University of Dayton

College of Arts and Sciences

Department of Mathematics

MTH 547: Design of Experiments – Fall 2021

Final Project: Effect on Water Filtration Time

Instructor: Dr. Maher Qumsiyeh

Student Name: Serge Alhalbi **Student ID:** 101682971

Date: 12/6/2021

I- FFD 1:
SAS code:

```
Data FFD1; /*Unreplicated 2^7-4 FFD with resolution 3 (max)*/
Input y   A B C D E F G;
/*AB= A*B; AC=A*C; BC=B*C; ABC= A*B*C;*/
Lines;
68.4  -1  -1  -1  +1  +1  +1  -1
77.7  +1  -1  -1  -1  -1  +1  +1
66.4  -1  +1  -1  -1  +1  -1  +1
81.0  +1  +1  -1  +1  -1  -1  -1
78.6  -1  -1  +1  +1  -1  -1  +1
41.2  +1  -1  +1  -1  +1  -1  -1
68.7  -1  +1  +1  -1  -1  +1  -1
38.7  +1  +1  +1  +1  +1  +1  +1
;

proc glm data = FFD1; /*Half normal plot*/
model y = A|B|C /solution;
ods output ParameterEstimates=PE1;
run;
quit;
data PE2;
set PE1;
estimate = abs(estimate);
if _n_>1;
drop StdErr tValue Probt Biased;
run;
proc rank data = PE2 out = PE3;
var estimate;
ranks u; run;
data PE4;
set PE3;
zscore = probit (.5+.5*((u-0.5)/7)); run;
proc sgplot data = PE4;
scatter y = zscore x = estimate/datalabel = Parameter;
yaxis label ='Half Normal Scores';
title 'Half normal Probability Plot'; run;

proc glm data = FFD1; /*Testing the significance of the parameters*/
class A /*B*/ C E;
model y = A C E /*B*C*/; run;
proc glm data = FFD1; /*Best setting*/
class A C;
model y = A|C;
lsmeans A|C / pdiff = all adjust = Tukey lines;
run;
```

- It's an unreplicated **2^{7-4} FFD design** with 7 factors and 8 runs: (2^{k-p}) with k = 7 and p = 4.

  -The design generators are 4 since p = 4: ABD = +I, ACE = +I, BCF = +I, ABCG = +I.

  -Resolution: III since main factors are aliased with at least 2 way factor interactions. (Maximum resolution)

  -The defining relations are 15 since 2^p -1 = 15: ABD = +I, ACE = +I, BCF = +I, CDG = +I BCDE = +I, ACDF = +I, ABCG = +I, ABEF = +I, ADEG = +I, BDFG = +I, DEF = +I, BEG = +I, AFG = +I, ABCDEFG = +I, CEFG = +I.

  -The alias structure consisting of 7 relations since $2^{k-p} – 1 = 7$ is the following:
  A = BD = BCG = ABCF = ACDG = CE = ABEG = CDF = BEF = DEG = ADEF = FG
  B = BCDG = ABCE = CDE = ACG = BDEF = ABFG = AD = CF = DFG = EG = AEF
  C = ABCD = BCEG = BDE = ABG = BF = DG = CDEF = ACFG = ADF = AE = EFG
  D = BCE = AB = CG = BCDF = BDEG = BFG = ACDE = ACF = ADFG = AEG = EF
  E = BCD = BG = AC = ABDE = BCEF = CDEG = ABF = ADG = CFG = DF = AEFG
  F = BC = BDG = ACD = ABE = CEG = ABDF = CDFG = BEFG = DE = AG = ACEF
  G = ABC = CD = BE = ABDG = BCFG = BDF = ACEG = ADE = CEF = DEFG = AF

  -The equation of the model reduces to the following: (D, E, F, and G in terms of A, B, and C)
  $y = \beta_0 + \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_{AB} x_{AB} + \beta_{AC} x_{AC} + \beta_{BC} x_{BC} + \beta_{ABC} x_{ABC} + \varepsilon.$

