STAT850 HW7

Stewart Kerr

March 11, 2019

Problem 1

a,b. See attached page.

c.

```
#Problem 1c
#Load the data in
breadvolume <- read.csv("~/2019spring/STAT850/hw7/breadvolume.csv")
breadvolume$fat = factor(breadvolume$fat)
breadvolume$surfactant = factor(breadvolume$surfactant)
breadvolume$day = factor(breadvolume$day)

#Build model and do ANOVA
bv_lm <- lm(volume ~ fat*surfactant + day, data = breadvolume)
anova(bv_lm)
```

```
## Analysis of Variance Table
##
## Response: volume
##                Df Sum Sq Mean Sq F value    Pr(>F)
## fat             2 7.4526  3.7263 19.5738 0.0002378 ***
## surfactant      2 0.2972  0.1486  0.7807 0.4819115
## day             3 7.0939  2.3646 12.4212 0.0007443 ***
## fat:surfactant  2 5.4002  2.7001 14.1834 0.0009004 ***
## Residuals      11 2.0941  0.1904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table, I confirm that my answer to part b is correct. Also, with a F-statistic of 12.42 and a p-value of 0.007, there is strong evidence to suggest that there are differences based on days. I will now examine the effect of days using the `drop1` function.

```
#Problem 1c
#Compare to drop1
drop1(bv_lm, test = "F")
```

```
## Single term deletions
##
## Model:
## volume ~ fat * surfactant + day
##                Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                      2.0941 -28.4135
## day             3    7.7726 9.8667  -1.8626  13.610 0.0005071 ***
## fat:surfactant  2    5.4002 7.4943  -5.6379  14.183 0.0009004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we got a slightly different F-statistic for day with a slightly larger p-value. The reason for this is that `ANOVA()` gives us type-I sum of squares while `drop1()` gives us type-III sum of squares. In the case of an

unbalanced experiment, using type-I sum of squares for the F-test is not really useful because our results are dependent on which order we defined our factors in the model. Thus, the results from the `drop1()` function are correct for this unbalanced case.

```
#Problem 1c
#Look at interaction coefficients
summary(bv_lm)
```

```
##
## Call:
## lm(formula = volume ~ fat * surfactant + day, data = breadvolume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5668 -0.2084 -0.1451  0.2084  0.5499
##
## Coefficients: (2 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.2080     0.3014  20.595  3.9e-10 ***
## fat2                1.4811     0.3713   3.989 0.002124 **
## fat3                1.1004     0.4058   2.712 0.020234 *
## surfactant2         0.3421     0.3706   0.923 0.375699
## surfactant3         1.9535     0.4585   4.260 0.001342 **
## day2               -1.6168     0.2793  -5.789 0.000121 ***
## day3               -0.3071     0.2820  -1.089 0.299440
## day4               -0.7432     0.3134  -2.371 0.037060 *
## fat2:surfactant2       NA         NA      NA       NA
## fat3:surfactant2    0.2162     0.5247   0.412 0.688134
## fat2:surfactant3   -2.9758     0.5823  -5.110 0.000338 ***
## fat3:surfactant3       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4363 on 11 degrees of freedom
## Multiple R-squared:  0.9063, Adjusted R-squared:  0.8296
## F-statistic: 11.82 on 9 and 11 DF,  p-value: 0.00018
```

In both the `anova()` and `drop1()` output, the sum of squares for the fat:surfactant interaction is equal. This is because the interaction term is always added to the model *after* the main effects of fat and surfactant. Also, when looking at `summary(bv_lm)` output, we notice that a few of the interaction coefficients cannot be estimated. Again, this is because there is data that is missing. To quantity the strength of evidence of an interaction between fat and surfactant, I will return to the `drop1()` output above. In the row fat:surfactant, we are performing an F-test comparing the reduced model without any interaction terms (in other words, all interaction coefficients are zero) vs. the full model which has at least one significant (non-zero) interaction term. From the R output, we find an F-statistic of 14.18 corresponding to a p-value of 0.009. This provides strong evidence that there is an interaction between fat and surfactant.

