# Lsn 21

*Clark*

## Admin

Recall we previously looked a pisatcio bleaching where we had three values of air velocity and three values of drying temperature. Recall that we fit the model:

Why might we be less than satisfied by this model or this analysis?

As it turns out, there were three values for air velocity and three values of drying temperature.

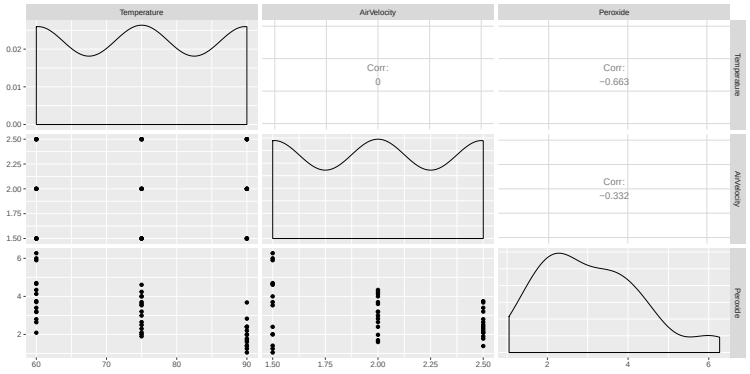
Let's think about designing this study a bit. Why might our researches have chosen 60, 75, and 90, as our temperatures and 1.5,2, and 2.5 as our air velocity values?

If our researchers had 45 pistachios to study, how should they allocate pistachios to temperature and air velocity? Why?

```
pistachio.raw<-read.table("http://www.isi-stats.com/isi2/data/pistachioStudy.txt",header=T)
pistachio.dat <- pistachio.raw %>% mutate(Peroxide=Peroxide..remaining.)%>%
  select(-Peroxide..remaining.)
pistachio.dat<-pistachio.dat %>% select(Temperature,AirVelocity,Peroxide)
```

One way to explore the data is to look at pairs plots

```
library(GGally)
pistachio.dat %>% ggpairs()
```




What is this output showing us?

```
library(rgl)
#x=pistachio.dat$Temperature
#y=pistachio.dat$AirVelocity
#z=pistachio.dat$Peroxide
#plot3d(x,y,z)
```

Note that this might lead to the model:

To fit this, we can fit:

```
pistachio.lm<-lm(Peroxide~AirVelocity+Temperature,data=pistachio.dat)
summary(pistachio.lm)
```

```
##
## Call:
## lm(formula = Peroxide ~ AirVelocity + Temperature, data = pistachio.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57222 -0.45822 -0.05822  0.52978  2.14478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.23322    1.03126   9.923 1.41e-12 ***
## AirVelocity -1.02400    0.31952  -3.205  0.00258 **
## Temperature -0.06820    0.01065  -6.403 1.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.8751 on 42 degrees of freedom
## Multiple R-squared:  0.5497, Adjusted R-squared:  0.5283
## F-statistic: 25.64 on 2 and 42 DF,  p-value: 5.289e-08
```

So our fitted model is:

Our conclusion based on the p values is:

Let's take a look at the univariate models:

```
uni1.lm<-lm(Peroxide~Temperature,data=pistachio.dat)
summary(uni1.lm)
```

```
##
## Call:
## lm(formula = Peroxide ~ Temperature, data = pistachio.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00322 -0.69322 -0.06722  0.56678  2.17678
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.18522    0.89239   9.172 1.11e-11 ***
## Temperature -0.06820    0.01174  -5.808 6.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9648 on 43 degrees of freedom
## Multiple R-squared:  0.4396, Adjusted R-squared:  0.4266
## F-statistic: 33.73 on 1 and 43 DF,  p-value: 6.957e-07
```

What do we note about the value of temperature? Is this surprising?

What about the SE? Why might this be?

```
uni2.lm<-lm(Peroxide~AirVelocity,data=pistachio.dat)
summary(uni2.lm)
```

```
##
```

```
## Call:
## lm(formula = Peroxide ~ AirVelocity, data = pistachio.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.53222 -0.67022 -0.05222  0.92978  2.68778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.1182     0.9062   5.648 1.19e-06 ***
## AirVelocity  -1.0240     0.4439  -2.307    0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.216 on 43 degrees of freedom
## Multiple R-squared:  0.1101, Adjusted R-squared:  0.08942
## F-statistic: 5.321 on 1 and 43 DF,  p-value: 0.02596
```

What is the relationship between $R^2$ in the full model and $R^2$ in our two univariate models? Why might this be?

