

STA 674 Assignment #5

Matt Massey

Question 1.

```
setwd("C:/Users/Melissa/Dropbox (Personal)/STA 674/code for homework/Assignment 5")
```

```
library(emmeans)
```

```
pest = read.csv("pesticides.csv", header = T)
```

```
KILLED = pest$Killed/100
```

```
# the following specifications are required
```

```
PESTICIDE = as.factor(pest$Pesticide)
```

```
PEST = data.frame(PESTICIDE, KILLED)
```

```
out = lm(KILLED ~ PESTICIDE)
```

```
summary(out)
```

```
##
```

```
## Call:
```

```
## lm(formula = KILLED ~ PESTICIDE)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -0.09000 -0.03000 -0.01667  0.01667  0.12000
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.46667      0.04047   11.531 4.25e-07 ***
```

```
## PESTICIDE2    0.09333      0.05723    1.631  0.1340
```

```
## PESTICIDE3    0.36333      0.05723    6.348 8.37e-05 ***
```

```
## PESTICIDE4    0.18000      0.05723    3.145  0.0104 *
```

```
## PESTICIDE5    0.37333      0.05723    6.523 6.70e-05 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 0.0701 on 10 degrees of freedom
```

```
## Multiple R-squared:  0.8688, Adjusted R-squared:  0.8164
```

```
## F-statistic: 16.56 on 4 and 10 DF, p-value: 0.0002075
```

```
#lsmeans and confidence interval without correction
```

```
lsmeans(results, "PESTICIDE")
```

```
## PESTICIDE lsmean      SE df lower.CL upper.CL
```

```
## 1          0.467 0.0405 10    0.376    0.557
```

```
## 2          0.560 0.0405 10    0.470    0.650
```

```
## 3          0.830 0.0405 10    0.740    0.920
```

```
## 4          0.647 0.0405 10    0.556    0.737
```

```
## 5          0.840 0.0405 10    0.750    0.930
```

```
##
```

```
## Confidence level used: 0.95
```

```
#lsmeans and confidence interval with Bonferroni correction
```

```
lsmeans(results, "PESTICIDE", level=(1-.05/5))
```

```
## PESTICIDE lsmean      SE df lower.CL upper.CL
```

```
## 1          0.467 0.0405 10    0.338    0.595
```

```
## 2          0.560 0.0405 10    0.432    0.688
```

```
## 3          0.830 0.0405 10    0.702    0.958
```

```
## 4      0.647 0.0405 10      0.518      0.775
## 5      0.840 0.0405 10      0.712      0.968
##
## Confidence level used: 0.99

library(DescTools)
## Warning: package 'DescTools' was built under R version 4.0.5

#ANOVA table, lsmeans, and Multiple Comparisons using Tukey
results = aov(KILLED ~ PESTICIDE, data=PEST)
summary(results)
##              Df Sum Sq Mean Sq F value    Pr(>F)
## PESTICIDE      4  0.3254  0.08136    16.56 0.000208 ***
## Residuals     10  0.0491  0.00491
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(results)
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = KILLED ~ PESTICIDE, data = PEST)
##
## $PESTICIDE
##              diff              lwr              upr              p adj
## 2-1  0.09333333 -0.095023428  0.281690094  0.5120660
## 3-1  0.36333333  0.174976572  0.551690094  0.0006182
## 4-1  0.18000000 -0.008356761  0.368356761  0.0626996
## 5-1  0.37333333  0.184976572  0.561690094  0.0004966
## 3-2  0.27000000  0.081643239  0.458356761  0.0057075
## 4-2  0.08666667 -0.101690094  0.275023428  0.5765402
## 5-2  0.28000000  0.091643239  0.468356761  0.0044297
## 4-3 -0.18333333 -0.371690094  0.005023428  0.0572923
## 5-3  0.01000000 -0.178356761  0.198356761  0.9997516
## 5-4  0.19333333  0.004976572  0.381690094  0.0436847
```

1A.

The cell means model for the pesticide study is:

$$y_{ij} = \mu_i + e_{ij}$$

where:

- $i =$ level i of factor 1; type of pesticide is the only factor in this experiment and it has 5 levels
- $j =$ numbered experimental unit; each pesticide is applied to three different mosquito tanks
- $y_{ij} =$ observed response for the j^{th} experimental unit assigned to the i^{th} treatment; percent of mosquitos killed in each tank
- $\mu_i =$ mean response for the i^{th} treatment; mean percent of mosquitos killed for each pesticide
- $e_{ij} =$ error for j^{th} experimental unit assigned to i^{th} treatment and we assume $e_{ij} \sim N(0, \sigma^2)$, independent (errors are normally distributed, have mean zero, constant variance, and independent); difference between observed percent of mosquitos killed in a tank from the mean percent of mosquitos killed in the tank for a particular pesticide

1B.

