have $i = 1, 2, 3$ *representing the level of cement and* $j = 1, 2, 3, 4$ *representing the level of aggregate.*

If we write the model in the matrix form it is:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ *is a column vector of length* 12 *and* $\mathbf{W}$ *is the design matrix with dimension* 33×12 *(we have* $N = 33$ *cell counts).*

```
/*Question 2*/
proc iml;

/*  design matrix */
    W = {1 0 0 0 0 0 0 0 0 0 0 0 , 1 0 0 0 0 0 0 0 0 0 0 0 , 1 0 0 0 0 0 0 0 0 0 0 0,
         0 1 0 0 0 0 0 0 0 0 0 0 , 0 1 0 0 0 0 0 0 0 0 0 0 , 0 1 0 0 0 0 0 0 0 0 0 0,
         0 0 1 0 0 0 0 0 0 0 0 0 , 0 0 1 0 0 0 0 0 0 0 0 0 ,
         0 0 0 1 0 0 0 0 0 0 0 0 , 0 0 0 1 0 0 0 0 0 0 0 0 , 0 0 0 1 0 0 0 0 0 0 0 0,
         0 0 0 0 1 0 0 0 0 0 0 0 , 0 0 0 0 1 0 0 0 0 0 0 0 , 0 0 0 0 1 0 0 0 0 0 0 0,
         0 0 0 0 0 1 0 0 0 0 0 0 , 0 0 0 0 0 1 0 0 0 0 0 0 , 0 0 0 0 0 1 0 0 0 0 0 0, 0 0 0 0 0 1 0 0 0 0 0 0,
         0 0 0 0 0 0 1 0 0 0 0 0 , 0 0 0 0 0 0 1 0 0 0 0 0 ,
         0 0 0 0 0 0 0 1 0 0 0 0 , 0 0 0 0 0 0 0 1 0 0 0 0 , 0 0 0 0 0 0 0 1 0 0 0 0,
         0 0 0 0 0 0 0 0 1 0 0 0 , 0 0 0 0 0 0 0 0 1 0 0 0 ,
         0 0 0 0 0 0 0 0 0 1 0 0 ,
         0 0 0 0 0 0 0 0 0 0 1 0 , 0 0 0 0 0 0 0 0 0 0 1 0 , 0 0 0 0 0 0 0 0 0 0 1 0,
         0 0 0 0 0 0 0 0 0 0 0 1 , 0 0 0 0 0 0 0 0 0 0 0 1 , 0 0 0 0 0 0 0 0 0 0 0 1 , 0 0 0 0 0 0 0 0 0 0 0 1};
    print W;
/*response y1*/
y = {21, 27, 19,  19,  19,  22,  19,  16,  23,  24,  23,  25,  23,  24,  23,  20,  24,  18,  19,  18,  28,  27,  25,  20,  24,
     28,  14,  16,  12,  23,  25,  22,  22};
/*create labels for matrices*/
cY = {"y"};
cW = {"mu_11" "mu_12" "mu_13" "mu_14" "mu_21" "mu_22" "mu_23" "mu_24" "mu_31" "mu_32" "mu_33" "mu_34"};
mattrib y colname=cY W colname=cW;

print W y;
```

The W design matrix looks like this:

W

	mu_11	mu_12	mu_13	mu_14	mu_21	mu_22	mu_23	mu_24	mu_31	mu_32	mu_33	mu_34
ROW1	1	0	0	0	0	0	0	0	0	0	0	0
ROW2	1	0	0	0	0	0	0	0	0	0	0	0
ROW3	1	0	0	0	0	0	0	0	0	0	0	0
ROW4	0	1	0	0	0	0	0	0	0	0	0	0
ROW5	0	1	0	0	0	0	0	0	0	0	0	0
ROW6	0	1	0	0	0	0	0	0	0	0	0	0
ROW7	0	0	1	0	0	0	0	0	0	0	0	0
ROW8	0	0	1	0	0	0	0	0	0	0	0	0
ROW9	0	0	0	1	0	0	0	0	0	0	0	0
ROW10	0	0	0	1	0	0	0	0	0	0	0	0
ROW11	0	0	0	1	0	0	0	0	0	0	0	0
ROW12	0	0	0	0	1	0	0	0	0	0	0	0
ROW13	0	0	0	0	1	0	0	0	0	0	0	0
ROW14	0	0	0	0	1	0	0	0	0	0	0	0
ROW15	0	0	0	0	0	1	0	0	0	0	0	0
ROW16	0	0	0	0	0	1	0	0	0	0	0	0
ROW17	0	0	0	0	0	1	0	0	0	0	0	0
ROW18	0	0	0	0	0	1	0	0	0	0	0	0
ROW19	0	0	0	0	0	0	1	0	0	0	0	0
ROW20	0	0	0	0	0	0	1	0	0	0	0	0
ROW21	0	0	0	0	0	0	0	1	0	0	0	0
ROW22	0	0	0	0	0	0	0	1	0	0	0	0
ROW23	0	0	0	0	0	0	0	1	0	0	0	0
ROW24	0	0	0	0	0	0	0	0	1	0	0	0
ROW25	0	0	0	0	0	0	0	0	1	0	0	0
ROW26	0	0	0	0	0	0	0	0	0	1	0	0
ROW27	0	0	0	0	0	0	0	0	0	0	1	0
ROW28	0	0	0	0	0	0	0	0	0	0	1	0
ROW29	0	0	0	0	0	0	0	0	0	0	1	0
ROW30	0	0	0	0	0	0	0	0	0	0	0	1
ROW31	0	0	0	0	0	0	0	0	0	0	0	1
ROW32	0	0	0	0	0	0	0	0	0	0	0	1
ROW33	0	0	0	0	0	0	0	0	0	0	0	1

We can find

$$s^2 = \frac{SSE}{v_E} = \frac{(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\mu}})'(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\mu}})}{N - ab} = \frac{(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\mu}})'(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\mu}})}{33 - 12}$$

To test main effect of cement (factor A), our null hypothesis is:

$$H_0 : \begin{cases} 2(\mu_{11} + \mu_{12} + \mu_{13} + \mu_{14}) = (\mu_{21} + \mu_{22} + \mu_{23} + \mu_{24}) + (\mu_{31} + \mu_{32} + \mu_{33} + \mu_{34}) \\ \mu_{21} + \mu_{22} + \mu_{23} + \mu_{24} = \mu_{31} + \mu_{32} + \mu_{33} + \mu_{34} \end{cases}$$

This is equivalent to the following generalized linear hypothesis:

$$H_0 : \mathbf{A}\boldsymbol{\mu} = 0$$

with $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}, \mu_{31}, \mu_{32}, \mu_{33}, \mu_{34})'$ *and*

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

Our test statistic is:

$$F = \frac{SSA/v_A}{SSE/v_E} = \frac{(\mathbf{A}\hat{\boldsymbol{\mu}})'[\mathbf{A}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{A}']^{-1}\mathbf{A}\hat{\boldsymbol{\mu}}/v_A}{SSE/v_E}$$

with

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} \\ v_A &= (a-1) = 2 \\ v_E &= N - ab = 33 - 12 = 21 \end{aligned}$$

