ST 518 Homework 7

Eric Warren

October 28, 2023

Contents

1 Problem 1

A person's blood-clotting ability is typically expressed in terms of a "prothrombin time," which is defined to be the interval between the initiation of the prothrombin-thrombin (two proteins) reaction and the formation of the final clot. Does *aspirin* affect this function? Measurements made before administration of two tablets and three hours after.

First we are going to read in the data to do our future analysis.

```
library(tidyverse)
drug <- read_table("prothrombin.dat")
(
  drug <- drug %>%
    mutate(subject = row_number(),
           difference = before - after) %>%
    dplyr::select(subject, everything())
)
```

```
## # A tibble: 12 x 4
##    subject before after difference
##      <int>  <dbl> <dbl>      <dbl>
##  1       1   12.3  12         0.300
##  2       2   12    12.3      -0.300
##  3       3   12    12.5      -0.5
##  4       4   13    12         1
##  5       5   13    13         0
##  6       6   12.5  12.5       0
##  7       7   11.3  10.3       1
##  8       8   11.8  11.3       0.5
##  9       9   11.5  11.5       0
## 10      10   11    11.5      -0.5
## 11      11   11    11         0
## 12      12   11.3  11.5      -0.200
```

1.1 Part A

Carry out a paired t-test of the hypothesis that prothrombin time is unaffected by aspirin.

To do this in R, we are going to use the `t.test()` function that will be shown below. We are looking at the difference between the before and after effects of taking aspirin on the prothrombin time. In this case, we can say our null hypothesis is $H_0 : \mu_{before} - \mu_{after} = 0$ with our alternative hypothesis saying that $H_A : \mu_{before} - \mu_{after} \neq 0$. We are going to do this test below.

```
# Transform the data to make it applicable for R t.test
drug_transform <- drug %>%
  pivot_longer(cols = c("before", "after"),
               names_to = "time",
               values_to = "response") %>%
  mutate(subject = as.factor(subject)) %>%
  select(- difference)

# Do t.test
t.test(formula = drug_transform$response ~ drug_transform$time,
```

```
       alternative = "two.sided",
       mu = 0,
       paired = TRUE,
       var.equal = TRUE,
       conf.level = 0.95) -> t_test_result

t_test_result
```

```
##
##  Paired t-test
##
## data:  drug_transform$response by drug_transform$time
## t = -0.73998, df = 11, p-value = 0.4748
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.4305585  0.2138918
## sample estimates:
## mean difference
##      -0.1083333
```

As we can see from our statistical test, we get a test statistic value (t) of -0.7399797, a degrees of freedom of 11, and a resulting p-value is 0.4748097. This p-value is higher than most significance levels we would use for hypothesis testing. Because of this we are failing to reject the null hypothesis that prothrombin time is unaffected by aspirin. As a result we can say that the difference in prothrombin time does not differ significantly after taking aspirin.

1.2 Part B

Carry out an F-test of the same hypothesis treating subjects as blocks in an analysis for a RCBD.

Here we are going to make a model by saying $Y_{ij} = \mu + \alpha_i + B_j + E_{ij}$ where $i = 1, 2$ (represents the before and after taking aspirin) and $j = 1, 2, ..., 12$ (the number of subjects). We are going to fit this model below also using our transformed data.

```
problem1_blocking_model <- lm(response ~ subject + time, drug_transform)
```