The estimated kill rates and 95% confidence intervals without adjustment for each pesticide are:

<u>Treatment</u>	<u>mean</u>	<u>95% CI</u>	
Pesticide 1 (control)	0.467	0.376	0.557
Pesticide 2	0.560	0.470	0.650
Pesticide 3	0.830	0.740	0.920
Pesticide 4	0.647	0.556	0.737
Pesticide 5	0.840	0.750	0.930

In the experiment we have five treatments of pesticide, which means we have five confidence intervals and the overall coverage probability would be $0.773781 (0.95^5)$. So, even though we have a 95% confidence interval on each single comparison, the overall experiment-wise confidence interval drops to a much lower 77% confidence interval, and would not likely meet the significance level specifications for an experiment.

1C.

The simultaneous confidence intervals for the kill rates of each pesticide using Bonferroni correction to maintain experiment-wise error rate of 5% are:

<u>Treatment</u>	<u>99% CI ($\alpha_c = 0.05$)</u>	
Pesticide 1 (control)	0.338	0.595
Pesticide 2	0.432	0.688
Pesticide 3	0.702	0.958
Pesticide 4	0.518	0.775
Pesticide 5	0.712	0.968

We see that these confidence intervals cover a wider range of kill rates as compared to the confidence intervals calculated in 1B without correction. The Bonferroni adjustment used here requires a 5% probability of committing a type 1 error in all tests of the experiment, rather than each individual test as in question 1B. The Bonferroni adjustment achieves this with the equation:

$$\alpha_c = \alpha_e / n$$

where,

α_c = comparison-wise error rate

α_e = experiment-wise error rate

n = number of tests

In this case, to ensure a 95% confidence in the results of the experiment, we need to calculate the confidence intervals of each comparison at a 0.01 significance level ($\alpha_c = 0.05 / 5 = 0.01$). This results in much wider confidence intervals with the correction.

1D.

The calculated estimates for each pesticide suggest that pesticides 3 and 5 have the highest kill rates at 83% and 84%, respectively.

If we use Bonferroni's correction to maintain an experiment-wise error of $\alpha_e = 0.05$, we see that the confidence intervals of pesticides 3 and 5 are above pesticides 1 and 2, although there is minor overlap with pesticide 4.

Tukey's honest significant differences method shows that there is no significant difference between pesticides 3 and 4 with a P value of 0.0572923, although it is close. We do see a significant difference between pesticides 4 and 5, which has a P value of 0.0436847. Comparison of pesticides 3 and 5, however, show no significant difference with a P value of 0.9997516.

Therefore, we conclude that pesticides 3 and 5 have the highest kill rate, but we do not see a significant difference between these two pesticides.