We have SAS code as following:

```
A = {2 2 2 2 -1 -1 -1 -1 -1 -1 -1 -1, 0 0 0 0 1 1 1 1 -1 -1 -1 -1};
print A;

est_mu = inv(t(W)*W)*t(W)*y;
N = 33;
a = 3;
b = 4;
I = i(N);
H = W*inv(t(W)*W)*t(W);
SSE = t(y)*(I - H)*y;
s_2 = SSE/(N - a*b);
print SSE s_2;
```

A

	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12
ROW1	2	2	2	2	-1	-1	-1	-1	-1	-1	-1	-1
ROW2	0	0	0	0	1	1	1	1	-1	-1	-1	-1

SSE	s_2
97.75	4.6547619

Continue with the test, we give output for F statistics, SSA value and p value:

```
SSA = t(A*estMu)*inv(A*inv(t(W)*W)*t(A))*A*estMu;
v_A = factor_A - 1;
/*F statistic and p value for testing main effect of cement*/
F_A = (SSA/v_A);
p_A = 1 - CDF('F', F_A, v_A, v_E);
print SSA F_A p_A;
```

SSA	F_A	p_A
18.025332	9.0126658	0.0014934

the p value is significant (0.0015) *so we reject* H_0 *and conclude that there is a different main effect between different types of cement.*

Similarly we have the null hypothesis for testing the main effect of aggregate (factor B) as:

$$H_0 : \begin{cases} 3(\mu_{11}+\mu_{21}+\mu_{31}) = (\mu_{12}+\mu_{22}+\mu_{32}) + (\mu_{13}+\mu_{23}+\mu_{33}) + (\mu_{14}+\mu_{24}+\mu_{34}) \\ 2(\mu_{12}+\mu_{22}+\mu_{32}) = (\mu_{13}+\mu_{23}+\mu_{33}) + (\mu_{14}+\mu_{24}+\mu_{34}) \\ \mu_{13}+\mu_{23}+\mu_{33} = \mu_{14}+\mu_{24}+\mu_{34} \end{cases}$$

which is equivalent to the general linear hypothesis:

$$H_0 : \mathbf{B}\boldsymbol{\mu} = 0$$

with

$$B = \begin{pmatrix} 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Our test statistic is:

$$F = \frac{SSB/v_B}{SSE/v_E} = \frac{(\mathbf{B}\hat{\boldsymbol{\mu}})'[\mathbf{B}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{B}']^{-1}\mathbf{B}\hat{\boldsymbol{\mu}}/v_B}{SSE/v_E}$$

with

$$v_B = (b-1) = 3$$

Corresponding SAS code and output:

```
/*testing main effect of aggreagate*/
SSB = t(B*estMu)*inv(B*inv(t(W)*W)*t(B))*B*estMu;
v_B = factor_B - 1;
/*F statistic and p value for testing main effect of cement*/
F_B = (SSB/v_B)/s_2;
p_B = 1 - CDF('F', F_B, v_B, v_E);
print SSB F_B p_B;
```

SSB	F_B	p_B
256.76322	18.387136	4.3666E-6

the p value is highly sigificant so we reject the null hypothesis and conclude there is a main effect difference from different types of aggregate.

Finally we can use hadamad product from A and B to get the orthogonal contrast matrix for testing interaction. We have

$$C = \begin{pmatrix} 6 & -2 & -2 & -2 & -3 & 1 & 1 & 1 & -3 & 1 & 1 & 1 \\ 0 & 4 & -2 & -2 & 0 & -2 & 1 & 1 & 0 & -2 & 1 & 1 \\ 0 & 0 & 2 & -2 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 3 & -1 & -1 & -1 & -3 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{pmatrix}$$

The F statistic is:

$$F = \frac{SSC/v_{AB}}{SSE/v_E} = \frac{(\mathbf{C}\hat{\boldsymbol{\mu}})'[\mathbf{C}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\mu}}/v_{AB}}{SSE/v_E}$$

with

$$v_{AB} = (a-1) \times (b-1) = 2 \times 3 = 6$$

Corresponding SAS code and output:

```
/*testing interaction between cement and aggregate*/
SSAB = t(C*estMu)*inv(C*inv(t(W)*W)*t(C))*C*estMu;
v_AB = (factor_A - 1)*(factor_B - 1);
/*F statistic and p value for testing main effect of cement*/
F_AB = (SSAB/v_AB)/s_2;
p_AB = 1 - CDF('F', F_AB, v_AB, v_E);
print SSAB F_AB p_AB;
```

SSAB	F_AB	p_AB
84.498553	3.0255236	0.0271952

The p value is significant (0.027) *and we reject the null and conclude that there is interaction between cement and aggregate.*

For part (*b*)*:*

The "constrained" cell means model here with additivity is:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon} \text{ subject to } \mathbf{C}\mu = 0$$

where $\mathbf{C}$ *is the contrast matrix for testing interaction that we got from part* (a)*.*

We can reparametrize the model using:

$$\mathbf{A}_{new} = \begin{pmatrix} \mathbf{K} \\ \mathbf{C} \end{pmatrix}$$

here the first row of K *correspond to mulitiple of overall mean, and the reamaing rows of* K *could include the contrasts for main effects. So we have*

$$\mathbf{K} = \begin{pmatrix} \mathbf{j}' \\ \mathbf{A} \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{pmatrix}$$

and

$$\mathbf{A}_{new} = \begin{pmatrix} \mathbf{K} \\ \mathbf{C} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \\ 6 & -2 & -2 & -2 & -3 & 1 & 1 & 1 & -3 & 1 & 1 & 1 \\ 0 & 4 & -2 & -2 & 0 & -2 & 1 & 1 & 0 & -2 & 1 & 1 \\ 0 & 0 & 2 & -2 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 3 & -1 & -1 & -1 & -3 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{pmatrix}$$

thanks to the previous work in part (a)*, all the rows of* $\mathbf{A}_{new}$ *are mutually orthogonal and the rank of* $\mathbf{A}_{new}$ *is full rank (rank 12).*

So we can reparametrize the model now with

$$\begin{aligned} \mathbf{y} &= \mathbf{W}\mathbf{A}_{new}^{-1}\mathbf{A}_{new}\boldsymbol{\mu} + \boldsymbol{\epsilon} \text{ subject to } \mathbf{C}\boldsymbol{\mu} = \mathbf{0} \\ &= \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon} \text{ subject to } \mathbf{C}\boldsymbol{\mu} = \mathbf{0} \end{aligned}$$

where $\mathbf{Z} = \mathbf{W}\mathbf{A}_{new}^{-1}$ *and* $\boldsymbol{\delta} = \mathbf{A}_{new}\boldsymbol{\mu}$*.*

We follow the notation from the textbook so the hypothesis for testing the main effect of cement (factor A) is:

$$H_0 : \mathbf{A}\hat{\boldsymbol{\mu}}_c = \mathbf{0}$$

with F statistic:

$$F = \frac{(\mathbf{A}\hat{\boldsymbol{\mu}}_c)'[\mathbf{A}\mathbf{K}^*(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}(\mathbf{K}^*)'\mathbf{A}']^{-1}\mathbf{A}\hat{\boldsymbol{\mu}}_c/v_A}{SSE_c/v_{E_c}}$$