d.

```
#Problem 1d
#change the baseline
breadvolume$fat = factor(breadvolume$fat, levels = c("3","2","1"))
bv_contrasts = contrasts(breadvolume$fat)

#Now repeat c
bv_lm2 <- lm(volume ~ fat*surfactant + day, data = breadvolume, contrasts = bv_contrasts)
```

```r
drop1(bv_lm2, test = "F")
```

```
## Single term deletions
##
## Model:
## volume ~ fat * surfactant + day
##                Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                      2.0941 -28.4135
## day             3    7.7726 9.8667  -1.8626  13.610 0.0005071 ***
## fat:surfactant  2    5.4002 7.4943  -5.6379  14.183 0.0009004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

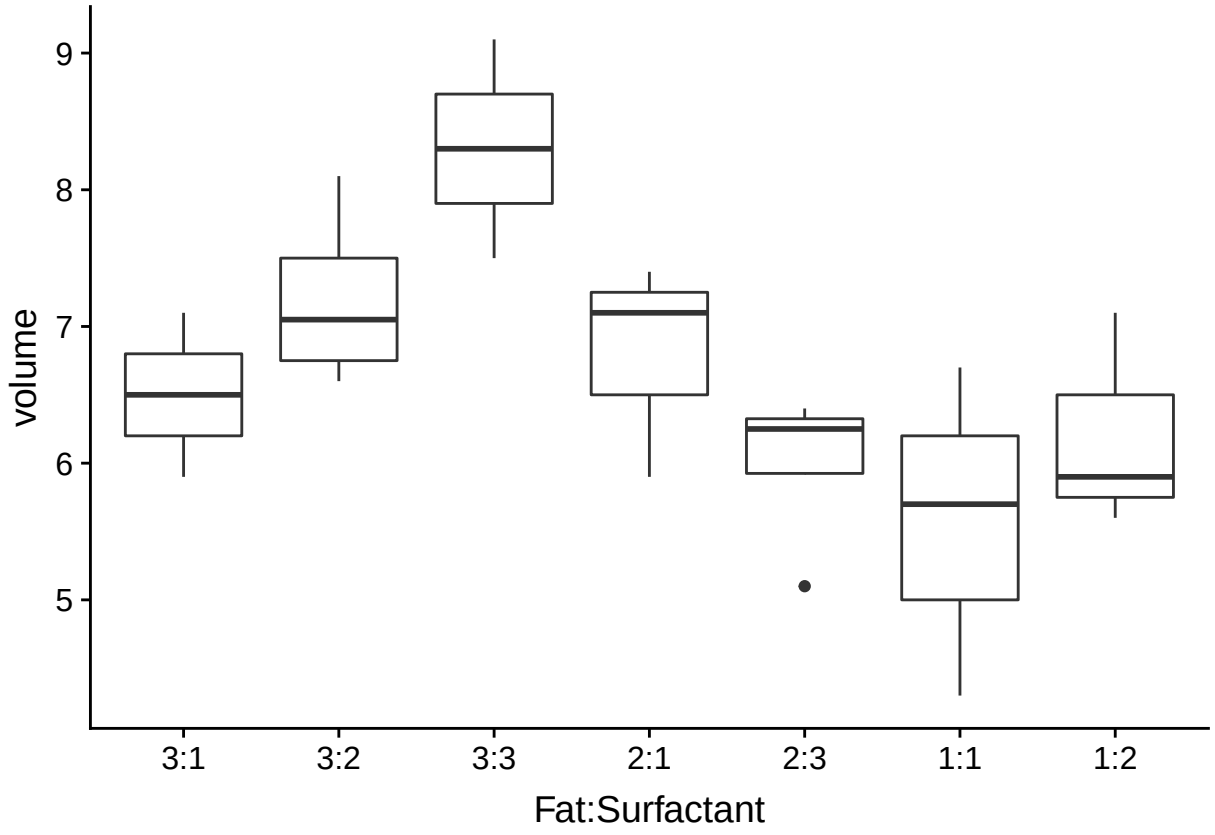
```r
summary(bv_lm2)
```