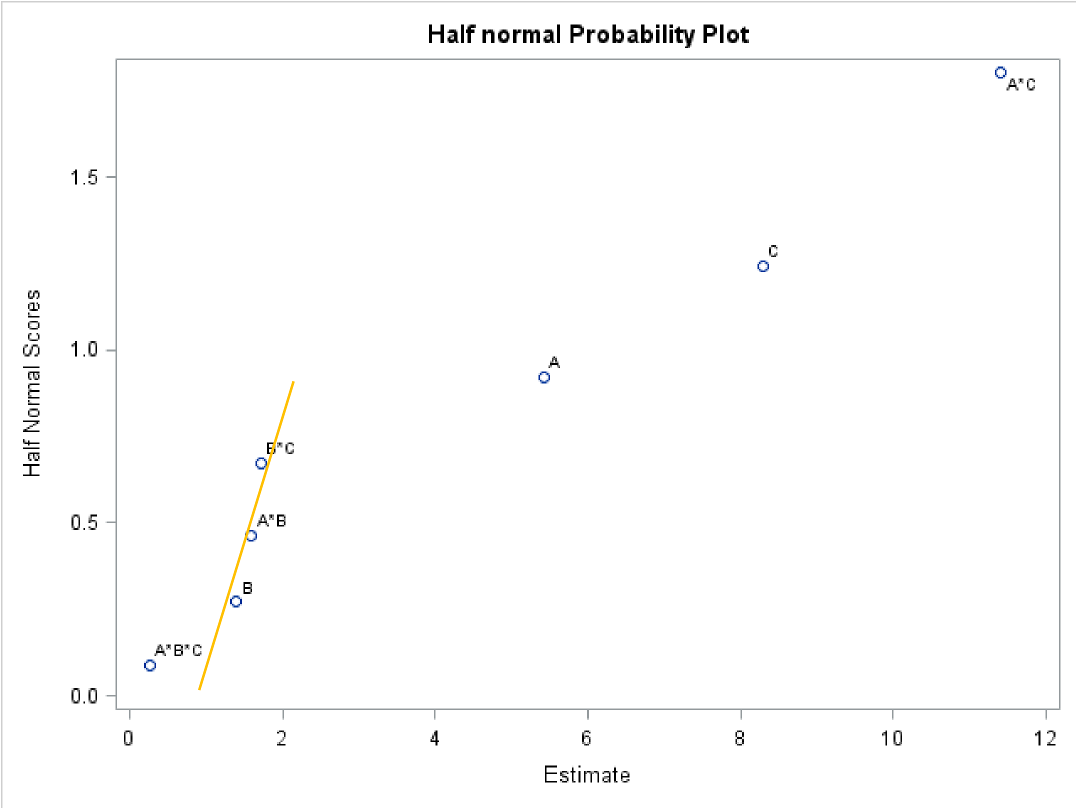
  -Degree of Freedom: Where: $df_{Factor/Interaction} = 1$
  $SS_{Total} = SS_A + SS_B + \cdots + SS_{ABC} + SS_{Error}$

  $df_{Total} = df_A + df_B + \cdots + df_{ABC} + df_{Error}$

  $2^{k-p} - 1 = 7 + 0$ (Respectively) => $df_{Error} = 0$

- Model Analysis:

  -Half normal plot: (since $df_{Error} = 0$)



Using SAS: ***A, C, and AC should be significant***

-The equation of the model reduces to the following: (E = AC)

$$y = \beta_0 + \beta_A x_A + \beta_C x_C + \beta_{AC} x_{AC} + \varepsilon.$$

-Degree of Freedom: Where: $df_{Factor/Interaction} = 1$

$SS_{Total} = SS_A + SS_B + SS_{AC} + SS_{Error}$

$df_{Total} = df_A + df_B + df_{AC} + df_{Error}$

$2^{k-p} - 1 = 3 + 4$ (Respectively) => $df_{Error} = 4$

-Testing if the model is fit ($\alpha = 0.05$):

$H_0: \beta_A = \beta_B = \beta_{AC} = 0$ Vs $H_a: At\ least\ one\ is\ \neq 0$

$$Test\ Statistic = F = \frac{MS_{Model}}{MS_E} = 40.91$$

$p-value = 0.0018$ (**Small**) => null hypothesis is rejected => **Model is fit.**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1827.953750	609.317917	40.91	0.0018
Error	4	59.575000	14.893750		
Corrected Total	7	1887.528750			

R-Square	Coeff Var	Root MSE	y Mean
0.968438	5.929314	3.859242	65.08750

-R-Square and R-Square Adjusted:

* $R^2 = \frac{SS_{Model}}{SS_{Total}} = 0.9684$ => 96.84% of the variations in the response y are explained by the model.