Similarly we have the hypothesis for testing the main effect of aggregate (factor B) under constrained model:

$$H_0 : \mathbf{B}\hat{\boldsymbol{\mu}}_c = \mathbf{0}$$

with F statistic:

$$F = \frac{(\mathbf{B}\hat{\boldsymbol{\mu}}_c)'[\mathbf{B}\mathbf{K}^*(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}(\mathbf{K}^*)'\mathbf{B}']^{-1}\mathbf{B}\hat{\boldsymbol{\mu}}_c/v_B}{SSE_c/v_{E_c}}$$

The following SAS code compute the necessary quantities needed in order to get F statistics, including $\mathbf{K}, \mathbf{K}^*, \mathbf{A}_{new}, \mathbf{Z}, \mathbf{Z}_1, \hat{\boldsymbol{\mu}}_c$ *and* SSE_c*. Specifically,*

$$\mathbf{K}^* = \mathbf{K}'(\mathbf{K}\mathbf{K}')^{-1}$$

and $\mathbf{Z}_1$ *is the submatrix of* $\mathbf{Z}$ *composed of the first* 6 *columns. With* $\mathbf{Z}_1$ *and* $\mathbf{K}^*$ *we can compute*

$$\hat{\boldsymbol{\mu}}_c = \mathbf{K}^*(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{y}$$

Finally, $v_{E_c} = v_E + 6$ *since* $\mathbf{C}$ *is of rank* 6*.*

The SAS code is given as following:

```
/*for part (b)*/

/*compute K, A_new(A in the textbook), Z, Z1, K_star, mu_c, SSE_c, vE_c*/
j = {1 1 1 1 1 1 1 1 1 1 1 1 };
K = t(t(j)||t(A)||t(B));
A_new = t(t(K)||t(C));
print A_new;

rankA_new=round(trace(ginv(A_new)*A_new));
print rankA_new;

Z = W*inv(A_new);
Z1 = Z[, 1:6];
K_star = t(K)*inv(K*t(K));
est_mu_c = K_star*inv(t(Z1)*Z1)*t(Z1)*y;

SSE_c = t(y - W*est_mu_c)*(y - W*est_mu_c);
vE_c = v_E + 6;

/*test main effect of cement(A) and aggregate(B) for constrained model*/
/*output F statistic and p values*/

F_Ac = t(A*est_mu_c)*inv(A*K_star*inv(t(Z1)*Z1)*t(K_star)*t(A))*A*est_mu_c/v_A/SSE_c*vE_c;
p_Ac = 1 - CDF('F', F_Ac, v_A, vE_c);

F_Bc = t(B*est_mu_c)*inv(B*K_star*inv(t(Z1)*Z1)*t(K_star)*t(B))*B*est_mu_c/v_B/SSE_c*vE_c;
p_Bc = 1 - CDF('F', F_Bc, v_B, vE_c);

print SSE_c vE_c F_Ac p_Ac F_Bc p_Bc;
```

In the output we show SSE_c*, two F statistics for testing the main effects, and their corresponding p values:*

SSE_c	vE_c	F_Ac	p_Ac	F_Bc	p_Bc
182.24855	27	1.8911275	0.170355	13.468117	0.0000146

The F statistic for testing the main effect of cement gives value 1.89 *with p value* 0.17*, which is not significant and we conclude that under the constrained additive model, there is* **No** *main effect from cement. The F statistic for testing the main effect of aggregate gives value* 13.468 *with p value* 0.0000146*, which is significant, so we conclude that under the constrained model there is a main effect coming from aggregate.*

For part (c):

Suppose the combination of type 3 *cement and type B aggregate is not available, we are missing information for* μ_{32}*. To help us analyze and find the right contrasts for our hypothesis, we look at the folloing table:*

CEMENT/AGGREGATE	A	B	C	D
1	μ_{11}	μ_{12}	μ_{13}	μ_{14}
2	μ_{21}	μ_{22}	μ_{23}	μ_{24}
3	μ_{31}		μ_{33}	μ_{34}

Our parameter is now

$$\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}, \mu_{31}, \mu_{33}, \mu_{34})'$$

Notice the length is 11 *now instead of* 12*.*

We make hypothesis for the main effect from cement:

$$H_0 : \begin{cases} 2(\mu_{11} + \mu_{13} + \mu_{14}) = (\mu_{21} + \mu_{23} + \mu_{24}) + (\mu_{31} + \mu_{33} + \mu_{34}) \\ \mu_{21} + \mu_{23} + \mu_{24} = \mu_{31} + \mu_{33} + \mu_{34} \end{cases}$$

the contrast matrix is:

$$A = \begin{pmatrix} 2 & 0 & 2 & 2 & -1 & 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & -1 & -1 & -1 \end{pmatrix}_{2\times 11}$$

we also make hypothesis for the main effect from aggregate:

$$H_0 : \begin{cases} 3(\mu_{11} + \mu_{21}) = (\mu_{12} + \mu_{22}) + (\mu_{13} + \mu_{23}) + (\mu_{14} + \mu_{24}) \\ 2(\mu_{12} + \mu_{22}) = (\mu_{13} + \mu_{23}) + (\mu_{14} + \mu_{24}) \\ \mu_{13} + \mu_{23} = \mu_{14} + \mu_{24} \end{cases}$$

the contrast matrix is:

$$B = \begin{pmatrix} 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix}_{3\times 11}$$

Finally we make hypothesis for interaction as:

$$H_0 : \begin{cases} \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} \\ \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23} \\ \mu_{12} - \mu_{22} = \mu_{14} - \mu_{24} \\ \mu_{11} - \mu_{31} = \mu_{13} - \mu_{33} \\ \mu_{11} - \mu_{31} = \mu_{14} - \mu_{34} \end{cases}$$

The corresponding contrast matrix is:

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}_{5\times 11}$$

We use proc glm as in the bakery study example:

```
proc glm data=q2;
    class cement aggregate;
    model strength = cement*aggregate/noint;
    contrast 'H_0 for cement'
        cement*aggregate 2 0 2 2 -1 0 -1 -1 -1 -1 -1,
        cement*aggregate 0 0 0 0 1 0 1 1 -1 -1 -1;
    contrast 'H_0 for aggregate'
        cement*aggregate 3 -1 -1 -1 3 -1 -1 -1 0 0 0,
        cement*aggregate 0 2 -1 -1 0 2 -1 -1 0 0 0,
        cement*aggregate 0 0 1 -1 0 0 1 -1 0 0 0;
    contrast 'H_0 for cement-aggregate'
        cement*aggregate 1 -1 0 0 -1 1 0 0 0 0 0,
        cement*aggregate 0 1 -1 0 0 -1 1 0 0 0 0,
        cement*aggregate 0 1 0 -1 0 -1 0 1 0 0 0,
        cement*aggregate 1 0 -1 0 0 0 0 0 -1 1 0,
        cement*aggregate 1 0 0 -1 0 0 0 0 -1 0 1;
run;
```

The output for the contrast hypothesis test is:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
H_0 for cement	2	46.2452381	23.1226190	4.97	0.0171226
H_0 for aggregate	3	139.1229167	46.3743056	9.96	0.0002748
H_0 for cement-aggregate	5	11.7430184	2.3486037	0.50	0.7694696

From the output we see the main effect from cement and aggregate are both significant, however no enough evidence support the existence of interaction (fail to reject with high p-value of 0.77*).*

We also check the type IV sum of square with default glm output:

```
proc glm data=q2;
    class cement aggregate;
    model strength = cement aggregate cement*aggregate/ss4;
run;
```

Source	DF		Type IV SS	Mean Square	F Value	Pr > F
cement	2	*	46.2452381	23.1226190	4.97	0.0171226
aggregate	3	*	257.5515128	85.8505043	18.44	0.0000043
cement*aggregate	5		11.7430184	2.3486037	0.50	0.7694696

comparing with the sum of square from contrasts, we got the same sum of square for cement main effect and interaction, but different sum of square for the aggregate main effect.

