Ximan Liu
1935858
DSC 423
HW4

I have completed this work independently. The solutions given are entirely my own work.

1)
a)
Voltage = 1.0666667 - 0.1155417 * Volume + 0.6400000 * Salinity + 1.1800000 * Surfactant + 0.0012552 * V2 - 0.0078333 * VSL - 0.0120000 * VSF + e

P.S.
 # V2: WATEROIL$Volume^2. A second-order term
 # VSL: WATEROIL$Volume * WATEROIL$Salinity. An interaction term
 # VSF: WATEROIL$Volume * WATEROIL$Surfactant. An interaction term

```
Call:
lm(formula = Voltage ~ Volume + Salinity + Surfactant + V2 +
    VSL + VSF, data = WATEROIL)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54000 -0.09000  0.01333  0.12500  0.64000 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.0666667  0.1951827   5.465 0.000144 ***
Volume      -0.1155417  0.0232047  -4.979 0.000320 ***
Salinity     0.6400000  0.1781766   3.592 0.003700 ** 
Surfactant   1.1800000  0.2672650   4.415 0.000843 ***
V2           0.0012552  0.0003047   4.119 0.001423 ** 
VSL         -0.0078333  0.0028172  -2.781 0.016634 *  
VSF         -0.0120000  0.0042258  -2.840 0.014906 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3381 on 12 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8007 
F-statistic: 13.05 on 6 and 12 DF,  p-value: 0.0001211
```

b)
To build up the final model, the first step is to build up a simple first-order model as a draft. Then I use cor(WATEROIL) to check the correlations with variables of the dataset. And then I set up squared variables as second-order terms. Additionally, I use anova function to check out interaction terms. After that I build the initial model m3 including first-order terms, second-order terms, and interaction terms. Then step by steps to drop variables and get the final model as showed above.

R code:

```
m1 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart, data = WATEROIL)
summary(m1)

cor(WATEROIL)

WATEROIL$V2 <- WATEROIL$Volume^2
WATEROIL$SL2 <- WATEROIL$Salinity^2
WATEROIL$T2 <- WATEROIL$Temperature^2
WATEROIL$D2 <- WATEROIL$Delay^2
WATEROIL$SF2 <- WATEROIL$Surfactant^2
WATEROIL$ST2 <- WATEROIL$SpanTriton^2
WATEROIL$SP2 <- WATEROIL$SolidPart^2

m2 <- lm(Voltage ~ (Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart)^2, data = WATEROIL)
anova(m2)

WATEROIL$VSL <- WATEROIL$Volume * WATEROIL$Salinity
WATEROIL$VT <- WATEROIL$Volume * WATEROIL$Temperature
WATEROIL$VD <- WATEROIL$Volume * WATEROIL$Delay
WATEROIL$VSF <- WATEROIL$Volume * WATEROIL$Surfactant
WATEROIL$VST <- WATEROIL$Volume * WATEROIL$SpanTriton
WATEROIL$VSP <- WATEROIL$Volume * WATEROIL$SolidPart
WATEROIL$SLT <- WATEROIL$Salinity * WATEROIL$Temperature
WATEROIL$SLD <- WATEROIL$Salinity * WATEROIL$Delay

m3 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + SL2 + T2 + D2 + SF2 + ST2 + SP2 + VSL + VT + VD + VSF + VST + VSP + SLT + SLD,
data = WATEROIL)
summary(m3)

# Drop SLT, SL2, T2, D2, SF2, ST2, SP2
```

```
m4 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + VSL + VT + VD + VSF + VST + VSP + SLD, data = WATEROIL)
summary(m4)

# Drop VT
m5 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + VSL + VD + VSF + VST + VSP + SLD, data = WATEROIL)
summary(m5)

# Drop SLD
m6 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + VSL + VD + VSF + VST + VSP, data = WATEROIL)
summary(m6)

# Drop VSP
m7 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + VSL + VD + VSF + VST, data = WATEROIL)
summary(m7)

# Drop VST
m8 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + VSL + VD + VSF, data = WATEROIL)
summary(m8)

# Drop VD
m8 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton +
SolidPart + V2 + VSL + VSF, data = WATEROIL)
summary(m8)

# Drop SolidPart
m9 <- lm(Voltage ~ Volume + Salinity + Temperature + Delay + Surfactant + SpanTriton + V2 +
VSL + VSF, data = WATEROIL)
summary(m9)

# Drop Delay
m10 <- lm(Voltage ~ Volume + Salinity + Temperature + Surfactant + SpanTriton + V2 + VSL +
VSF, data = WATEROIL)
summary(m10)

# Drop SpanTriton
m11 <- lm(Voltage ~ Volume + Salinity + Temperature + Surfactant + V2 + VSL + VSF, data =
WATEROIL)
summary(m11)
```

# Drop Temperature
m12 <- lm(Voltage ~ Volume + Salinity + Surfactant + V2 + VSL + VSF, data = WATEROIL)
summary(m12)
plot(m12)

c)

For second-order terms, I tried all the possible potential variables (showed as below) and finally only keep V2 (Volume^2) as a significant second-order term. And I do look at scatter plots to determine which second-order terms to evaluate.
For the first strategy, it has high workload since we have at least 7 second-order term to check. If there are relatively few second-order terms to check, this may be a better way. Vice versa, using scatter plots can identify the second-order terms we needed in a fast speed, so it is better to use this strategy here.