```
##
## Call:
## lm(formula = volume ~ fat * surfactant + day, data = breadvolume,
##     contrasts = bv_contrasts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5668 -0.2084 -0.1451  0.2084  0.5499
##
## Coefficients: (2 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.3084     0.3387  21.581 2.36e-10 ***
## fat2               0.3806     0.4257   0.894 0.390401
## fat1              -1.1004     0.4058  -2.712 0.020234 *
## surfactant2        0.5584     0.3921   1.424 0.182213
## surfactant3        1.9535     0.4585   4.260 0.001342 **
## day2              -1.6168     0.2793  -5.789 0.000121 ***
## day3              -0.3071     0.2820  -1.089 0.299440
## day4              -0.7432     0.3134  -2.371 0.037060 *
## fat2:surfactant2       NA         NA      NA       NA
## fat1:surfactant2  -0.2162     0.5247  -0.412 0.688134
## fat2:surfactant3  -2.9758     0.5823  -5.110 0.000338 ***
## fat1:surfactant3       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4363 on 11 degrees of freedom
## Multiple R-squared:  0.9063, Adjusted R-squared:  0.8296
## F-statistic: 11.82 on 9 and 11 DF,  p-value: 0.00018
```

In this problem I have made fat 3 the baseline and repeated my analysis from c. The differences are that the coefficients for fat have changed (as expected) and the coefficients for the interaction term have switched. It's probably best to set the baseline fat level to 3 because we observe all surfactants for fat 3 while for the other fat levels we have missing surfactant data.

e. For this problem, I will first consider boxplots of fat:surfactant combinations across the different days.

```r
#Problem 1e
breadvolume$fs = breadvolume$fat:breadvolume$surfactant
qplot(fs, volume, geom = "boxplot", data = breadvolume, xlab = "Fat:Surfactant")
```



From the boxplots, it appears that fat 3 and surfactant 3 is best. I will use Tukey's HSD to determine if the difference is statistically significant.

```
#Problem 1e
bv_aov <- aov(volume ~ fat:surfactant + day, data = breadvolume)
TukeyHSD(bv_aov, "fat:surfactant", ordered = TRUE)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = volume ~ fat:surfactant + day, data = breadvolume)
##
## $`fat:surfactant`
##               diff         lwr      upr     p adj
## 1:2-1:1 0.44444444 -0.90404789 1.792937 0.9277881
## 2:3-1:1 0.47666667 -0.78473241 1.738066 0.8645695
## 3:1-1:1 1.15500000 -0.35266027 2.662660 0.1884832
## 2:1-1:1 1.39000000  0.04150766 2.738492 0.0417813
## 3:2-1:1 1.67666667  0.41526759 2.938066 0.0074905
## 3:3-1:1 2.88166667  1.37400640 4.389327 0.0003693
## 2:2-1:1         NA          NA       NA        NA
## 1:3-1:1         NA          NA       NA        NA
## 2:3-1:2 0.03222222 -1.22917686 1.293621 1.0000000
## 3:1-1:2 0.71055556 -0.79710471 2.218216 0.6913035
## 2:1-1:2 0.94555556 -0.40293678 2.294048 0.2647293
## 3:2-1:2 1.23222222 -0.02917686 2.493621 0.0572104
```

