**Final Model**

```
Call:
lm(formula = Voltage ~ Volume + Salinity + Temperature + Surfactant +
    VSQ + Temperature * Salinity, data = WATEROIL)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63000 -0.18000  0.01333  0.16500  0.73000

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.067e+00  2.309e-01   4.620 0.000590 ***
Volume               -1.414e+03  4.060e+02  -3.482 0.004532 **
Salinity              3.974e-01  1.158e-01   3.431 0.004972 **
Temperature           2.684e-02  2.045e-02   1.312 0.213927
Surfactant            4.600e-01  9.996e-02   4.602 0.000609 ***
VSQ                   1.553e+05  5.213e+04   2.980 0.011491 *
Salinity:Temperature -1.684e-02  7.015e-03  -2.401 0.033468 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3999 on 12 degrees of freedom
Multiple R-squared:  0.8141,    Adjusted R-squared:  0.7212
F-statistic: 8.759 on 6 and 12 DF,  p-value: 0.0008173
```

In model 7 shown above we can first look at the f-test that states the p-value is .0008173. This tells us we can reject the null hypothesis, at least one of the betas is not equal to zero. If we then look at each o of the induvial betas in the t-test, we can see that all of them look ok (less than .05) except for Temperature. Despite Temperature being above .05, we will leave this in our model because out interaction variable of Salinity: Temperature passes the p-test and significantly increases our adjusted R-squared (as will be explained in a later paragraph). With our p-values passing the t-test we can assume we will then use our estimations that are also show in the model. The adjusted R-Squared is .7212, which means that 72.12% of Y is explained by our model. How the final model was created will be explained the analysis below.

In order to get to the final Model (shown in Model 7), there had to be analysis through R in order to figure out what specific variables were supposed to be kept in our model. First, we created all the second order variables that were possible in the data set such as the below:

```
WATEROIL$VSQ<- WATEROIL$Volume*WATEROIL$Volume
WATEROIL$SalSQ<- WATEROIL$Salinity*WATEROIL$Salinity
WATEROIL$TempSQ<-WATEROIL$Temperature*WATEROIL$Temperature
WATEROIL$DelSQ<-WATEROIL$Delay*WATEROIL$Delay
WATEROIL$SurfSQ<-WATEROIL$Surfactant*WATEROIL$Surfactant
WATEROIL$SpanSQ<- WATEROIL$SpanTriton*WATEROIL$SpanTriton
WATEROIL$SolSQ<- WATEROIL$SolidPart*WATEROIL$SolidPart
```

After this it was necessary to find what variables needed to be left in. This led us to our next step of performing backward elimination stepwise regression. In this process we started with our full model using the following formula:

**modl2<lm(Voltage~Volume+Salinity+Temperature+Delay+Surfactant+SpanTriton+SolidPart+VSQ+SalSQ+TempSQ+DelSQ+SurfSQ+SpanSQ+SolSQ, data = WATEROIL)**

Then we applied our backward regression stepwise formula:

**step<- stepAIC(modl2,direction = "backward")**

This led us to keep five of our independent variables as shown in the below:

**Voltage ~ Volume + Salinity + Temperature + Surfactant + VSQ**

These variables gave us the lowest possible AIC, meaning that with the given variables, the least amount of possible information is lost. The result of our backwards step elimination has now reduced our model that explains the data the best it can.

Given some domain knowledge and a little bit research (link in MLA format below), it was found out that Temperature and Salinity are somewhat dependent on each other. This led to our interaction variable that was placed in the final model (Temperature*Salinity). With adding this in our final model, we increased our adjusted R-squared from .619 to .7212. That means that variability in Y is explained nearly 10% greater by adding in the interaction variable.

"Home." MarineBio Conservation Society, 8 June 2019, https://marinebio.org/oceans/temperature/.