STAT 523 HW 8

Md Muhtasim Billah

3/25/2020

Question 1

(From Jay L. Devore, Exercise 40, Chapter 11, Section 11.4) In a study of processes used to remove impurities from cellulose goods ("Optimization of Rope-Range Bleaching of Cellulosic Fabrics," Textile Research J., 1976: 493–496), the following data resulted from a 24 experiment involving the desizing process. The four factors were enzyme concentration (A), pH (B), temperature (C), and time (D). The ANOVA table for analyzing the dataset is given. Follow the instructions below.

a. Identify the effects (main and interactions) that are significant at the $\alpha = 0.10$ level.
b. Compute the coefficient of determination R^2 and give a brief interpretation. (See Formula 11K and pages L-71 to L-72) of lecture notes.)
c. Computed R^2_{adj}, the adjusted R^2, (See Formula 11K and pages L-71 to L-72) of lecture notes.)
d. Assume that only the main and interaction effects of C and D are important (two-way interaction model). Complete the ANOVA table below in this case. Hint: You can derive the missing values from the previous ANOVA table.
e. Compute R^2_{adj} for the ANOVA table in part (d).
f. Assume that only the main effects of C and D (two-way additive model) are important. Below is the corresponding ANOVA table. Compute R^2_{adj}.
g. Based on the R^2_{adj} values, which of the 3 fitted models do you prefer? Give a short explanation for your choice.

Answer

a.

The dataset for the problem comes from a 2^P factorial experiment which has 4 factors A (conc), B (pH), C (temperature) and D (time). Each factor have two levels with two replicates for each combination. The response is the Starch % by Weight. The ANOVA table for the dataset can generated in R as below.

```
library(Devore7) #loading the library for Devore's textbook exercise
data(ex11.40) ## load the dataset for the specific problem
out = aov(sizing~conc*pH*tempture*time, data=ex11.40) #run ANOVA with this datset
summary(out) #print out summary
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## conc                  1   0.17   0.168   0.075 0.7876
## pH                    1   1.94   1.940   0.866 0.3659
## tempture              1   8.16   8.161   3.642 0.0744 .
## time                  1  13.08  13.082   5.838 0.0280 *
## conc:pH               1   3.42   3.419   1.526 0.2346
## conc:tempture         1   0.26   0.263   0.117 0.7364
## pH:tempture           1   0.74   0.738   0.329 0.5740
## conc:time             1   0.91   0.911   0.407 0.5327
```

```
## pH:time                1   0.78   0.781   0.349 0.5631
## tempture:time          1   6.77   6.771   3.022 0.1013
## conc:pH:tempture       1   0.02   0.020   0.009 0.9259
## conc:pH:time           1   0.78   0.775   0.346 0.5647
## conc:tempture:time     1   0.62   0.622   0.277 0.6056
## pH:tempture:time       1   1.76   1.758   0.785 0.3889
## conc:pH:tempture:time  1   0.00   0.004   0.002 0.9666
## Residuals             16  35.85   2.241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the ANOVA table, all the p-values are listed for the main effect of the four factors as well as for their interactions. For the main effect of the factors, the null hypothesis, H_0 is that there is no main effect of the different levels of that factor. For the interactions among the factors, the null hypothesis, H_0 is that there is no interaction. At $\alpha = 0.10$, from the ANOVA table above, it is visible that only the factors C and D have significant main effect. It is also visible that there are no interactions among any of the factors at this significance level.

To summarize, at $\alpha = 0.10$,

i) Only main effects of C and D are significant.
ii) None of the interactions are significant.

b.

The coefficient of determination is a common number used to compare model fits which is often given as a percentage,

$$R^2 = (1 - \frac{SSE}{SST}) \times 100\%$$

From the given ANOVA table, $SSE = 35.851$ and $SST = 75.264$. Plugging the values, we get,

$$R^2 = (1 - \frac{35.851}{75.264}) \times 100\% = 52.37\%$$

Interpretation:

This value of R^2 indicates that 52.37% of the variations in the response due to the factor effects (main and interactions) are explained by this model which is not good enough. A R^2 value close to 100% is ideal which means that most of the variance is explained by the model. If the model gives a R^2 value much smaller than 100%, reducing the model (excluding some of the main or interaction effects of the factors) can improve the R^2 which in term may lead to a better estimation.

c.

A better diagnostic of the model is provided by the adjusted R^2 which accounts for the size of the model (degree of freedom) which is given as,

$$R^2_{adj} = (1 - \frac{total\ df}{error\ df} \times \frac{SSE}{SST}) \times 100\%$$

From the given ANOVA table, $total\ df = 31$ and $error\ df = 16$. Plugging the values, we get,

$$R^2_{adj} = (1 - \frac{31}{16} \times \frac{35.851}{75.264}) \times 100\% = 7.71\%$$

It is seen that the R^2_{adj} has a very low value which indicates the model is not very good.

d.

The given ANOVA table is for the case when only the main and interaction effects of factor C (tempture) and D (time) are important. The missing SS and MS values can be found from the previous ANOVA table (for the full model) since they will not change. Then, the F-statistic for them can be found from the usual formula. The completed ANOVA table is given below.

Source	DF	SS	MS	F	p-value
C	1	8.16	8.161	4.836	0.036
D	1	13.082	13.082	7.752	0.010
C*D	1	6.771	6.771	4.013	0.055
Error	28	47.25	1.688		
Total	31	75.264			

The same ANOVA table was also be generated in R to check the values and the values were found to be exact match.