* $R_{adj}^2 = 1 - \frac{SS_E/df_E}{SS_{Total}/df_{Total}} = 0.9448$ => 94.48% of the variations in the response y are explained by the model (Used since the number of regressors affecting the response is high).

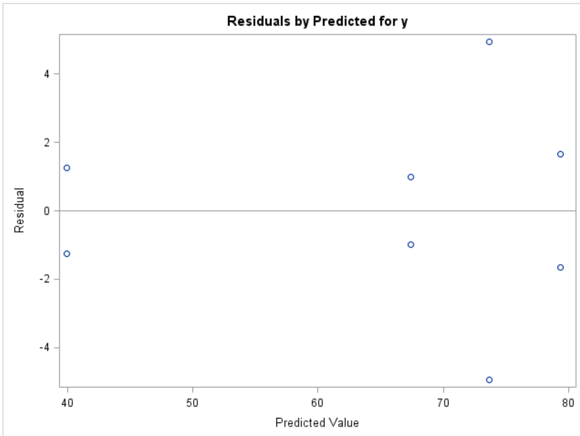
-Testing the significance of the parameters:

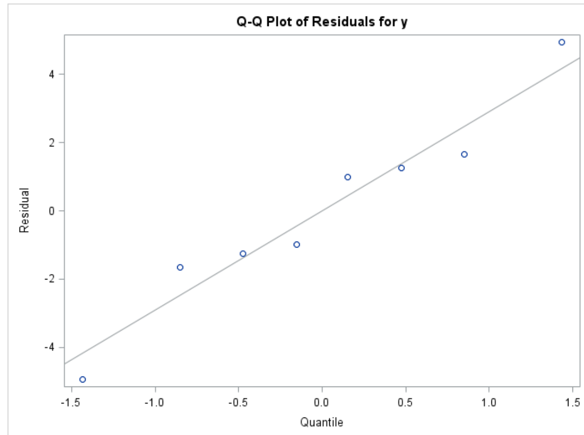
$H_0: Parameter = 0$ Vs $H_a: Parameter\ \neq 0$ (One by one)

${\beta_A}_{F-value} = 15.88$, ${\beta_A}_{P-value} = 0.0163 < \alpha$ (Small) => $\boldsymbol{\beta_A}$ **is significant: A is active.**
${\beta_C}_{F-value} = 36.89$, ${\beta_C}_{P-value} = 0.0037 < \alpha$ (Small) => $\boldsymbol{\beta_C}$ **is significant: C is active.**
${\beta_{AC}}_{F-value} = 69.96$, ${\beta_{AC}}_{P-value} = 0.0011 < \alpha$ => $\boldsymbol{\beta_{AC}}$ **is significant: AC or E is active.**

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	236.531250	236.531250	15.88	0.0163
C	1	549.461250	549.461250	36.89	0.0037
E	1	1041.961250	1041.961250	69.96	0.0011