If we look at $R^2$ of the two univariate models, which covariate explains more variation, Temperature or Air Velocity?

Looking at $\hat{\beta}_1$ and $\hat{\beta}_2$ which value is bigger? Is this surprising in light of what we found when looking at $R^2$ values?

The biggest issue here is we are on different scales. In order to get rid of scaling we standardize, which is:

When we have standardized coefficients we can compare values to determine which one has a larger effect.

```
pistachio.std<-pistachio.raw %>%mutate(Peroxide=Peroxide..remaining.)%>%
  select(std.temp,std.air,std.peroxide,Peroxide)
std.lm<-lm(Peroxide~std.temp+std.air,data=pistachio.std)
summary(std.lm)
```

```
##
## Call:
## lm(formula = Peroxide ~ std.temp + std.air, data = pistachio.std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57222 -0.45822 -0.05822  0.52978  2.14478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0702     0.1304  23.537  < 2e-16 ***
## std.temp     -0.8447     0.1319  -6.403 1.04e-07 ***
```

```
## std.air      -0.4228     0.1319  -3.205  0.00258 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8751 on 42 degrees of freedom
## Multiple R-squared:  0.5497, Adjusted R-squared:  0.5283
## F-statistic: 25.64 on 2 and 42 DF,  p-value: 5.289e-08
```

Note here our intercept actually means something.

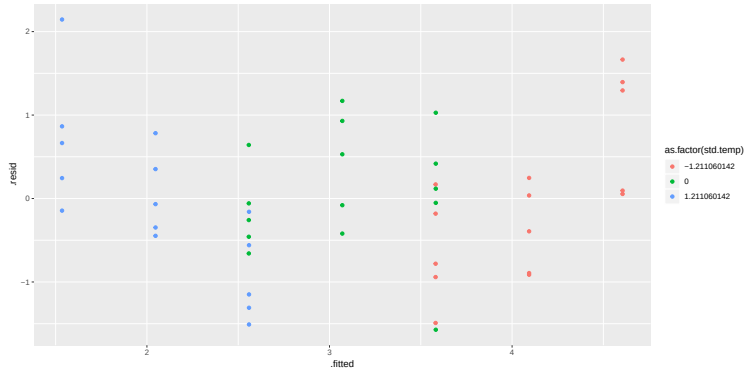
Let's see this:

```
#install.packages("p3d", repos="http://R-Forge.R-project.org")
#library(p3d)
#Plot3d(std.lm)
```

Here we note that our response surface is a plane.

To see if this is a good model, we can examine the residuals:

```
std.lm %>%ggplot(aes(x=.fitted,y=.resid,color=as.factor(std.temp)))+geom_point()
```




Do we have any concerns? What happens when temp is high?

A model with an interaction term:

For a fixed $X_2$ what happens as we change $X_1$ by 1 unit?

Conversely, for a fixed $X_1$ what happens when we change $X_2$ by 1 unit?

```r
inter.lm<-lm(Peroxide~std.temp*std.air,data=pistachio.std)
summary(inter.lm)
```

```
##
## Call:
## lm(formula = Peroxide ~ std.temp * std.air, data = pistachio.std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57222 -0.45822 -0.05822  0.52978  1.33528
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.0702     0.1016  30.211  < 2e-16 ***
## std.temp          -0.8447     0.1028  -8.219 3.32e-10 ***
## std.air           -0.4228     0.1028  -4.114 0.000183 ***
## std.temp:std.air   0.5519     0.1039   5.310 4.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6817 on 41 degrees of freedom
## Multiple R-squared:  0.7332, Adjusted R-squared:  0.7137
## F-statistic: 37.56 on 3 and 41 DF,  p-value: 7.743e-12
```

```r
#Plot3d(inter.lm)
```

What's going on here? What are we testing?

Here we see that $R^2$ is still additive

Note that our $\mathbf{X}$ matrix becomes:

Which is nested with the $\mathbf{X}$ matrix without an interaction, so we can test whether a model with an interaction is preferable through:

```r
anova(std.lm,inter.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Peroxide ~ std.temp + std.air
## Model 2: Peroxide ~ std.temp * std.air
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     42 32.160
## 2     41 19.054  1    13.106 28.2 4.112e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```