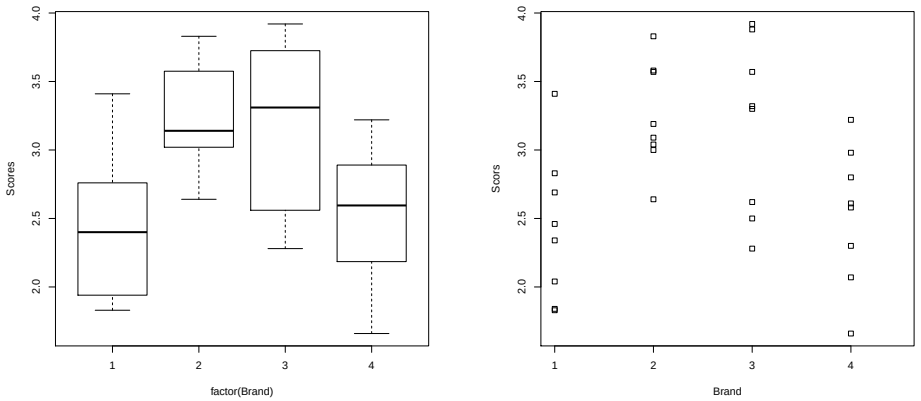
Spring 16 – AMS256 Homework 5

1. (a) The figures below show some differences in scores by brands. Consider the model, $y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \ldots, 4, j = 1, \ldots, 8$. With the constraint, $\sum_{i=1}^{2} \alpha_i = 0$, the estimates are $\hat{\mu} = 2.84$, $\hat{\alpha}_1 = -0.41344$, $\hat{\alpha}_2 = 0.399$, $\hat{\alpha}_3 = 0.33$ and $\hat{\alpha}_4 = 0.41422 - 0.399 - 0.33 = -0.31478$. The F test for $H : \alpha_1 = \ldots = \alpha_4$ has p-value 0.005434. We conclude from the small p-value for the F-statistic that there is some difference between the brands. There is need to further investigate which brands differ.




```
> options(contrasts=c("contr.sum", "contr.poly"))
> g <- lm(Scores ~ factor(Brand), dat)
> summary(g)

Call:
lm(formula = Scores ~ factor(Brand), data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89375 -0.40687  0.04125  0.35219  0.98000

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.84344    0.09272  30.666   <2e-16 ***
factor(Brand)1 -0.41344    0.16060  -2.574   0.0156 *
factor(Brand)2  0.39906    0.16060   2.485   0.0192 *
factor(Brand)3  0.33031    0.16060   2.057   0.0491 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5245 on 28 degrees of freedom
Multiple R-squared:  0.3589,Adjusted R-squared:  0.2902
F-statistic: 5.225 on 3 and 28 DF,  p-value: 0.005434

> anova(g)
Analysis of Variance Table

Response: Scores
              Df Sum Sq Mean Sq F value   Pr(>F)
factor(Brand)  3 4.3128 1.43761  5.2253 0.005434 **
```

```
Residuals     28 7.7035 0.27512
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(b) *(Please read Section 4.2 of Ronald's book before answering this question. This may be helpful to follow my solution.)* Recall a contrast is a function of $\sum_{i=1}^{a} c_i\alpha_i$ with $\sum_{i=1}^{a} c_i = 0$. Here our $a = 4$. Since the first column of $\boldsymbol{X}$, $\mathbf{1}$ is the column associated with μ, we can have three linearly independent contrasts at most (or we can say that from the ANOVA table, Brands have three degrees of freedom so we can have three linearly independent contrasts). Let's consider three orthogonal contrasts, $\boldsymbol{c}_1^T = [0, 1, -1, 0]$, $\boldsymbol{c}_2^T = [1, 0, 0, -1]$ and $\boldsymbol{c}_3^T = [1/2, -1/2, -1/2, 1/2]$. We can easily check the three contrasts are orthogonal, that is, $\sum_i c_{ki}c_{k'i} = 0$, $k \neq k'$, $k, k' = 1, 2, 3$. The first contrast, $\boldsymbol{c}_1$ is for testing a difference between two inexpensive brands, the second $\boldsymbol{c}_2$ for testing a difference between the two costly brands, and the third $\boldsymbol{c}_3$ for testing a difference between the mean of Brands 1 and 4 (costly) and the mean of Brands 2 and 3 (inexpensive). The first two contrasts compare brands in the same price category and the third compares the two price categories. Using the formulas (see the R output given below), the sums of squares for the three contrasts are 0.0189, 0.038025, 4.2559 (check the sums of these three SS is equal to SS by brands in the above ANOVA table).