Question 2.

```
kuehl = read.csv("kuehl_6_6.csv", header = T)

library(multcomp)

## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##      geyser

PROPORTION = as.factor(kuehl$PROPORTION)
SURFACE     = as.factor(kuehl$SURFACE)
FILLER      = as.factor(kuehl$FILLER)
LOSS        = kuehl$LOSS
KUEHL = data.frame(PROPORTION, SURFACE, FILLER, LOSS)

#ANOVA table
result.anova = aov(LOSS ~ PROPORTION*SURFACE*FILLER, data=KUEHL)
summary(result.anova)

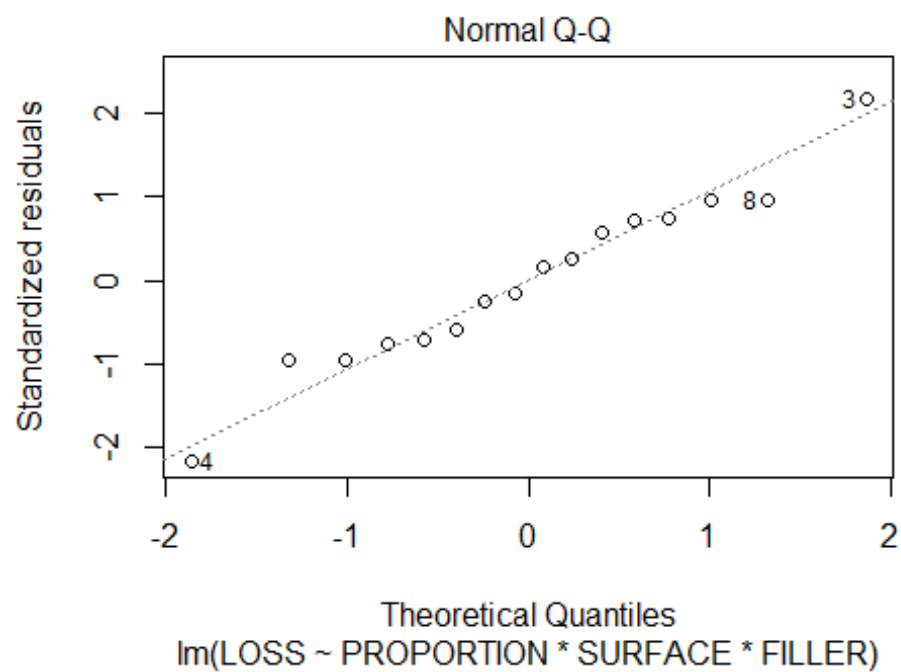
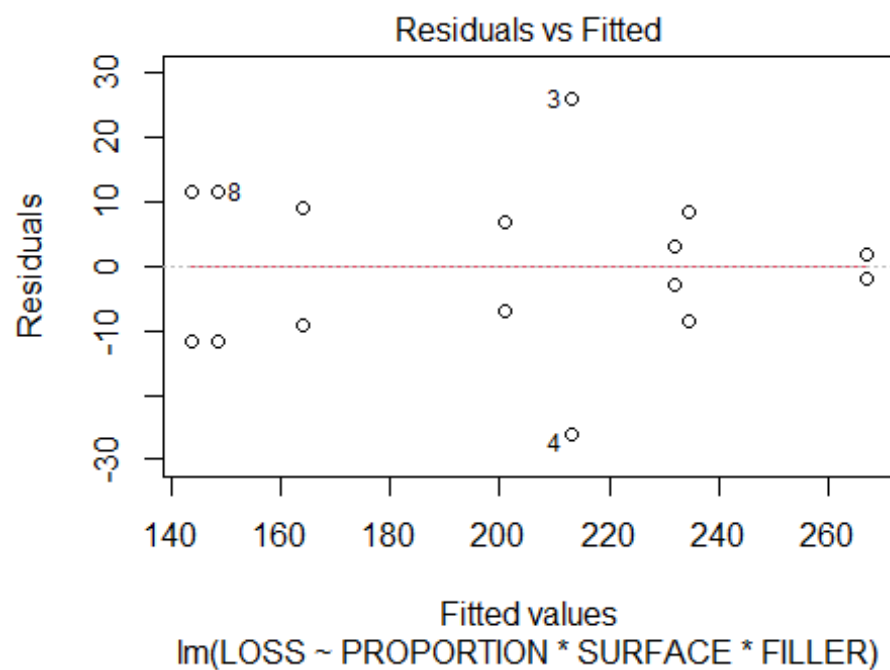
##              Df Sum Sq Mean Sq F value    Pr(>F)
## PROPORTION      1   5663    5663  19.598 0.002206 **
## SURFACE          1  12939   12939  44.782 0.000154 ***
## FILLER           1   3875    3875  13.411 0.006381 **
## PROPORTION:SURFACE 1    150     150   0.519 0.491634
## PROPORTION:FILLER  1   3452    3452  11.946 0.008615 **
## SURFACE:FILLER     1   1743    1743   6.033 0.039556 *
## PROPORTION:SURFACE:FILLER 1    203     203   0.703 0.426184
## Residuals        8   2311     289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

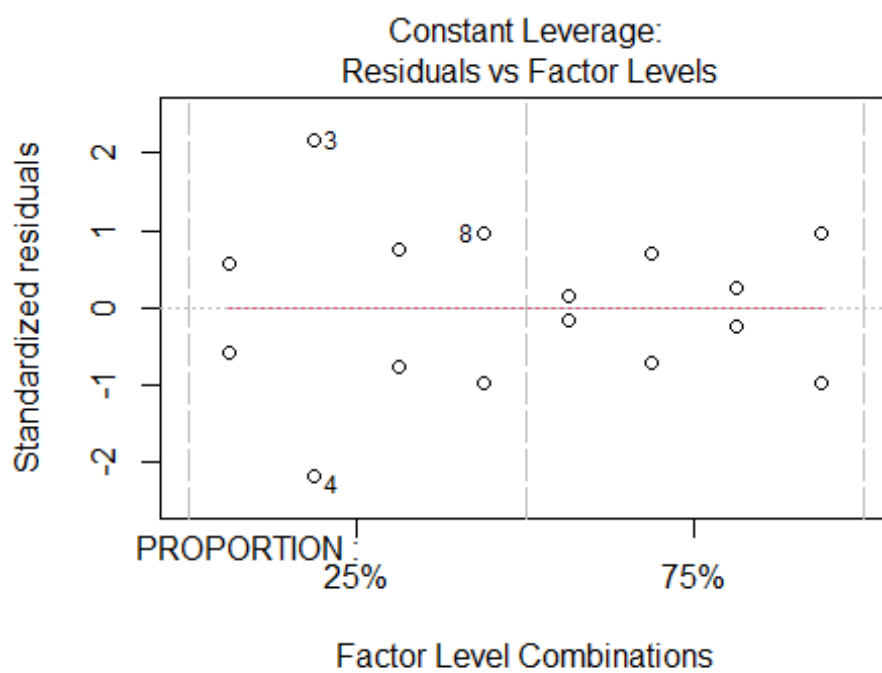