For part (d):

With the same assumption as (c), to test for interaction:

First consider using Full-Reduced model approach:

The full model is the unconstrained model:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon} = (\mathbf{W}_1, \mathbf{O}) \begin{pmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_e \end{pmatrix} + \boldsymbol{\epsilon}$$

with

$$SSE_u = \mathbf{y}'[\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}']\mathbf{y}$$

of degree of freedom $(n - ab + m)$.

On the other hand there is reduced model that is additive subject to a constraint:

$$\mathbf{y} = \mathbf{W}\mathbf{A}_{repar}^{-1}\mathbf{A}_{repar}\boldsymbol{\mu} + \boldsymbol{\epsilon} \text{ subject to } \mathbf{G}\boldsymbol{\mu} = 0$$

Here we use the notation $\mathbf{A}_{repar}$ *to indicate this is the matrix for reparametrization, and later we will use* $\mathbf{A}$, $\mathbf{B}$ *and* $\mathbf{C}$ *to denote the contrast matrix for main effects and interaction.*

We have:

$$\mathbf{A}_{repar} = \begin{pmatrix} \mathbf{K} \\ \mathbf{G} \end{pmatrix}$$

In our case, we have:

$$\mathbf{K} = \begin{pmatrix} \mathbf{j}' \\ \mathbf{A} \\ \mathbf{B} \end{pmatrix}$$
$$\mathbf{G} = \mathbf{C}$$

where $\mathbf{A}$ *is the contrast matrix for main effect from cement,* $\mathbf{B}$ *is the contrast matrix for the main effect from aggregate, and* $\mathbf{C}$ *is the contrast matrix for interaction.*

We rearrange $\mathbf{W}$ *so the empty cell appears in the last column, and we do the same rearrangement for response* $\mathbf{y}$.

So the parameters are reordered as:

$$\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}, \mu_{31}, \mu_{33}, \mu_{34}, \mu_{32})$$

Our hypothesis stays the same for each part so does the contrast matrix. (we are doing full-reduced model approach here so our contrast would be proposed as if there is no empty sell). We did make a re-arrangement of the parameter so the contrast matrices would be slighty different than before. For main effect of cement, we still have:

$$H_0 : \begin{cases} 2(\mu_{11} + \mu_{12} + \mu_{13} + \mu_{14}) = (\mu_{21} + \mu_{22} + \mu_{23} + \mu_{24}) + (\mu_{31} + \mu_{32} + \mu_{33} + \mu_{34}) \\ \mu_{21} + \mu_{22} + \mu_{23} + \mu_{24} = \mu_{31} + \mu_{32} + \mu_{33} + \mu_{34} \end{cases}$$

But the contrast matrix is now:

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

For main effect of aggregate, we have:

$$H_0 : \begin{cases} 3(\mu_{11} + \mu_{21} + \mu_{31}) = (\mu_{12} + \mu_{22} + \mu_{32}) + (\mu_{13} + \mu_{23} + \mu_{33}) + (\mu_{14} + \mu_{24} + \mu_{34}) \\ 2(\mu_{12} + \mu_{22} + \mu_{32}) = (\mu_{13} + \mu_{23} + \mu_{33}) + (\mu_{14} + \mu_{24} + \mu_{34}) \\ \mu_{13} + \mu_{23} + \mu_{33} = \mu_{14} + \mu_{24} + \mu_{34} \end{cases}$$

the contrast matrix is now:

$$\mathbf{B} = \begin{pmatrix} 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 & 3 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 & 0 & 2 & -1 & -1 & 0 & -1 & -1 & 2 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \end{pmatrix}$$

and finally for the interaction, we use hadamad product to get the contrst matrix as:

$$\mathbf{C} = \begin{pmatrix} 6 & -2 & -2 & -2 & -3 & 1 & 1 & 1 & -3 & 1 & 1 & 1 \\ 0 & 4 & -2 & -2 & 0 & -2 & 1 & 1 & 0 & 1 & 1 & -2 \\ 0 & 0 & 2 & -2 & 0 & 0 & -1 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 3 & -1 & -1 & -1 & -3 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 & 1 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \end{pmatrix}$$

We can then get $\mathbf{K}$ *and compute*

$$\mathbf{K}^* = \mathbf{K}'(\mathbf{K}\mathbf{K}')^{-1}$$

and then we have

$$\mathbf{Z}_1 = \mathbf{W}\mathbf{K}^*$$

and is able to get

$$SSE_a = \mathbf{y}'[\mathbf{I} - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1']\mathbf{y}$$

so we can get the testing statsitic for interaction as

$$F = \frac{(SSE_a - SSE_u)/[3 \times 2 - 1]}{SSE_u/(32 - 12 + 1)} \sim F(5, 21) \text{ under the null}$$

The SAS code are as following:

```
/*contrasts matrix for testing main effect of cement(factor A),
  main effect of aggregate(factor B), and interaction effect(A*B)*/
A = {2 2 2 2 -1 -1 -1 -1 -1 -1 -1 -1, 0 0 0 0 1 1 1 1 -1 -1 -1 -1};
print A;

B = {3 -1 -1 -1 3 -1 -1 -1 3 -1 -1 -1, 0 2 -1 -1 0 2 -1 -1 0 -1 -1 2,
     0 0 1 -1 0 0 1 -1 0 1 -1 0};
print B;

C = {6 -2 -2 -2 -3 1 1 1 -3 1 1 1 , 0 4 -2 -2 0 -2 1 1 0 1 1 -2,
     0 0 2 -2 0 0 -1 1 0 -1 1 0, 0 0 0 0 3 -1 -1 -1 -3 1 1 1 ,
     0 0 0 0 0 2 -1 -1 0 1 1 -2, 0 0 0 0 0 0 1 -1 0 -1 1 0};
print C;

m = 1;
N = 33;
factor_A = 3;
factor_B = 4;
I = i(N);
H = W*ginv(t(W)*W)*t(W);
SSE_u = t(y)*(I - H)*y;
print SSE_u;

/*compute K, A_new(A in the textbook), Z, Z1, K_star, mu_c, SSE_c, vE_c*/
j = {1 1 1 1 1 1 1 1 1 1 1 1 };
K = t(t(j)||t(A)||t(B));
A_repar = t(t(K)||t(C));
print A_repar;

rankA_repar=round(trace(ginv(A_repar)*A_repar));
print rankA_repar;

K_star = t(K)*inv(K*t(K));
Z1 = W*K_star;
SSE_a = t(y)*(I - Z1*inv(t(Z1)*Z1)*t(Z1))*y;
F_int = (SSE_a - SSE_u)/(6 - 1)/SSE_u*(33-12);
p_int = 1 - CDF('F', F_int, 5, 21);
print  F_int p_int;
```

I got the following output:

F_int	p_int
0.055935	0.9976966

The p value is highly insignificant, and we fail to reject the null and conclude that there is not enough evidence to support the existence of interaction.