```
library(Devore7) #loading the library for Devore's textbook exercise
data(ex11.40) ## load the dataset for the specific problem
out = aov(sizing~tempture+time+tempture:time, data=ex11.40) #run ANOVA with this datset
summary(out) #print out summary
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## tempture       1   8.16   8.161   4.836 0.03630 *
## time           1  13.08  13.082   7.752 0.00951 **
## tempture:time  1   6.77   6.771   4.013 0.05493 .
## Residuals     28  47.25   1.688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e.

From the ANOVA table in (d), $SSE = 47.25$, $SST = 75.264$ and *total df* $= 31$, *error df* $= 28$. Plugging the values, we get the adjusted R^2,

$$R^2_{adj} = (1 - \frac{31}{28} \times \frac{47.25}{75.264}) \times 100\% = 30.49\%$$

f.

The corresponding ANOVA table is given for when only the main effects of C and D (two-way additive model) are important. From the table, $SSE = 54.0214$, $SST = 75.2638$ and *total df* $= 31$, *error df* $= 29$. Plugging the values, we get the adjusted R^2 for this model,

$$R^2_{adj} = (1 - \frac{31}{29} \times \frac{54.0214}{75.2638}) \times 100\% = 23.27\%$$

g.

The R^2_{adj} values for the 3 fitted models are given in the table below.

Model	Model Description	R^2_{adj} (%)
1	Factors A, B, C, D (Main and Interaction Effects)	7.71
2	Factors C, D (Main and Interaction Effects)	30.49
3	Factors C, D (Main Effect)	23.27

Looking at the R^2_{adj} values, Model 2 will be my preferred choice which can be expressed as below:

$$X = \mu + \delta_k + \chi_l + \gamma_{kl}CD$$

Explanation:

It is knwon that higher the value of R^2_{adj}, better the goodness of fir for a model. Among the three models discussed in this problem, Model 2 has the highest value of R^2_{adj} which is 30.49%. This means that 30.49% of the variance in the response was explained by this model. So, Model 2 is the best of these three models to explain the variablity of the given response variable and thus it is my preferred choice.

Question 2

A Fractional Factorial Design. In a study of p = 5 factors with two levels each, only 8 different combinations of levels of A, B, C, D, and E will be run (i.e. q = 2 and the experiment is a quarter fractional). Use A, B, D as the independent factors and with generators C = ABD and E = AD.

a. List out the 8 treatments that will be run.
b. Which treatments, used in the experiment, have all of B and D at their low levels?
c. What effects will be aliased with the A main effect?

Answer

a.

Given,
Number of factors, $p = 5$.
For a quarter fractional experiment, $q = 2$.
Number of treatments for this fractional factorial design $= 2^{p-q} = 2^{5-2} = 2^3 = 8$.
Independent factors $= A,\ B,\ D$.
Generators are, $C = ABD$ and $E = AD$.

The 8 treatments that will be run are given in the table below.

No.	A	B	D	C = ABD	E = AD	Treatment
1	−	−	−	−	+	e
2	−	+	−	+	+	bce
3	−	−	+	+	−	cd
4	−	+	+	−	−	bd
5	+	−	−	+	−	ac
6	+	+	−	−	−	ab
7	+	−	+	−	+	ade
8	+	+	+	+	+	abcde

b.

Looking at the table, we can see that only treatment No. 1 and 5 have all of B and D at their low levels.

c.

For determining the alias structure, let's take the q generators and apply multiplication so that I is on the left hand side of the equation. Thus,

$$C = ABD$$

$$C * C = C * ABD$$

$$I = ABCD$$

And,

$$E = AD$$

$$E * E = E * AD$$

$$I = ADE$$

Now, multiplying both the equations (LHS x LHS and RHS x RHS),

$$I * I = ABCD * ADE$$

$$I = BCE$$

Combining, we get the following relation,

$$I = ABCD = ADE = BCE$$

The above equation is called the "defining relation" for the experiment. It tells us exactly what factor effects are aliased or confounded together. Here, we see that I is equivalent to $2^q - 1 = 2^2 - 1 = 3$ products which was expected. There should be $2^{p-q} = 2^{5-2} = 2^3 = 8$ aliased relations (including the defining one).

Now, to determine what effects will be aliased with the main effect of A, we multiply the "defining relation" with A to get,

$$A * I = A * ABCD = A * ADE = A * BCE$$

$$A = BCD = DE = ABCE$$

The above equation means that,

- The main effect α_2 of A is aliased with the BCD, DE and ABCE effects.
- We cannot estimate α_2, ${\gamma_{222}}^{BCD}$, ${\gamma_{22}}^{DE}$ and ${\gamma_{2222}}^{ABCE}$ individually but we can estimate $\alpha_2 + {\gamma_{222}}^{BCD} + {\gamma_{22}}^{DE} + {\gamma_{2222}}^{ABCE}$.