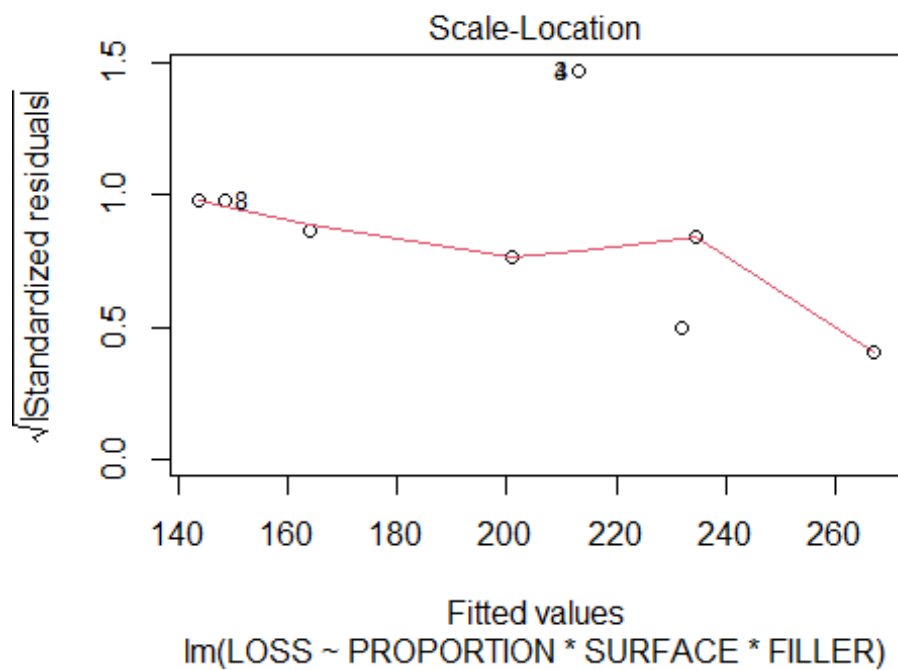
#LSMeans and confidence intervals (the emmean column is LSMeans)
result_emm = emmeans(result_rg1, ~(PROPORTION*SURFACE*FILLER))
confint(result_emm, level = (1 - 0.05/8))

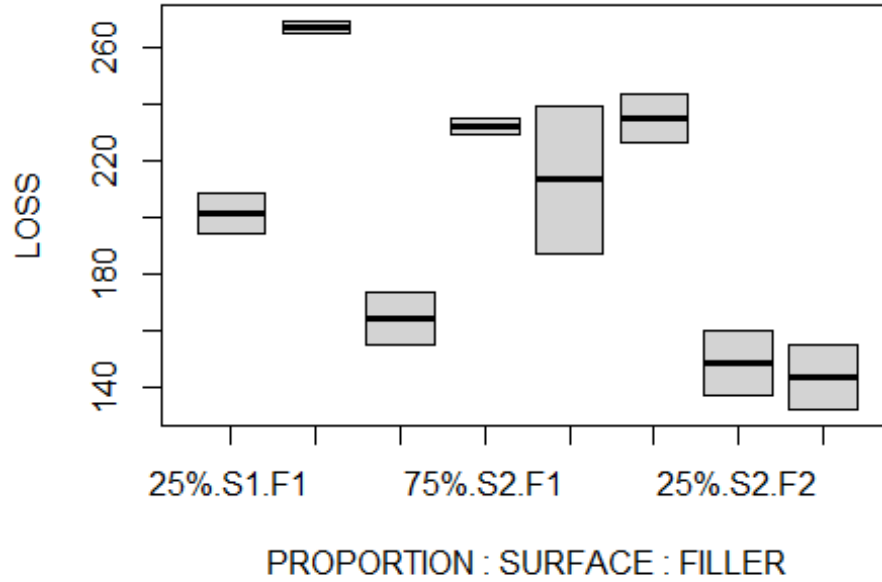
## PROPORTION SURFACE FILLER emmean SE df lower.CL upper.CL
## 25%         S1      F1      201 12 8    156.8    245
## 75%         S1      F1      267 12 8    222.8    311
## 25%         S2      F1      164 12 8    119.8    208
## 75%         S2      F1      232 12 8    187.8    276
## 25%         S1      F2      213 12 8    168.8    257
## 75%         S1      F2      234 12 8    190.3    279
## 25%         S2      F2      148 12 8    104.3    193
## 75%         S2      F2      144 12 8     99.3    188
##
## Confidence level used: 0.99375

###check assumptions
result_lm = lm(LOSS ~ PROPORTION*SURFACE*FILLER)
result_rg1 = ref_grid(result_lm)

#summary(result_lm)
plot(result_lm)
boxplot(LOSS ~ PROPORTION*SURFACE*FILLER)
```