Now consider the side condition approach:

$$\gamma^*_{32} = \mu_{32} - \bar{\mu}_{3\cdot} - \bar{\mu}_{\cdot 2} + \bar{\mu}_{\cdot\cdot} = 0$$

We then we have:

$$\implies \mu_{32} - \frac{1}{4}(\mu_{31} + \mu_{32} + \mu_{33} + \mu_{34}) - \frac{1}{3}(\mu_{12} + \mu_{22} + \mu_{32}) + \frac{1}{12}\sum_{i=1}^{3}\sum_{j=1}^{4}\mu_{ij} = 0$$

$$\implies 12\mu_{32} - 3\mu_{3\cdot} - 4\mu_{\cdot 2} + \mu_{\cdot\cdot} = 0$$

this simplies to

$$(1, -3, 1, 1, 1, -3, 1, 1, -2, -2, -2, 6)\boldsymbol{\mu} = 0$$

with

$$\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}, \mu_{31}, \mu_{33}, \mu_{34}, \mu_{32})$$

(don't forget that the last column of $\mathbf{W}$ *is* 0 *and* μ_{32} *is the last parameter now).*

So

$$\mathbf{T} = (1, -3, 1, 1, 1, -3, 1, 1, -2, -2, -2, 6)$$

is our $m \times ab$ *(*$m = 1$*,* $a = 3$*,* $b = 4$*) matrix with the only row that corresponds to* $\mu_{32} - \bar{\mu}_{3\cdot} - \bar{\mu}_{\cdot 2} + \bar{\mu}_{\cdot\cdot}$ *for* 1 *empty cell. We can get*

$$\hat{\boldsymbol{\mu}} = (\mathbf{W}'\mathbf{W} + \mathbf{T}'\mathbf{T})^{-1}\mathbf{W}'\mathbf{y}$$

and the the F *statistic for testing interaction would be*

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\mu}})'\{\mathbf{C}[cov(\hat{\boldsymbol{\mu}})/\sigma^2]\mathbf{C}'\}^{-}(\mathbf{C}\hat{\boldsymbol{\mu}})/[2 \times 3 - 1]}{SSE/(32 - 12 + 1)} \sim F(5, 21) \text{ under null}$$

The following is the SAS code:

```
/*side condition approach*/
T = {1 -3 1 1 1 -3 1 1 -2 -2 -2 6};
/*equation 15.52*/
est_mu = inv(t(W)*W + t(T)*T)*t(W)*y;

SSE = t(y - W*est_mu)*(y - W*est_mu);
/*equation 15.53*/
cov_mu = inv(t(W)*W + t(T)*T)*t(W)*W*inv(t(W)*W + t(T)*T);
/*equation 15.54*/
F_int = t(C*est_mu)*ginv(C*cov_mu*t(C))*C*est_mu/(3*2 - 1)/SSE*(32 - 12 + 1);
p_int = 1 - CDF('F', F_int, 5, 21);
print SSE F_int p_int;
```

SSE	F_int	p_int
881.75	0.055935	0.9976966

As we can see it gives the same p value as in the full-reduced models approach, and we fail to reject the null hypothesis and conclude that there is no enough evidence showing the interaction between aggregate and cement.

We also notice that the SSE in the side condition approach is the same as the SSE_u *in the full-reduced model approach.*

Thus completed part (d).

For part (e)*:*

Consider the data with only one factor aggregator, the cell mean coding gives the model as

$$y_{ij} = \mu_i + \epsilon_{ij}$$

here $i = 1, 2, 3, 4$ *and* $j = 1, 2, \ldots, n_i$.

We consider this is a linear model with matrix form

$$\mathbf{y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

To test on main effect, we have

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

and the test statistic is

$$F = \frac{SSB/(k-1)}{SSE/(n-k)} = \frac{(\hat{\boldsymbol{\mu}}'\mathbf{W}\mathbf{y} - N\bar{y}_{\cdot\cdot}^2)/(4-1)}{(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\mu}}'\mathbf{W}\mathbf{y})/(33-4)} \sim F(3, 29) \text{ under } H_0$$

We have the following code and output:

```
/*5(e)*/
proc iml;
    /*cell means approach*/
    W1 ={1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
    W2 = {0,0,0,0,0,0,0,0,1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
    W3 = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
            1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
    W4 = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
           0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1};
    W = W1||W2||W3||W4;
    y = {21, 27, 19, 25, 23, 24, 20, 24, 19, 19, 22, 23, 20, 24, 18, 28,
         19, 16, 19, 18, 14, 16, 12, 23, 24, 23, 28, 27, 25, 23, 25, 22, 22};
    N = nrow(y);
    est_mu = inv(t(W)*W)*t(W)*y;
    SSB = t(est_mu)*t(W)*y - sum(y)##2/N;
    SSE = t(y)*y - t(est_mu)*t(W)*y;
    F = SSB/(4 - 1)/SSE*(N - 4);
    p = 1 - CDF('F', F, 3, N - 4);
    print SSE F p;
```

SSE	F	p
207.77857	12.946762	0.0000152

The p-value here is significant so we reject the null and conclude there is a main effect from the aggregate.

For reference cell coding, the model is:

$$\begin{aligned} y_{1j} &= \mu + \epsilon_{1j} \\ y_{2j} &= \mu + \alpha_2 + \epsilon_{2j} \\ &\vdots \\ y_{kj} &= \mu + \alpha_k + \epsilon_{kj} \end{aligned}$$

We will have a different design matrix than the cell means coding. The matrix form of the model now is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The SAS code is the following:

```
/*the reference cell coding*/
X1 = {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};
X2 = {0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0,
      0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
X3 = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,
    1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
X4 = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      0, 1, 1, 1, 1, 1,1,1,1,1,1};
X = X1||X2||X3||X4;

est_mu = inv(t(X)*X)*t(X)*y;
SSB = t(est_mu)*t(X)*y - sum(y)##2/N;
SSE = t(y)*y - t(est_mu)*t(X)*y;
F = SSB/(4 - 1)/SSE*(N - 4);
p = 1 - CDF('F', F, 3, N - 4);
print SSE F p;
```

The output is the same as the cell means model:

SSE	F	p
207.77857	12.946762	0.0000152

and we conclude there is a main effect from aggregate.

Now for effect cell coding:

$$\begin{aligned} y_{1j} &= \mu - \alpha_2 - \ldots - \alpha_k + \epsilon_{1j} \\ y_{2j} &= \mu + \alpha_2 + \epsilon_{2j} \\ &\vdots \\ y_{kj} &= \mu + \alpha_k + \epsilon_{kj} \end{aligned}$$

The SAS code is the following:

```
/*effect coding*/
X1 = {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};
X2 = {-1, -1, -1, -1, -1, -1, -1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0,
      0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
X3 = {-1, -1, -1, -1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,
    1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
X4 = {-1, -1, -1, -1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      0, 1, 1, 1, 1, 1,1,1,1,1,1,1};
X = X1||X2||X3||X4;

est_mu = inv(t(X)*X)*t(X)*y;
SSB = t(est_mu)*t(X)*y - sum(y)##2/N;
SSE = t(y)*y - t(est_mu)*t(X)*y;
F = SSB/(4 - 1)/SSE*(N - 4);
p = 1 - CDF('F', F, 3, N - 4);
print SSE F p;
```

The output is the same as the cell means coding and reference cell coding:

SSE	F	p
207.77857	12.946762	0.0000152

These results are not surprising, since all these different codings are essentially reparametrization of each other and they give equivalent models, so SSR and SSE are invariant under these different models and we should have the same F statistic for making the inference on main effect.