```
## 3:3-1:2 2.43722222  0.92956196 3.944882 0.0015860
## 2:2-1:2         NA          NA       NA        NA
## 1:3-1:2         NA          NA       NA        NA
## 3:1-2:3 0.67833333 -0.75195878 2.108625 0.6851392
## 2:1-2:3 0.91333333 -0.34806574 2.174732 0.2354017
## 3:2-2:3 1.20000000  0.03217138 2.367829 0.0425772
## 3:3-2:3 2.40500000  0.97470789 3.835292 0.0011370
## 2:2-2:3         NA          NA       NA        NA
## 1:3-2:3         NA          NA       NA        NA
## 2:1-3:1 0.23500000 -1.27266027 1.742660 0.9992947
## 3:2-3:1 0.52166667 -0.90862545 1.951959 0.8839759
## 3:3-3:1 1.72666667  0.07510759 3.378226 0.0383463
## 2:2-3:1         NA          NA       NA        NA
## 1:3-3:1         NA          NA       NA        NA
## 3:2-2:1 0.28666667 -0.97473241 1.548066 0.9910995
## 3:3-2:1 1.49166667 -0.01599360 2.999327 0.0531883
## 2:2-2:1         NA          NA       NA        NA
## 1:3-2:1         NA          NA       NA        NA
## 3:3-3:2 1.20500000 -0.22529211 2.635292 0.1237275
## 2:2-3:2         NA          NA       NA        NA
## 1:3-3:2         NA          NA       NA        NA
## 2:2-3:3         NA          NA       NA        NA
## 1:3-3:3         NA          NA       NA        NA
## 1:3-2:2         NA          NA       NA        NA
```