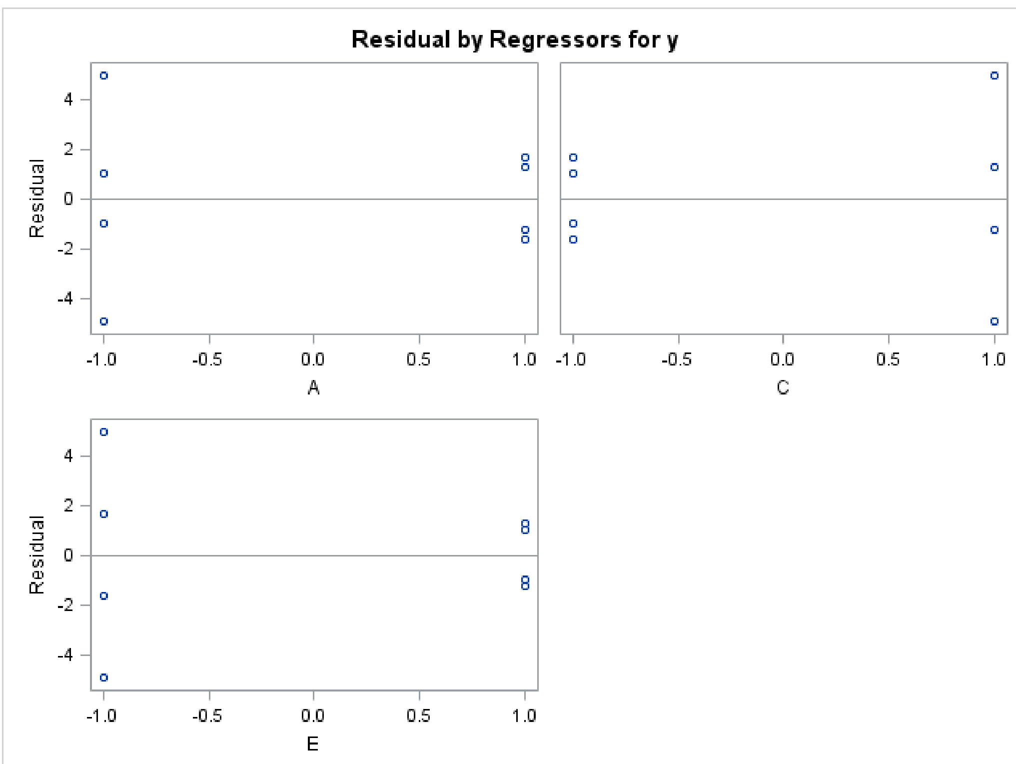
- Adequacy Plots:





No pattern: The residuals have equal variances

Almost straight line: The residuals seem to be normally distributed



We can't decide whether the variances are equal amongst the level of each factor since we have few points

- Settings for optimal response: (Minimum time)

Tukey Comparison Lines for Least Squares Means of A*C				
LS-means with the same letter are not significantly different.				
	y LSMEAN	A	C	LSMEAN Number
A	79.35	1	-1	3
A				
A	73.65	-1	1	2
A				
A	67.40	-1	-1	1
B	39.95	1	1	4

A	C	y LSMEAN	LSMEAN Number
-1	-1	67.4000000	1
-1	1	73.6500000	2
1	-1	79.3500000	3
1	1	39.9500000	4

Least Squares Means for effect A*C Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: y				
i/j	1	2	3	4
1		0.4609	0.1140	0.0071
2	0.4609		0.5236	0.0033
3	0.1140	0.5236		0.0018
4	0.0071	0.0033	0.0018	

As we can see from the grouping table one and only one setting minimizes the response y:

Group B: $\mu_4 \neq (\mu_1 = \mu_2 = \mu_3)$ => A = 1; C = 1 for an average of 39.95 minutes.

Or: Water supply source: Well; Temperature: High.

-Moreover, the p-values from the SAS table assure that this is correct:
* $(\mu_4 = \mu_1)_{p-value} = 0.0071 < \alpha$ (Small) It means that $\mu_4 = \mu_1$ is rejected.

* $(\mu_4 = \mu_2)_{p-value} = 0.0033 < \alpha$ (Small) It means that $\mu_4 = \mu_1$ is rejected.

* $(\mu_4 = \mu_3)_{p-value} = 0.0018 < \alpha$ (Small) It means that $\mu_4 = \mu_1$ is rejected.