```
> ybar <- aggregate(Scores~Brand, dat, mean)$Scores
>
> c1 <- c(0, 1, -1, 0)
> c2 <- c(1, 0, 0, -1)
> c3 <- c(1/2, -1/2, -1/2, 1/2)
>
> n <- 8  ## number of replicates
> SS1 <- (sum(c1*ybar))^2/(sum(c1^2)/n)
> SS1
[1] 0.01890625
>
> SS2 <- (sum(c2*ybar))^2/(sum(c2^2)/n)
> SS2
[1] 0.038025
>
> SS3 <- (sum(c3*ybar))^2/(sum(c3^2)/n)
> SS3
[1] 4.255903
>
> SS1 + SS2 + SS3
[1] 4.312834
>
```

(c) We next do the F-test for each of the contrast identified in (b) (equivalently, t-test). We reject the first two contrasts at significance level 0.05 and fail to reject the third contrast. This implies that there is no statistically significant difference between the brands that cost similarly (Brand2 vs Brand3 and Brand1 vs Brand4), but there is a significant difference between the two groups of brands (inexpensive and very costly).

```
> SS1/(summary(g)$sig)^2
[1] 0.06871888
> SS2/(summary(g)$sig)^2
[1] 0.1382101
> SS3/(summary(g)$sig)^2
[1] 15.46901
>
```

```
> qf(0.95, 1, 28)  ## critical value
[1] 4.195972
```

(d) The mean difference estimates are given in the output below. We compare the estimated difference with the critical values$\times\hat{\sigma}\sqrt{2/n}$ for the LSD method and the Bonferroni method. Using the LSE method, we find that the 2-1, 3-1, 4-3 differences are significant since the absolute difference is greater than 0.7246976. Using the Bonferroni, we don't observe any significant difference between any pair of brands. Tukey's method finds that the $1-2$, $1-3$ and $1-4$ are significant as the corresponding intervals do not contain zero.

```
> ## LSE with alpha=0.01
> qt(1-0.01/2, 28)*(summary(g)$sig)*sqrt(2/n)
[1] 0.7246976
>
> ## Bonferroni with alpha=0.012
> qt(1-0.012/(2*6), 28)*(summary(g)$sig)*sqrt(2/n)
[1] 0.8938282
>
> ## Tukey HSD with alpha=0.01
> TukeyHSD(aov(Scores ~ factor(Brand), dat), conf.level = 0.99)
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = Scores ~ factor(Brand), data = dat)

$`factor(Brand)`
        diff         lwr       upr     p adj
2-1  0.81250 -0.08313327 1.7081333 0.0215658
3-1  0.74375 -0.15188327 1.6393833 0.0395259
4-1  0.09750 -0.79813327 0.9931333 0.9821036
3-2 -0.06875 -0.96438327 0.8268833 0.9935443
4-2 -0.71500 -1.61063327 0.1806333 0.0504445
4-3 -0.64625 -1.54188327 0.2493833 0.0880646
```

2. (a) The R output from regressing the voltage on all the seven independent variables are given the below. The overall F test for $H: \alpha_1 = \ldots = \alpha_7 = 0$ (that is, the test of whether any of the predictors have significance in the model) yields p-value 0.007428. The p-value is small, so the null hypothesis is rejected. $\hat{\beta}_j$, $j = 1, \ldots, 7$ is the effect of $\boldsymbol{x}_j$ when all the other predictors are held constant. For example, as disperse phase volume (x_1) increases by 1 unit, the estimated change in the voltage (y) is -0.0022429. The R^2 for the full model is 0.625. Note that from the sequential sum of squares in the ANOVA table, possibly a smaller model is more desirable.

```
> g <- lm(voltage ~ dphasevolume + salinity + temperature +
timedelay + concentration + spantriton + solidparticles, dat)
> summary(g)

Call:
lm(formula = voltage ~ dphasevolume + salinity + temperature +
    timedelay + concentration + spantriton + solidparticles,
    data = dat)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.68444 -0.23788  0.03217  0.13755  0.74783

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.998082   0.247542   4.032 0.001975 **
dphasevolume   -0.022429   0.005039  -4.451 0.000977 ***
salinity        0.155711   0.074291   2.096 0.060018 .
temperature    -0.017187   0.011860  -1.449 0.175188
timedelay      -0.009527   0.009619  -0.990 0.343279
concentration   0.421421   0.100782   4.182 0.001533 **
spantriton      0.417123   0.437702   0.953 0.361070
solidparticles -0.155244   0.148582  -1.045 0.318516
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4635 on 11 degrees of freedom
Multiple R-squared:  0.771,Adjusted R-squared:  0.6253
F-statistic: 5.292 on 7 and 11 DF,  p-value: 0.007428

> anova(g)
Analysis of Variance Table

Response: voltage
               Df Sum Sq Mean Sq F value   Pr(>F)
dphasevolume    1 1.4016  1.4016  6.5239 0.026794 *
salinity        1 1.9263  1.9263  8.9663 0.012202 *
temperature     1 0.1171  0.1171  0.5452 0.475736
timedelay       1 0.0446  0.0446  0.2075 0.657630
concentration   1 4.0771  4.0771 18.9776 0.001143 **
spantriton      1 0.1565  0.1565  0.7283 0.411627
solidparticles  1 0.2345  0.2345  1.0917 0.318516
Residuals      11 2.3632  0.2148
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