If we focus on just the 3:3-x:x differences above, *most* of them are significant - thus I would still highly recommend the 3:3 treatment above all others. The 3:3-3:2 difference may not be significant, so fat level 3 and surfactant level 2 might be a good suggestion as well. It's worth noting that we didn't observe fat 1 with surfactant 3 or fat 2 with surfactant 2 - it's possible that those interactions would yield higher average volume than the 3:3 treatment.

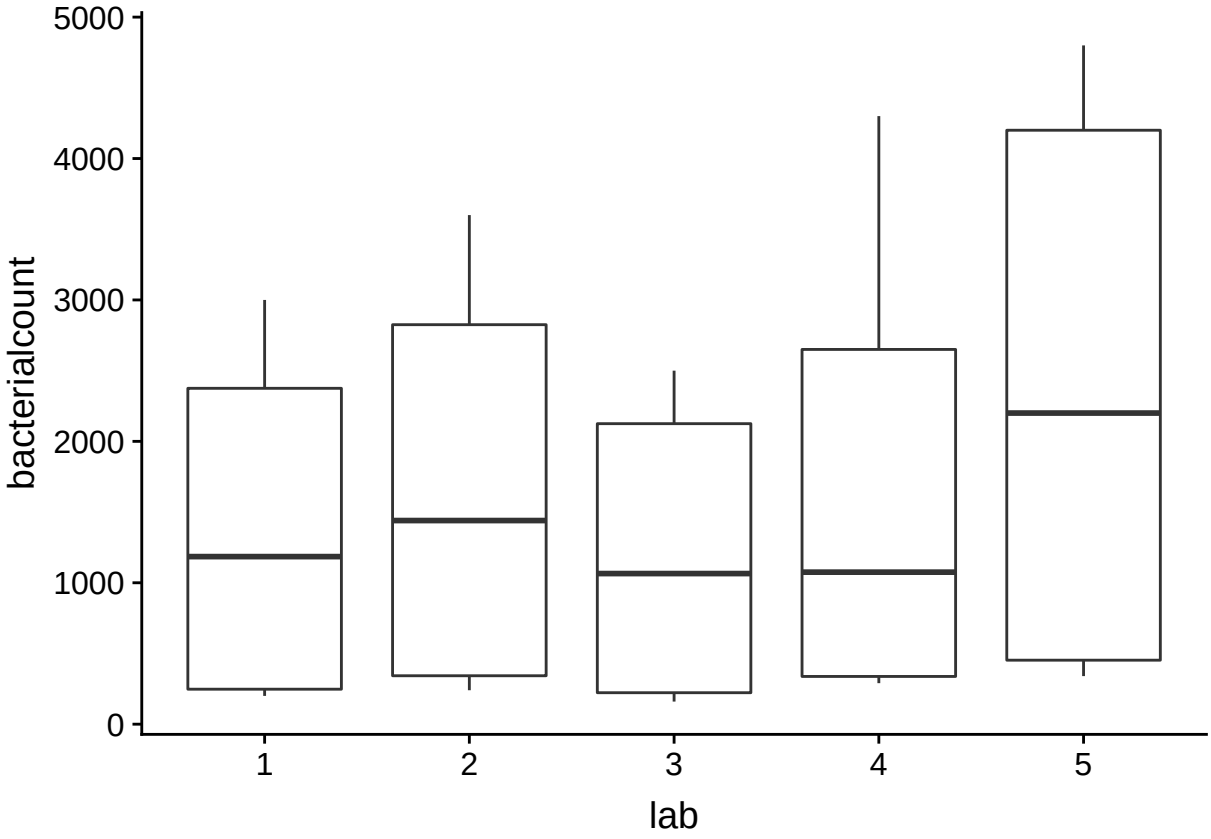
Problem 2

a. See attached sheet.

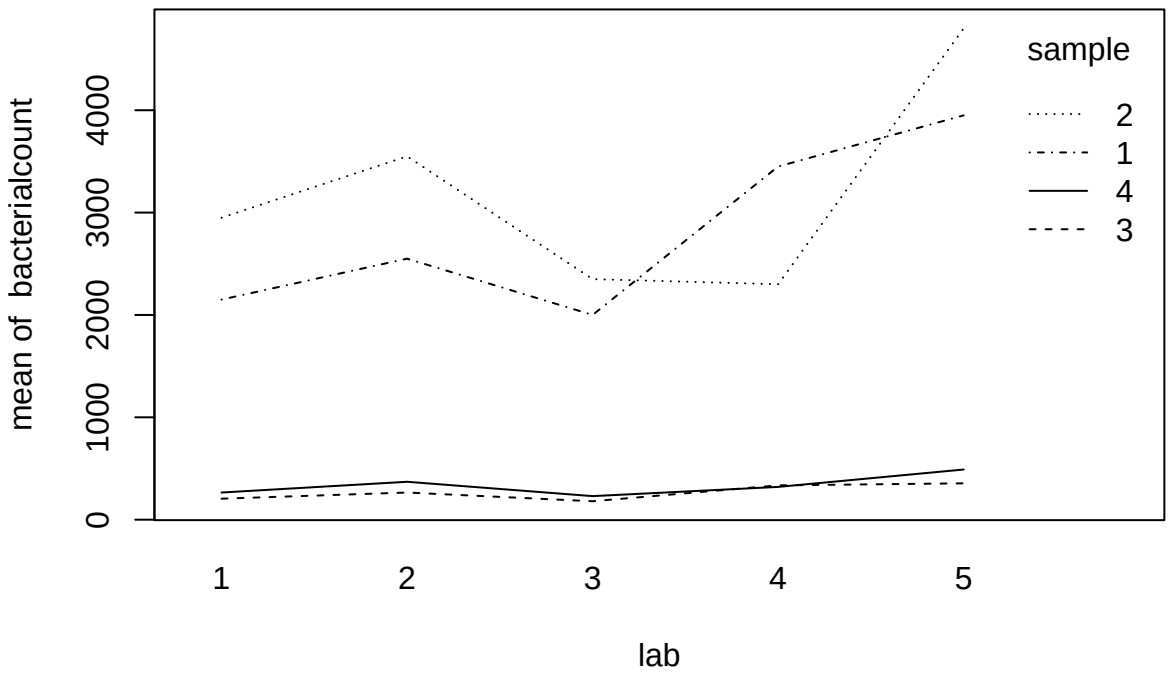
b. For this analysis, we are primarily interested in two things: differences between laboratories (i.e. the main effect of laboratory) and the interaction between laboratory and concentration of bacteria (which comes from sampling milk at different stages of spoilage). From the model I've written down for a, we have that all factors in this analysis are random effects. Below, I have fit the model using `lmer()` and can get the variances of each random effect from `summary(milk_lmer)`. I will use these variances to calculate F-statistics and perform tests. First, however, I will visualize the data.