II- FFD 2:
SAS code:

```
Data FFD2; /*Unreplicated 2^7-3 FFD with resolution 4 (max)*/
Input y    A  B  C  D  E  F  G;
Lines;
68.4  -1  -1  -1  +1  +1  +1  -1
77.7  +1  -1  -1  -1  -1  +1  +1
66.4  -1  +1  -1  -1  +1  -1  +1
81.0  +1  +1  -1  +1  -1  -1  -1
78.6  -1  -1  +1  +1  -1  -1  +1
41.2  +1  -1  +1  -1  +1  -1  -1
68.7  -1  +1  +1  -1  -1  +1  -1
38.7  +1  +1  +1  +1  +1  +1  +1
66.7  +1  +1  +1  -1  -1  -1  +1
65.0  -1  +1  +1  +1  +1  -1  -1
86.4  +1  -1  +1  +1  -1  +1  -1
61.9  -1  -1  +1  -1  +1  +1  +1
47.8  +1  +1  -1  -1  +1  +1  -1
59.0  -1  +1  -1  +1  -1  +1  +1
42.6  +1  -1  -1  +1  +1  -1  +1
67.6  -1  -1  -1  -1  -1  -1  -1
;

proc glm data = FFD2; /*Half normal plot*/
model y = A|B|C|D /solution;
ods output ParameterEstimates = PE1; run; quit;
data PE2;
set PE1;
estimate = abs(estimate);
if _n_>1;
drop StdErr tValue Probt Biased; run; quit;
proc rank data = PE2 out = PE3;
var estimate;
ranks u; run;
data PE4;
set PE3;
zscore = probit (.5+.5*((u-0.5)/15)); run;
proc sgplot data = PE4;
scatter y = zscore x = estimate/datalabel = Parameter;
yaxis label = 'Half Normal Scores';
title 'Half normal Probability Plot'; run;

Proc glm data = FFD2 plots = all diagnostics(unpack); /*Testing the significance of the parameters*/
Class A B C D E F;
Model y =  A /*ABCD is */A*E /*BCD is */E / p; Run;
proc glm data = FFD2; /*Best setting*/
class A E;
model y = A|E;
lsmeans A|E / pdiff = all adjust = Tukey lines; run;
```

- It's an unreplicated **2^{7-3} FFD design** with 7 factors and 8 runs: (2^{k-p}) with k = 7 and p = 3.

  -The design generators are 3 since p = 3: BCDE = +I, ACDF = +I, ABCG = +I.

  -Resolution: IV since main factors are aliased with at least 3 way factor interactions. (Maximum resolution)

  -The defining relations are 7 since 2^p -1 = 7: BCDE = +I, ACDF = +I, ABCG = +I, ABEF = +I, ADEG = +I, BDFG = +I, CEFG = +I.

  -The alias structure consisting of 15 relations since $2^{k-p} - 1 = 15$ is the following:
  A = BEF = BCG = CDF = DEG
  B = AEF = ACG = DFG = CDE
  C = ABG = BDE = ADF = EFG
  D = BFG = BCE = AEG = ACF
  E = ABF = BCD = ADG = CFG
  F = ABE = BDG = ACD = CEG
  G = ABC = BDF = ADE = CEF
  AB = BCDF = BDEG = ACDE =ADFG = EF = CG
  AC = BG = ABDE = BCEF = AEFG = DF = CDEG
  AD = ABCE = ABFG = BDEF = BCDG = CF = EG
  AE = BF = ABCD = BCEG = ACFG = DG = CDEF
  AF = BE = ABDG = BCFG = ACEG = CD = DEFG
  AG = BC = ABDF = BEFG = ACEF = DE = CDFG
  BD = ABEG = ABCF = ADEF = ACDG = FG = CE
  ABD = BCF = BEG = ACE = AFG = DEF = CDG

  -The equation of the model reduces to the following: (E, F, and G in terms of A, B, C, and D)
  $y = \beta_0 + \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_D x_D + \beta_{BCD} x_{BCD} + \beta_{ACD} x_{ACD} + \beta_{ABC} x_{ABC} + \beta_{AB} x_{AB} + \beta_{AC} x_{AC} + \beta_{AD} x_{AD} + \beta_{ABCD}\, x_{ABCD} + \beta_{CD}\, x_{CD} + \beta_{BC}\, x_{BC} + \beta_{BD}\, x_{BD} + \beta_{ABD}\, x_{ABD} + \varepsilon.$