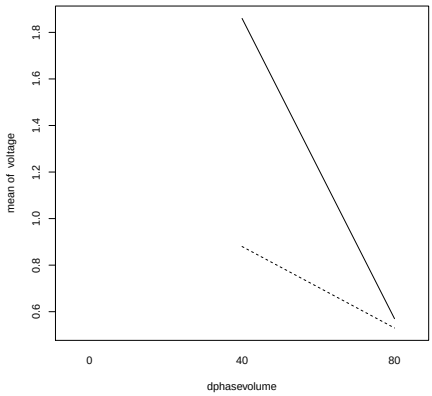
(b) From the notation, $\hat{\beta}_4$ is the coefficient associated with the interactions between disperse phase volume (x_1) and percentage of salinity (x_2), and $\hat{\beta}_5$ is the coefficient associated with the interaction between disperse phase volume (x_1) and percentage in surfactant concentration (x_5). Their estimates, $\hat{\beta}_4$ and $\hat{\beta}_5$ are -0.002804 and 0.001579, respectively (see the R output below). The increase in voltage with a unit increase in disperse phase volume becomes smaller the higher the percentage of salinity and larger the higher the percentage in surfactant concentration. For example, with the percentage of salinity 1 ($x_2 = 1$) and the percentage in surfactant concentration 2 ($x_5 = 2$), the estimated change in the voltage by 1 unit increase in disperse phase volume is -0.01803 ($= -0.022753 - 0.006677 \times 1 + 2 \times 0.0057$). The p-values for the tests $\beta_4 = 0$ and $\beta_5 = 0$ are 0.03924 and 0.01553, respectively. They seem statistically significant marginally. The interaction can be checked graphically from the figures below. Panel (a) of the figure is a plot of voltage (y) vs disperse phase volume (x_1) at the two levels of the percentage of salinity (x_2) to check an interaction between x_1 and x_2. Panel (b) of the figure is a plot of voltage (y) vs disperse phase volume (x_1) at the two levels of the percentage in surfactant concentration (x_5) to check an interaction between x_1 and x_5. You may find my R commands that I

used to make the figures from the below.

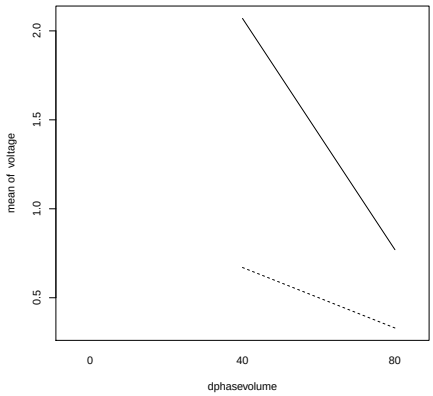
The adjusted R^2 under the full model without interactions and the model with interactions are 0.6253 and 0.5318. $\hat{\sigma}$ =0.4635 and 0.5182, respectively. Those statistics show that the full model fits better. (we will discuss model comparison next week – using AIC, BIC and so on. We can use those here to compare the full vs the model with the interactions)

It would be more desirable to know if those interactions influence the response in important ways by utilizing a priori knowledge.

```
> g1 <- lm(voltage ~ dphasevolume + salinity + concentration + I(dphasevolume*salinity)
+ I(dphasevolume*concentration), dat)
> summary(g1)

Call:
lm(formula = voltage ~ dphasevolume + salinity + concentration +
    I(dphasevolume * salinity) + I(dphasevolume * concentration),
    data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8016 -0.2576  0.1343  0.1855  1.0826

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.905732   0.285463   3.173  0.00734 **
dphasevolume                   -0.022753   0.008318  -2.736  0.01700 *
salinity                        0.304719   0.236600   1.288  0.22023
concentration                   0.274741   0.227048   1.210  0.24780
I(dphasevolume * salinity)     -0.002804   0.003790  -0.740  0.47253
I(dphasevolume * concentration) 0.001579   0.003947   0.400  0.69563
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5047 on 13 degrees of freedom
Multiple R-squared:  0.6792,Adjusted R-squared:  0.5558
F-statistic: 5.505 on 5 and 13 DF,  p-value: 0.006142

> anova(g1)
Analysis of Variance Table

Response: voltage
                               Df Sum Sq Mean Sq F value   Pr(>F)
dphasevolume                    1 1.4016  1.4016  5.5036 0.035492 *
salinity                        1 1.9263  1.9263  7.5639 0.016530 *
concentration                   1 3.5422  3.5422 13.9089 0.002524 **
I(dphasevolume * salinity)      1 0.0994  0.0994  0.3904 0.542924
I(dphasevolume * concentration) 1 0.0408  0.0408  0.1600 0.695631
Residuals                      13 3.3107  0.2547
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```



(a) x_1 vs y by x_2



(b) x_1 vs y by x_5

```
> with(dat, interaction.plot(dphasevolume, salinity, voltage, legend=F))
> with(dat, interaction.plot(dphasevolume, concentration, voltage, legend=F))
```