Question 3.

Solution 3. *For part (i):*

We have the following data:

TABLE 7.4 Chemical Reaction Data

y_1	y_2	x_1	x_2	x_3
41.5	45.9	162	23	3
33.8	53.3	162	23	8
27.7	57.5	162	30	5
21.7	58.8	162	30	8
19.9	60.6	172	25	5
15.0	58.0	172	25	8
12.2	58.6	172	30	5
4.3	52.4	172	30	8
19.3	56.9	167	27.5	6.5
6.4	55.4	177	27.5	6.5
37.6	46.9	157	27.5	6.5
18.0	57.3	167	32.5	6.5
26.3	55.0	167	22.5	6.5
9.9	58.9	167	27.5	9.5
25.0	50.3	167	27.5	3.5
14.1	61.1	177	20	6.5
15.2	62.9	177	20	6.5
15.9	60.0	160	34	7.5
19.6	60.6	160	34	7.5

Our linear regression model for y_1 is:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

The following SAS code create the data set and we print out design matrix $\mathbf{X}$ *and response vector* y_1*:*

```
/*Question 3*/
proc iml;
/*  design matrix */
    X = {1 162 23 3, 1 162 23 8, 1 162 30 5, 1 162 30 8,
         1 172 25 5, 1 172 25 8, 1 172 30 5, 1 172 30 8,
         1 167 27.5 6.5, 1 177 27.5 6.5, 1 157 27.5 6.5, 1 167 32.5 6.5,
         1 167 22.5 6.5, 1 167 27.5 9.5, 1 167 27.5 3.5, 1 177 20 6.5,
         1 177 20 6.5, 1 160 34 7.5, 1 160 34 7.5};
/*response y1*/
y1 = {41.5, 33.8, 27.7, 21.7, 19.9, 15.0, 12.2, 4.3, 19.3, 6.4, 37.6,
      18.0, 26.3, 9.9, 25.0, 14.1, 15.2, 15.9, 19.6};
/*create labels for matrices*/
cY = {"y1"};
cX = {"1" "x1" "X2" "X3" };
mattrib y1 colname=cY X colname=cX;

print X y1;
```

X				y1
1	x1	X2	X3	y1
1	162	23	3	41.5
1	162	23	8	33.8
1	162	30	5	27.7
1	162	30	8	21.7
1	172	25	5	19.9
1	172	25	8	15
1	172	30	5	12.2
1	172	30	8	4.3
1	167	27.5	6.5	19.3
1	177	27.5	6.5	6.4
1	157	27.5	6.5	37.6
1	167	32.5	6.5	18
1	167	22.5	6.5	26.3
1	167	27.5	9.5	9.9
1	167	27.5	3.5	25
1	177	20	6.5	14.1
1	177	20	6.5	15.2
1	160	34	7.5	15.9
1	160	34	7.5	19.6

To answer question 7.54*:*

For part (a)*:*

we first check that the design matrix is full rank (rank 4):

```
rankX=round(trace(ginv(X)*X));

print rankX;
```

```
rankX

     4
```

We then have:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}_1$$
$$s^2 = \frac{SSE}{n-k-1} = \frac{SSE}{19-3-1} = \frac{SSE}{4} = \mathbf{y}_1'\Big(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big)\mathbf{y}_1/4$$

The following SAS code compute $\hat{\beta}$ and s^2:

```
/*find dimension of design matrix X*/
dX =dimension(X);
print dX;

n = 19;
k = 3;
/*compute I,  H, SSE, hat_beta, and s^2*/
I = i(n);
H = X*inv(t(X)*X)*t(X);
SSE = t(y1)*(I - H)*y1;
s_2 = SSE/(n - k - 1);
hat_beta = inv(t(X)*X)*t(X)*y1;

print hat_beta s_2;
```

```
dX

19          4


hat_beta        s_2

332.11098 5.3449026
-1.545961
-1.424559
-2.237366
```

So we have:

$$\hat{\boldsymbol{\beta}} = (332.11098, -1.545961, -1.424559, -2.237366)'$$
$$s^2 = 5.3449026$$

For part (b):

Since we have:

$$cov(\boldsymbol{\beta}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Thus an estimate would be:

$$cov(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

The following SAS code provide this computation:

```
/*beta covariance matrix estimate*/
cov_beta_hat = s_2*inv(t(X)*X);

print cov_beta_hat;
```

cov_beta_hat

349.42569	-1.811104	-1.670412	-0.109109
-1.811104	0.0098079	0.0068077	-0.002303
-1.670412	0.0068077	0.0218005	-0.009425
-0.109109	-0.002303	-0.009425	0.1154798

the output above is $cov(\hat{\boldsymbol{\beta}})$.

For part (c)*:*

We have:

$$R^2 = \frac{SSR}{SSE} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2}{SST}$$
$$R_a^2 = \frac{(n-1)R^2 - k}{n-k-1}$$

The following SAS code provide this computation:

```
/*compute SSR*/
SSR = t(hat_beta)*t(X)*y1 - sum(y1)##2/n;
print SSR;

/*R^2 and R^2_a*/
R_2 = SSR/(SSE + SSR);
R_2_adj = ((n - 1)*R_2 - k)/(n - k - 1);

print R_2 R_2_adj;
```

SSR

1707.158

R_2	R_2_adj
0.9551434	0.9461721

So we got:

$$R^2 = 0.9551434$$
$$R_{adj}^2 = 0.9461721$$

For part (d) and part (e):

We consider second order model:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \epsilon$$

The following SAS create design matrix for this model and compute $\hat{\beta}, s^2, R^2$ *and* R_a^2.

```
/*create second order terms*/

/*square terms*/
x1_2 = X[, 2]##2;
x2_2 = X[, 3]##2;
x3_2 = X[, 4]##2;
print x1_2 x2_2 x3_2;
/*mixed terms*/
x12 = X[, 2]#X[, 3];
x13 = X[, 2]#X[, 4];
x23=  X[, 3]#X[, 4];

print x1_2 x2_2 x3_2 x12 x13 x23;

X_new = X||x1_2||x2_2||x3_2||x12||x13||x23;

print X_new;

rankx_new = round(trace(ginv(X_new)*X_new));

print rankx_new;

k_new = 9;
H_new = X_new*inv(t(X_new)*X_new)*t(X_new);
SSE_new = t(y1)*(I - H_new)*y1;
s_2_new = SSE_new/(n - k_new - 1);
hat_beta_new = inv(t(X_new)*X_new)*t(X_new)*y1;
SSR_new = t(hat_beta_new)*t(X_new)*y1 - sum(y1)##2/n;
R_2_new = SSR_new/(SSE_new + SSR_new);
R_2_adj_new = ((n - 1)*R_2_new - k_new)/(n - k_new - 1);

print SSE_new s_2_new hat_beta_new SSR_new R_2_new R_2_adj_new;
```

we have also checked that the design matrix for the second order model is also with full rank 10.

```
rankx_new

10


  SSE_new   s_2_new hat_beta_new   SSR_new   R_2_new R_2_adj_new

46.208256 5.1342507    964.92906 1741.1233 0.9741468   0.9482936
                       -7.442128
                        -11.5077
                       -2.140127
                       0.0124571
                       0.0332188
                       -0.294014
                        0.053507
                         0.03804
                       -0.101633
```

So we have found that:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 964.92906 \\ -7.442128 \\ -11.5077 \\ -2.140127 \\ 0.0124571 \\ 0.0332188 \\ -0.294014 \\ 0.053507 \\ 0.03804 \\ -0.101633 \end{bmatrix}$$

and

$$s^2 = MSE = 5.1342507$$

we have also foud

$$R^2 = 0.9741468$$
$$R^2_{adj} = 0.9482936$$

Thus finished solving part (a)(ie. Exercise 7.54).