-Degree of Freedom: Where: $df_{Factor/Interaction} = 1$

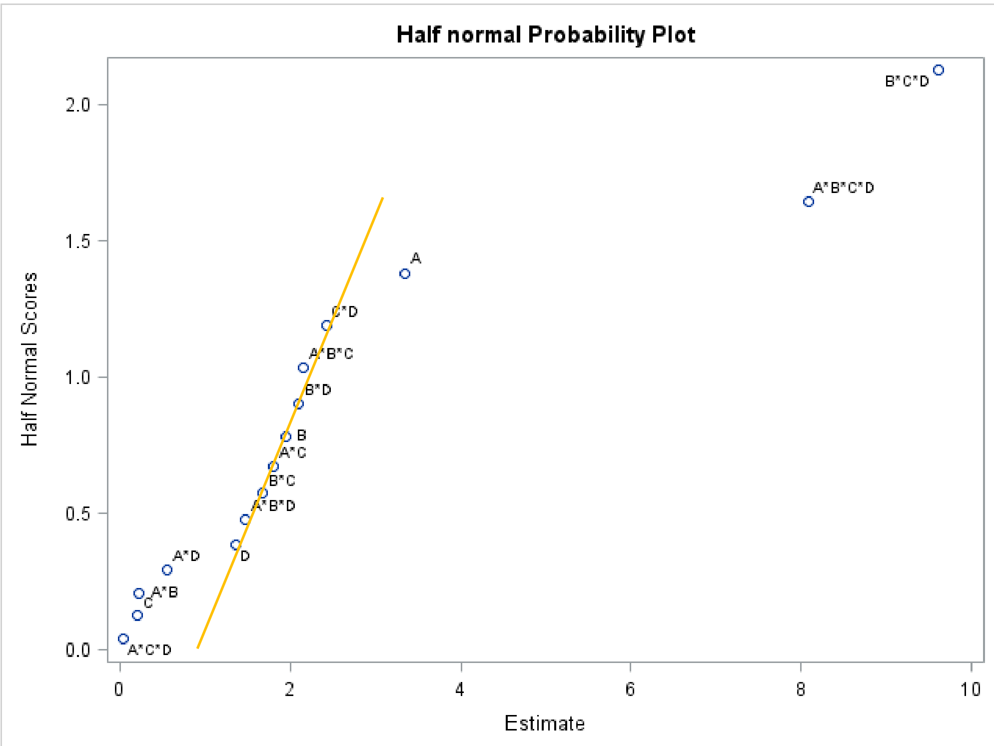
$SS_{Total} = SS_A + SS_B + \cdots + SS_{ABD} + SS_{Error}$

$df_{Total} = df_A + df_B + \cdots + df_{ABD} + df_{Error}$

$2^{k-p} - 1 = 15 + 0$ (Respectively) => $df_{Error} = 0$

- Model Analysis:

  -Half normal plot: (since $df_{Error} = 0$)



*Using SAS: **A, ABCD, and BCD should be significant***

-The equation of the model reduces to the following: (AE = ABCD, E = BCD)

$y = \beta_0 + \beta_A x_A + \beta_E x_E + \beta_{AE} x_{AE} + \varepsilon.$

-Degree of Freedom: Where: $df_{Factor/Interaction} = 1$

$SS_{Total} = SS_A + SS_E + SS_{AE} + SS_{Error}$

$df_{Total} = df_A + df_E + df_{AE} + df_{Error}$

$2^{k-p} - 1 = 3 + 12$ (Respectively) => $df_{Error} = 12$

-Testing if the model is fit ($\alpha = 0.05$):

$\mathrm{H_0}: \beta_A = \beta_E = \beta_{AE} = 0$ Vs $\mathrm{H_a}: At\ least\ one\ is\ \neq 0$

$$Test\ Statistic = F = \frac{MS_{Model}}{MS_E} = 23.13$$

$p-value < .0001$ (**Small**) => null hypothesis is rejected => **Model is fit.**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2700.276875	900.092292	23.13	<.0001
Error	12	467.052500	38.921042		
Corrected Total	15	3167.329375			

R-Square	Coeff Var	Root MSE	y Mean
0.852541	9.808271	6.238673	63.60625

-R-Square and R-Square Adjusted:

* $\mathrm{R}^2 = \frac{SS_{Model}}{SS_{Total}} = 0.8525$ => 85.25% of the variations in the response y are explained by the model.
* $\mathrm{R_{adj}}^2 = 1 - \frac{SS_E/df_E}{SS_{Total}/df_{Total}} = 0.8157$ => 81.57% of the variations in the response y are explained by the model (Used since the number of regressors affecting the response is high).