```
#Problem 2b
#Load the data in
milk <- read.csv("~/2019spring/STAT850/hw7/milkcontamination.csv")
milk$lab = factor(milk$lab)
milk$sample = factor(milk$sample)

#Visualize data
qplot(lab,bacterialcount, geom = "boxplot", data = milk)
```




```
with(milk, interaction.plot(lab,sample,bacterialcount))
```



From the boxplots, it's not readily apparent that one lab significantly differs in their bacterial count estimations from the other labs. Laboratory 5 *might* estimate a higher bacterial count on average, but it's not clearly different. From the interaction plot, we observe that all labs perform similarly on samples 3 and 4, but they perform differently on samples 1 and 2. The lines for samples 1 and 2 are not parallel, so I would say that there is evidence for an interaction between laboratory and concentration of bacteria.

```
#Problem 2b
#Fit the lmer model
milk_lmer <- lmer(bacterialcount ~ 1 + (1|lab) + (1|sample) + (1|lab:sample), data = milk)
summary(milk_lmer)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: bacterialcount ~ 1 + (1 | lab) + (1 | sample) + (1 | lab:sample)
##    Data: milk
##
## REML criterion at convergence: 613.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.31698 -0.14368  0.02496  0.09849  3.00761
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  lab:sample (Intercept)  269168   518.8
##  lab        (Intercept)  130561   361.3
##  sample     (Intercept) 2405018  1550.8
```

```
##  Residual                101935   319.3
## Number of obs: 40, groups:  lab:sample, 20; lab, 5; sample, 4
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept) 1653.250    802.106    3.229   2.061    0.125
## convergence code: 0
## Model failed to converge with max|grad| = 0.00447186 (tol = 0.002, component 1)
```

Now I will formally test for the effects of interest. First, to test the main effect of laboratory. For the random effect model, we want to test the following hypothesis for lab:

$$H_0 : \sigma_L^2 = 0$$

$$H_0 : \sigma_L^2 \neq 0$$

We can test this hypothesis using a F-test where the f-statistic is calculated as $F = \frac{MSL}{MSLC}$ where MSL is the mean squared error for the laboratory random effect and is calculated as $E(MSL) = \sigma_\epsilon^2 + n\sigma_{LC}^2 + cn\sigma_L^2$ and MSLC is the mean squared error for the interaction between lab and contamination and is calculated as $E(MSLC) = \sigma_\epsilon^2 + n\sigma_{LC}^2$. Using the R output above, we can calculate both of these quantities and perform our F test.

$$E(MSL) = \sigma_\epsilon^2 + n\sigma_{LC}^2 + cn\sigma_L^2 = 101935 + 2 \times 269168 + 4 \times 2 \times 130561 = 1684759$$

$$E(MSLC) = \sigma_\epsilon^2 + n\sigma_{LC}^2 = 101935 + 2 \times 269168 = 640271$$

Thus, our F-statistic is $F = \frac{MSL}{MSLC} = \frac{1684759}{640271} = 2.63$ with 4 degrees of freedom in the numerator and 12 degrees of freedom in the denominator. The corresponding p-value is thus `pf(1684759/640271,4,12, lower.tail = FALSE)` $= 0.086948$. Given this p-value, there is weak evidence to reject the null hypothesis that there is no random effect from laboratory. This partially agrees with my observation from the histograms above. Next, we will conduct an F-test for the interaction term.

Specifically, we are now testing the hypothesis $H_0 : \sigma_{LC}^2 = 0$. Our test-statistic is defined as $F = \frac{MSLC}{MSE}$ where MSLC is the mean squares error for the interaction term (calculated above) and MSError is estimated with residual variance from the R output. Thus,

$$F = \frac{640271}{101935} = 6.28$$

The test statistic has 12 degrees of freedom in the numerator and 20 degrees of freedom in the denominator. The p-value for this test statistic is thus ~0.0002. There is strong evidence against the null hypothesis - that is, there is strong evidence that the interaction between lab and contamination is significant.

Now, we want to estimate the effects of interest. To do this, we must fit a model with fixed effects. Note that because the interaction term is significant, the main effect of laboratory must be included in the model.

```
#Problem 2b
milk_fixed <- lm(bacterialcount ~ lab*sample, data = milk)
summary(milk_fixed)
```