R code:
WATEROIL$V2 <- WATEROIL$Volume^2
WATEROIL$SL2 <- WATEROIL$Salinity^2
WATEROIL$T2 <- WATEROIL$Temperature^2
WATEROIL$D2 <- WATEROIL$Delay^2
WATEROIL$SF2 <- WATEROIL$Surfactant^2
WATEROIL$ST2 <- WATEROIL$SpanTriton^2
WATEROIL$SP2 <- WATEROIL$SolidPart^2

d)

I found VSL (Volume * Salinity) and VSF (Volume * Surfactant) as significant interaction terms. Yes, I did try all combinations of interaction terms. Also, I do not think it is an appropriate strategy because the workload is very large. As the number of independent terms increase, the number of interaction terms also increase.

R code:
WATEROIL$VSL <- WATEROIL$Volume * WATEROIL$Salinity
WATEROIL$VT <- WATEROIL$Volume * WATEROIL$Temperature
WATEROIL$VD <- WATEROIL$Volume * WATEROIL$Delay
WATEROIL$VSF <- WATEROIL$Volume * WATEROIL$Surfactant
WATEROIL$VST <- WATEROIL$Volume * WATEROIL$SpanTriton
WATEROIL$VSP <- WATEROIL$Volume * WATEROIL$SolidPart
WATEROIL$SLT <- WATEROIL$Salinity * WATEROIL$Temperature
WATEROIL$SLD <- WATEROIL$Salinity * WATEROIL$Delay

e)

```
Call:
lm(formula = Voltage ~ Volume + Salinity + Surfactant + V2 +
    VSL + VSF, data = WATEROIL)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54000 -0.09000  0.01333  0.12500  0.64000 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.0666667  0.1951827   5.465 0.000144 ***
Volume      -0.1155417  0.0232047  -4.979 0.000320 ***
Salinity     0.6400000  0.1781766   3.592 0.003700 ** 
Surfactant   1.1800000  0.2672650   4.415 0.000843 ***
V2           0.0012552  0.0003047   4.119 0.001423 ** 
VSL         -0.0078333  0.0028172  -2.781 0.016634 *  
VSF         -0.0120000  0.0042258  -2.840 0.014906 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3381 on 12 degrees of freedom
Multiple R-squared:  0.8671,	Adjusted R-squared:  0.8007 
F-statistic: 13.05 on 6 and 12 DF,  p-value: 0.0001211
```

The final model:
Voltage = 1.0666667 - 0.1155417 * Volume + 0.6400000 * Salinity + 1.1800000 * Surfactant + 0.0012552 * V2 - 0.0078333 * VSL - 0.0120000 * VSF + e

P.S.
 # V2: WATEROIL$Volume^2. A second-order term
 # VSL: WATEROIL$Volume * WATEROIL$Salinity. An interaction term
 # VSF: WATEROIL$Volume * WATEROIL$Surfactant. An interaction term

The value of adjusted R-squared (0.8007 or 80.07%) coefficient indicates the quantity of the variation in Voltage explained by the regression line. At this time, Adj-R2 (0.8007 or 80.07%) of the variation in Voltage is explained by Volume, Salinity and Surfactant. Adj-R2 (0.8007 or 80.07%) does not increase with the addition of a x-variable that does not improve the regression model. A higher Adj-R2 (0.8007 or 80.07%) typically indicates a better model.

T-Test:
Our p-value for the final model is 0.0001211, which is quite close to zero. Usually, a p-value with 0.05 or less is a good sign. Therefore, the small values of p-value for the intercept and slope indicates we can reject the null hypothesis and there is a relationship between Voltage (y) and Volume & Salinity & Surfactant (x-var).

F-Test:
Null hypothesis:
H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
Alternative hypothesis:
H_a: At least one coefficient $\beta_j \neq 0$
Test statistic:
F = 13.05
Therefore, F = 13.05 and with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between Voltage (y) and Volume & Salinity & Surfactant (x-var) is rejected. At least one x-variable has a significant effect on changes in balance. F-test gives strong support to the fitted model.