-Testing the significance of the parameters:

$\mathrm{H_0}: Parameter = 0$ Vs $\mathrm{H_a}: Parameter\ \neq 0$ (One by one)

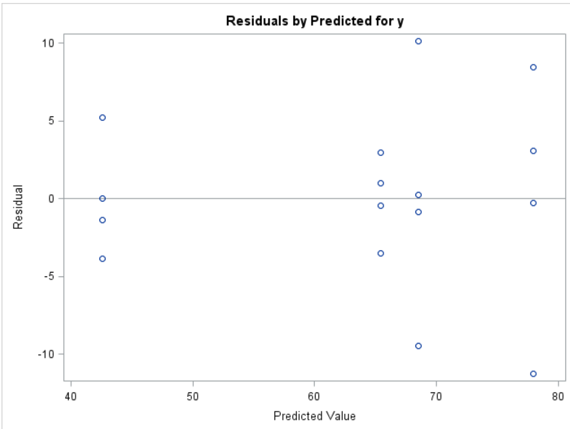
${\beta_A}_{F-value} = 4.60$, ${\beta_A}_{P-value} = 0.0532 \approx \alpha$ (Small) => $\boldsymbol{\beta_A}$ **is significant: A is active.**

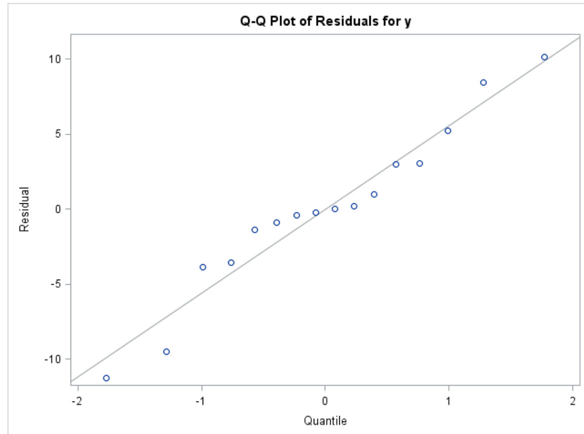
${\beta_E}_{F-value} = 37.94$, ${\beta_E}_{P-value} < .0001 < \alpha$ (Small) => $\boldsymbol{\beta_E}$ **is significant: E is active.**

${\beta_{AE}}_{F-value} = 26.85$, ${\beta_{AE}}_{P-value} = 0.0002 < \alpha$ (Small) => $\boldsymbol{\beta_{AE}}$ **is significant: AE is active.**