```
##
## Call:
## lm(formula = bacterialcount ~ lab * sample, data = milk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -850.00  -26.25    0.00   26.25  850.00
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2150.0      225.8   9.521 7.18e-09 ***
## lab2            400.0      319.4   1.252 0.224838
## lab3           -150.0      319.4  -0.470 0.643660
## lab4           1300.0      319.4   4.071 0.000596 ***
## lab5           1800.0      319.4   5.636 1.62e-05 ***
## sample2         800.0      319.4   2.505 0.021008 *
## sample3       -1945.0      319.4  -6.090 5.94e-06 ***
## sample4       -1885.0      319.4  -5.902 8.98e-06 ***
## lab2:sample2    200.0      451.6   0.443 0.662646
## lab3:sample2   -450.0      451.6  -0.996 0.330981
## lab4:sample2  -1950.0      451.6  -4.318 0.000335 ***
## lab5:sample2     50.0      451.6   0.111 0.912953
## lab2:sample3   -340.0      451.6  -0.753 0.460340
## lab3:sample3    125.0      451.6   0.277 0.784801
## lab4:sample3  -1170.0      451.6  -2.591 0.017482 *
## lab5:sample3  -1650.0      451.6  -3.653 0.001580 **
## lab2:sample4   -295.0      451.6  -0.653 0.521089
## lab3:sample4    115.0      451.6   0.255 0.801614
## lab4:sample4  -1245.0      451.6  -2.757 0.012169 *
## lab5:sample4  -1575.0      451.6  -3.487 0.002323 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319.4 on 20 degrees of freedom
## Multiple R-squared:  0.9774, Adjusted R-squared:  0.9559
## F-statistic: 45.52 on 19 and 20 DF,  p-value: 1.867e-12
```

The coefficient estimates are the effects of laboratory and interaction between laboratory and contamination. Note that we are using lab 1, sample 1, as the baseline. From these coefficients, it appears that lab 2 and 3 do not differ much from lab 1 but labs 4 and 5 differ significantly. This matches our expectations from the interaction plot above. Thus I would say that both of these labs are to blame. Specifically, I think the interaction between lab 4 and sample 2 (which has the highest coefficient among the interaction terms) might lead us to blame lab 4 the most.

Problem 3

See attached page.

Code

```
## ----include = FALSE-----------------------------------------------------
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(knitr)
library(car)
library(ggplot2)
library(MASS)
library(lme4)
library(lmerTest)
library(tidyr)
library(cowplot)
library(multcomp)
```

```r
set.seed(1104)             # make random results reproducible

this_file <- "kerr_stat850_hw07.Rmd"  # used to automatically generate code appendix

## ------------------------------------------------------------------------
#Problem 1c
#Load the data in
breadvolume <- read.csv("~/2019spring/STAT850/hw7/breadvolume.csv")
breadvolume$fat = factor(breadvolume$fat)
breadvolume$surfactant = factor(breadvolume$surfactant)
breadvolume$day = factor(breadvolume$day)

#Build model and do ANOVA
bv_lm <- lm(volume ~ fat*surfactant + day, data = breadvolume)
anova(bv_lm)

## ------------------------------------------------------------------------
#Problem 1c
#Compare to drop1
drop1(bv_lm, test = "F")

## ------------------------------------------------------------------------
#Problem 1c
#Look at interaction coefficients
summary(bv_lm)

## ------------------------------------------------------------------------
#Problem 1d
#change the baseline
breadvolume$fat = factor(breadvolume$fat, levels = c("3","2","1"))
bv_contrasts = contrasts(breadvolume$fat)

#Now repeat c
bv_lm2 <- lm(volume ~ fat*surfactant + day, data = breadvolume, contrasts = bv_contrasts)
drop1(bv_lm2, test = "F")
summary(bv_lm2)

## ------------------------------------------------------------------------
#Problem 1e
breadvolume$fs = breadvolume$fat:breadvolume$surfactant
qplot(fs, volume, geom = "boxplot", data = breadvolume, xlab = "Fat:Surfactant")

## ------------------------------------------------------------------------
#Problem 1e
bv_aov <- aov(volume ~ fat:surfactant + day, data = breadvolume)
TukeyHSD(bv_aov, "fat:surfactant", ordered = TRUE)

## ------------------------------------------------------------------------
#Problem 2b
#Load the data in
milk <- read.csv("~/2019spring/STAT850/hw7/milkcontamination.csv")
milk$lab = factor(milk$lab)
milk$sample = factor(milk$sample)
```

```r
#Visualize data
qplot(lab,bacterialcount, geom = "boxplot", data = milk)
with(milk, interaction.plot(lab,sample,bacterialcount))

## ---- warning=FALSE-------------------------------------------------------
#Problem 2b
#Fit the lmer model
milk_lmer <- lmer(bacterialcount ~ 1 + (1|lab) + (1|sample) + (1|lab:sample), data = milk)
summary(milk_lmer)

## -------------------------------------------------------------------------
#Problem 2b
milk_fixed <- lm(bacterialcount ~ lab*sample, data = milk)
summary(milk_fixed)

## ----code = readLines(purl(this_file, documentation = 1)), echo = T, eval = F----
## # this R markdown chunk generates a code appendix
```