For part (b):

If we assume the second order model is the correct full model, then when the second order terms are ignored, the underfitted first order reduced model will cause the bias on the estimate of coefficients $\hat{\boldsymbol{\beta}}$, the predicted values $\hat{\mathbf{y}}$ and the estimate of variance s^2. It has been shown also that s^2 for the reduced model will be biased upward.

For part (c):

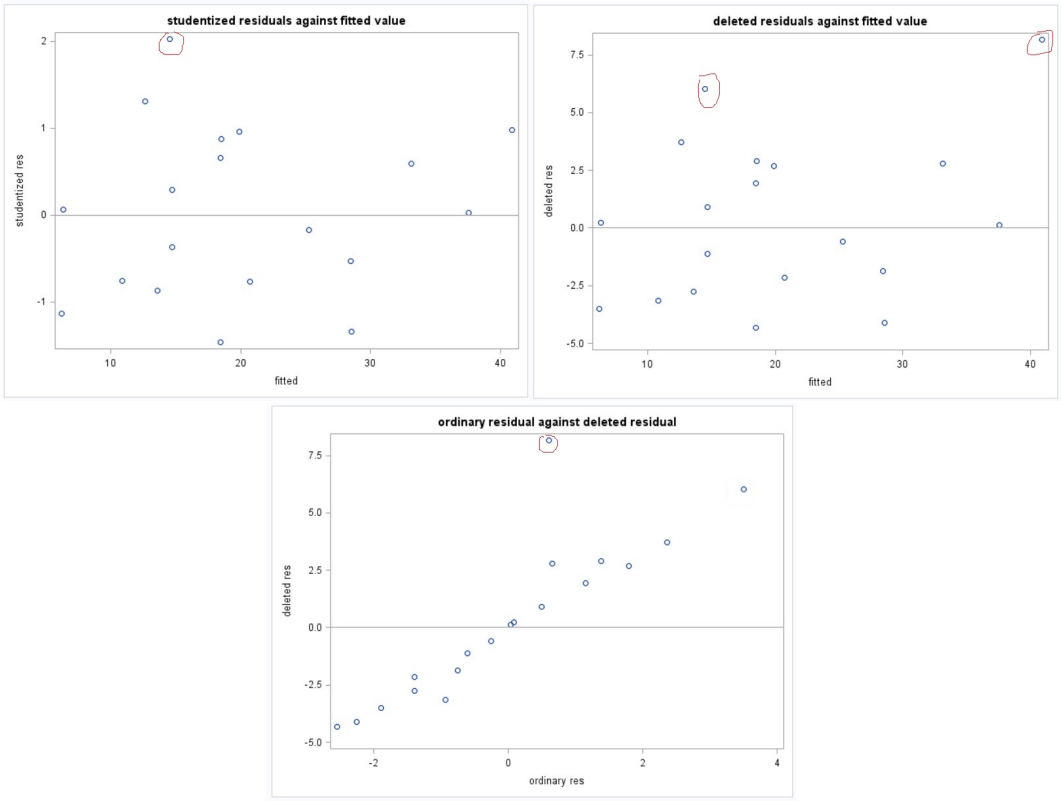
We use the second order model to conduct the residual analysis here.

We compute regular residuals $\hat{e}_i$, studentized residuals r_i, studentized external residuals t_i and deleted residuals $\hat{e}_{(i)}$ for later use. The code is as following:

```
/*residual, outlier and influential observation analysis*/

/*compute fitted values*/
y_hat = H_new*y1;

/*compute the residuals*/

/*regular residuals*/
residual = (I - H_new)*y1;

/*studentized residuals*/
residual_stu = (inv(I - diag(H_new)))##(0.5)*residual/s_new;

/*deleted residuals*/
residual_del = inv(I - diag(H_new))*residual;
print residual_del;

/*external studentized residuals*/

/*compute SSEs without a single observation*/
/*refer to equation 9.32 from textbook*/
SSE_ext = SSE_new*j(n, 1, 1)  - inv(I - diag(H_new))*(residual##2);
/*compute s*/
s_ext = (SSE_ext/(n - k_new - 2))##(0.5);
/*compute esternal studentized residuals*/
residual_ext = (inv(I - diag(H_new))##(0.5))*(residual # (s_ext##(-1)));
```

To locate outliers, we make a scatter plot of studentized residuals against fitted value, and deleted residuals against fitted value, as well as deleted residual against ordinary residuals:

```
/*scatter plot of residual against estimated mean*/
ods html;
ods listing close;
title "studentized residuals against fitted value";
run Scatter(y_hat, residual_stu)
    /*add reference line*/
    other = "refline 0/axis = y"
    label={"fitted" "studentized res"}
    ;
title "deleted residuals against fitted value";
run Scatter(y_hat, residual_del)
    /*add refernece line*/
    other = "refline 0/axis = y"
    label={"fitted" "deleted res"}
    ;
title "ordinary residual against deleted residual";
run Scatter(residual, residual_del)
    /*add refernece line*/
    other = "refline 0/axis = y"
    label={"ordinary res" "deleted res"}
    ;
ods html close;

ods listing;
```



We can see that both residual plots against fitted values display similar patterns, although some potential outliers behave differently. We marked the outlier candidates with red circles. The deleted residual vs regular residual shows a straightline pattern, with one point deviated far away, which suggested a potential influential observation. We will do further analysis in the following.