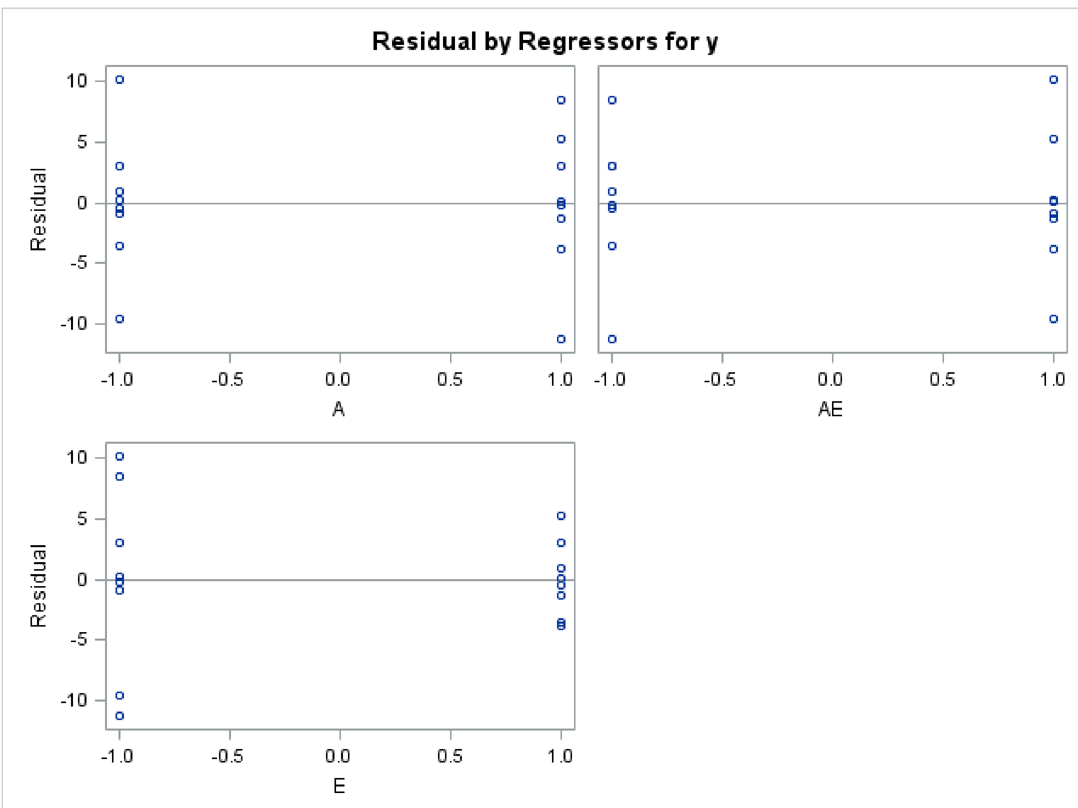
Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	178.890625	178.890625	4.60	0.0532
A*E	1	1044.905625	1044.905625	26.85	0.0002
E	1	1476.480625	1476.480625	37.94	<.0001

* Adequacy Plots:





No pattern: ==*The residuals have equal variances*== *Almost straight line:* ==*The residuals seem to be normally distributed*==



Equal variances amongst the residuals and the settings of the active factors and factor interactions

- Settings for optimal response: (Minimum time)

Tukey Comparison Lines for Least Squares Means of A*E				
LS-means with the same letter are not significantly different.				
	y LSMEAN	A	E	LSMEAN Number
A	77.950	1	-1	3
A				
A	68.475	-1	-1	1
A				
A	65.425	-1	1	2
B	42.575	1	1	4

A	E	y LSMEAN	LSMEAN Number
-1	-1	68.4750000	1
-1	1	65.4250000	2
1	-1	77.9500000	3
1	1	42.5750000	4

Least Squares Means for effect A*E Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: y				
i/j	1	2	3	4
1		0.8984	0.1933	0.0004
2	0.8984		0.0625	0.0011
3	0.1933	0.0625		<.0001
4	0.0004	0.0011	<.0001	

As we can see from the grouping table one and only one setting minimizes the response y:

Group B: $\mu_4 \neq (\mu_1 = \mu_2 = \mu_3)$ => A = 1; E = 1 for an average of 42.575 minutes.

Or: Water supply source: Well; Caustic Soda Addition Rate: Fast.

Moreover, the p-values from the SAS table assure that this is correct:
* $(\mu_4 = \mu_1)_{p-value} = 0.0071 < \alpha$ (Small) It means that $\mu_4 = \mu_1$ is rejected.
* $(\mu_4 = \mu_2)_{p-value} = 0.0033 < \alpha$ (Small) It means that $\mu_4 = \mu_1$ is rejected.
* $(\mu_4 = \mu_3)_{p-value} = 0.0018 < \alpha$ (Small) It means that $\mu_4 = \mu_1$ is rejected.

- Conclusion:

  Comparing the same experiment with 8 runs to the one with 16 runs. We find that the active factors are different. However, **factors A and E are active in both cases**.
  As well, the best setting remained the same in both designs:
  In design I, the best setting was A = 1; C = 1, but AC = E, this implies that the best setting is A = 1; E = 1. (Same as in design II)
  Thus, we can conclude that still running the experiment few runs is still sometime somewhat accurate.