2A.

The cell means model for the fabric study is:

$$y_{ijkl} = \mu_{ijk} + e_{ijkl}$$

where:

- $i =$ level i of factor 1; there are three factors for this experiment, type of filler, proportion of filler, and surface treatment, each having 2 levels; assume factor 1 is type of filler and there are two types of filler (2 levels)
- $j =$ level j of factor 2; there are three factors for this experiment, type of filler, proportion of filler, and surface treatment, each having 2 levels; assume factor 2 is proportion of filler, and there are two proportions at 25% and 75% (2 levels)
- $k =$ level k of factor 3; there are three factors for this experiment, type of filler, proportion of filler, and surface treatment, each having 2 levels; assume factor 3 is surface treatment and it has two types, S1 and S2 (2 levels)
- $l =$ numbered experimental units; there are 2 observations (experimental units) for each treatment in this experiment
- $\mu_{ijk} =$ mean response of units treated with level i of factor 1, level j of factor 2, and level k of factor 3; mean weight loss in mg for each set of treatments
- $y_{ijkl} =$ observed response for the l^{th} experimental unit treated with level i of factor 1, level j of factor 2, and level k of factor 3; actual weight loss in mg for a one observation of a treatment
- $e_{ijkl} =$ error for the l^{th} experimental unit treated with level i of factor 1, level j of factor 2, and level k of factor 3 ($y_{ijkl} - \mu_{ijk}$)

2B.

The least squares estimate of the mean weight loss after each abrasion test for each treatment, and using the Bonferroni correction to maintain an experiment-wise error rate of 0.05 are:

PROPORTION	SURFACE	FILLER	Mean	CI (lower)	CI (upper)
25%	S1	F1	201	156.8	245
75%	S1	F1	267	222.8	311
25%	S2	F1	164	119.8	208
75%	S2	F1	232	187.8	276
25%	S1	F2	213	168.8	257
75%	S1	F2	234	190.3	279
25%	S2	F2	148	104.3	193
75%	S2	F2	144	99.3	188

2C.

The assumptions for the cell means model here are: (1) the errors are normally distributed, (2) the errors have a mean of zero, (3) the errors have constant variance, and (4) the errors are independent.

In this experiment, the assumptions are almost met, as discussed below.

The errors are approximately normally distributed as can be seen by the Residuals vs. Fitted plots above. The Normal QQ plot does show a weak sigmoidal pattern of the data, which deviates from a straight line, however, I don't know the point at which a weak pattern indicates non-normality and seems subjective. Perhaps a more quantitative approach to assessing normality would be more appropriate here. If it is not-normal, then the dataset needs reviewed and possibly a data transformation applied.

The errors have a mean of approximately zero, as can be seen by the Residuals vs. Fitted plot.

The errors have a constant variance as be seen by the Residual and Standardized Residual plots.

Finally, the errors are independent, as they are not correlated or show any association with each other on any of the plots.

There are a few potential outliers, as can be seen in several of the plots above, and may need to be examined further.