To find influential observations, we compute the following:

1. *studentized (r_i), studentized external (t_i) and deleted ($\hat{e}_{(i)}$) residuals, the code of which is shown above.*
2. *PRESS: prediction sum of square*
3. *Cooks distance*

Our code is as following: also, as suggested by Hoaglin and Welsch(1978), the high leverage point is

$$\frac{2k+1}{n} = \frac{2 \times (9+1)}{19} = 1.052632$$

We print out a table similar to Table 9.1 *as the example from the book:*

table

	obs	response	fitted	residual	leverage(h_ii)	r_i	t_i	Cook
ROW1	1	41.5	40.899784	0.6002156	0.9264479	0.9767224	0.9739256	1.2016224
ROW2	2	33.8	33.152964	0.6470363	0.7680842	0.5929594	0.5702984	0.1164471
ROW3	3	27.7	28.455163	-0.755163	0.5982579	-0.525809	-0.503532	0.0411716
ROW4	4	21.7	19.908692	1.7913079	0.3295035	0.9654581	0.9613874	0.0458068
ROW5	5	19.9	18.52291	1.3770895	0.523456	0.880384	0.8682628	0.0851376
ROW6	6	15	12.642135	2.3578646	0.36603	1.3069109	1.3688886	0.0986143
ROW7	7	12.2	13.594751	-1.394751	0.4926836	-0.864209	-0.850852	0.0725314
ROW8	8	4.3	6.1894794	-1.889479	0.4608305	-1.135641	-1.156777	0.1102295
ROW9	9	19.3	20.693549	-1.393549	0.3519984	-0.764004	-0.744869	0.0317071
ROW10	10	6.4	6.3117728	0.0882272	0.59948	0.061525	0.0580185	0.0005666
ROW11	11	37.6	37.566748	0.0332521	0.7144815	0.0274639	0.0258943	0.0001887
ROW12	12	18	14.495935	3.5040647	0.4180725	2.0272118	2.5928141	0.2952439
ROW13	13	26.3	28.552101	-2.252101	0.4534991	-1.344479	-1.417956	0.1500008
ROW14	14	9.9	10.8338	-0.9338	0.7038411	-0.757274	-0.737859	0.1362875
ROW15	15	25	25.261046	-0.261046	0.5589153	-0.173467	-0.163821	0.0038129
ROW16	16	14.1	14.709428	-0.609428	0.4585575	-0.365517	-0.3472	0.0113151
ROW17	17	15.2	14.709428	0.4905716	0.4585575	0.2942305	0.278747	0.0073319
ROW18	18	15.9	18.450156	-2.550156	0.4086516	-1.463546	-1.580708	0.1480208
ROW19	19	19.6	18.450156	1.1498442	0.4086516	0.6599009	0.6377816	0.0300931

high_leverage	PRESS	SSE_new
1.0526316	219.75472	46.208256

The observation number here are manually added just to make it easier to point out which observation we are talking about. But it is not the real observation index since it is not given by the data.

There is no observation that has a leverage higher than the suggested high leverage. Observation 1, 4, 11, *and* 14 *have relatively higher leverage.From the leverage aspect, these points can be potentially influential to the model.*

If we look at Cook's distance, observation 1 *has a much larger value than the rest, implying that it might be most influential than the other observations. Observation* 12 *also has a relatively larger cooks' distance than the others.*

In terms of residuals, observation 12 *is the only one with studentized and studentized external residuals with magnitude larger than* 2.

So to summarize, we conclude that the two most influential observations are observation 1 *and* 12.

Our PRESS value is 219.75, *if we want to compare between models with different observations, the one with smaller PRESS value will be more preferrable.*

For part (iv)*:*

for any particular coefficient β_j, *we know that*

$$P\Big[-t_{\alpha/2,n-k-1} \leq \frac{\hat{\beta}_j - \beta_j}{s\sqrt{g_{jj}}} \leq t_{\alpha/2,n-k-1}\Big] = 1-\alpha.$$

which gives the $100(1-\alpha)\%$ *confidence interval for* β_j *as:*

$$\hat{\beta}_j \pm t_{\alpha/2,n-k-1} s\sqrt{g_{jj}}$$

Here g_{jj} *is the* j*th diagonal element of* $(\mathbf{X}'\mathbf{X})^{-1}$.

The following SAS code compute 95% *confidence intervals for* β_1 *to* β_9

```
/*individual confidence intervals for betas*/
g = inv(t(X_new)*X_new);
print g;
g11 = g[2, 2];
g22 = g[3, 3];
g33 = g[4, 4];
g44 = g[5, 5];
g55 = g[6, 6];
g66 = g[7, 7];
g77 = g[8, 8];
g88 = g[9, 9];
g99 = g[10, 10];
lower_1 = hat_beta_new[2] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g11);
upper_1 = hat_beta_new[2] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g11);
lower_2 = hat_beta_new[3] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g22);
upper_2 = hat_beta_new[3] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g22);
lower_3 = hat_beta_new[4] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g33);
upper_3 = hat_beta_new[4] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g33);
lower_4 = hat_beta_new[5] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g44);
upper_4 = hat_beta_new[5] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g44);
lower_5 = hat_beta_new[6] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g55);
upper_5 = hat_beta_new[6] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g55);
lower_6 = hat_beta_new[7] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g66);
upper_6 = hat_beta_new[7] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g66);
lower_7 = hat_beta_new[8] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g77);
upper_7 = hat_beta_new[8] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g77);
lower_8 = hat_beta_new[9] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g88);
upper_8 = hat_beta_new[9] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g88);
lower_9 = hat_beta_new[10] - tinv(0.975, n - k_new - 1)*s_new*sqrt(g99);
upper_9 = hat_beta_new[10] + tinv(0.975, n - k_new - 1)*s_new*sqrt(g99);
print lower_1 upper_1 lower_2 upper_2 lower_3 upper_3 lower_4 upper_4 lower_5 upper_5;
print lower_6 upper_6 lower_7 upper_7 lower_8 upper_8 lower_9 upper_9;
```

The output is:

```
          lower_1   upper_1   lower_2   upper_2   lower_3
             Col1      Col2      Col3      Col4      Col5

ROW1 -23.46535 8.5810898 -29.79275 6.7773418 -36.85452

          upper_3   lower_4   upper_4   lower_5   upper_5
             Col6      Col7      Col8      Col9     Col10

ROW1 32.574268  -0.03268 0.0575947 -0.081384  0.147822


 lower_6   upper_6   lower_7   upper_7   lower_8   upper_8   lower_9   upper_9

-0.802502 0.2144737 -0.027781 0.1347949 -0.186424 0.2625041 -0.443609 0.2403428
```

reorganize the answer, we got the 95% *confidence intervals for each* β_j *as:*

$\beta_1 : (-23, 46535, 8.5810898)$ $\beta_2 : (-29.79275, 6.7773418)$ $\beta_3 : (-36.85452, 32.574268)$

$\beta_4 : (-0.03268, 0.0575947)$ $\beta_5 : (-0.081384, 0.147822)$ $\beta_6 : (-0.802502, 0.2144737)$

$\beta_7 : (-0.027781, 0.1347949)$ $\beta_8 : (-0.186424, 0.2625041)$ $\beta_9 : (-0.443609, 0.2403428)$

Now for confidence interval of σ^2*, we have:*

$$P\left[\chi^2_{1-\alpha/2,n-k-1} \leq \frac{(n-k-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-k-1}\right] = 1-\alpha$$

which gives the $100(1-\alpha)\%$ *confidence interval for* σ^2 *as:*

$$\frac{(n-k-1)s^2}{\chi^2_{\alpha/2,n-k-1}} \leq \sigma^2 \leq \frac{(n-k-1)s^2}{\chi^2_{1-\alpha/2,n-k-1}}$$

The sas code and output are as following:

```
/*confidence interval for sigma^2*/
lower_sigma = (n - k_new - 1)*s_2_new/cinv(0.975, n - k_new - 1);
upper_sigma = (n - k_new - 1)*s_2_new/cinv(0.025, n - k_new - 1);
print lower_sigma upper_sigma;
```

lower_sigma	upper_sigma
2.4291027	17.1117

So the 95% *confidence interval for* σ^2 *is* $(2.4291027, 17.1117)$.

Thus finished the solution of Question 3.