Now we are going to do our test which is looking at the difference between the before and after effects of taking aspirin on the prothrombin time. In this case, we can say our null hypothesis is $H_0 : \alpha_1 - \alpha_2 = 0$ with our alternative hypothesis saying that $H_A : \alpha_1 - \alpha_2 \neq 0$. We are going to do an F-test by using the `anova()` function on our fitted model and see what the `time` variable gives us. We are going to do this test below.

```
problem1_anova <- anova(problem1_blocking_model)

problem1_anova
```

```
## Analysis of Variance Table
##
## Response: response
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## subject   11 10.2113 0.92830  7.2186 0.001377 **
```

```
## time       1  0.0704 0.07042  0.5476 0.474810
## Residuals 11  1.4146 0.12860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from our statistical test, we get a test statistic value (F) of 0.54757, a degrees of freedom of numerator 1 and denominator 11, and a resulting p-value is 0.4748097. As this is the same p-value as in **Part A**, this p-value is higher than most significance levels we would use for hypothesis testing. Because of this we are failing to reject the null hypothesis that prothrombin time is unaffected by aspirin. As a result we can say that the difference in prothrombin time does not differ significantly after taking aspirin.

1.3 Part C

Show that, in general, the paired t-test is equivalent to the F-test for the RCBD with block size equal to 2.

We know from our class (I think covered in Module 2 or 3) but also in the practice of statistics that there is a relationship between the F-distribution and t-distribution. It follows that the F-value of something with numerator degrees of freedom 1 and denominator degrees of freedom m is equal to the squared value t-value with m degrees of freedom; moreover, $F(1, m) = [t(m)]^2$. Well whenever we have a size of two for our blocks when doing a F-test, the numerator degrees of freedom will always be $2 - 1 = 1$. The denominator degrees of freedom will always be the observation size in the data $2n$, where n is the number of subjects, subtracting the number of degrees of freedom for subject + blocks + 1. Because of 2 blocks, degrees of freedom for blocks is $2 - 1 = 1$ and degrees of freedom for subject is $n - 1$ where n is the number of subjects. Therefore, the denominator degrees of freedom, $m = 2n - (1 + n - 1 + 1) = 2n - (n + 1) = n - 1$. Now if we can prove that in a paired t-test that the degrees of freedom, $m = n - 1$ then we will have shown that when the number of blocks is true, then we have $F(1, m) = [t(m)]^2$. For a two sample t-test we know the degrees of freedom is $m = n - 1$. Therefore, the degrees of freedom for a t-test is the same as the denominator degrees of freedom for a F-test when having 2 blocks. Therefore, with 2 blocks $F(1, m) = [t(m)]^2$ holds and why that in general the paired t-test is equivalent to the F-test for the RCBD with block size equal to 2.

1.4 Part D

Consider the linear mixed effects model (or just "mixed model"), $Y_{ij} = \mu + \alpha_i + B_j + E_{ij}$ where $B_j \sim^{iid} N(0, \sigma_B^2), E_{ij} \sim^{iid} N(0, \sigma^2)$ with $B \perp E$ for $i = 1, ..., a = 2, j = 1, ..., b = 12$.

1.4.1 Part i

Show that $E[MS(block)] = \sigma^2 + a\sigma_B^2$.

We from our lecture that $\hat{\sigma_B^2} = \frac{1}{a}(MS(Block) - MS(E))$. We are going to manipulate this formula to get $MS(Block)$ by itself. So,

$$\hat{\sigma_B^2} = \frac{1}{a}(MS(Block) - MS(E))$$

$$a\hat{\sigma_B^2} = MS(Block) - MS(E)$$

$$a\hat{\sigma_B^2} + MS(E) = MS(Block)$$

Now we want to find $E[MS(Block)]$ so using $MS(Block) = a\hat{\sigma_B^2} + MS(E)$, we get that $E[MS(Block)] = E[a\hat{\sigma_B^2} + MS(E)] = E[a\hat{\sigma_B^2}] + E[MS(E)] = a\sigma_B^2 + \sigma^2$ since we know from past lectures that $E[MS(E)] = \sigma^2$. So $E[MS(Block)] = a\sigma_B^2 + \sigma^2$.

1.4.2 Part ii

Use this result to estimate the variance component for subject effects in a mixed model for the prothrombin data.

We know that $E[MS(Block)] = a\sigma_B^2 + \sigma^2$ which means that $MS(Block) = a\hat{\sigma_B^2} + \hat{\sigma^2}$. We want to find $\hat{\sigma_B^2}$ so we can change our formula $MS(Block) = a\hat{\sigma_B^2} + \hat{\sigma^2}$ to get this by saying $MS(Block) = a\hat{\sigma_B^2} + \hat{\sigma^2}$ <===> $MS(Block) - \hat{\sigma^2} = a\hat{\sigma_B^2}$ <===> $\frac{MS(Block) - \hat{\sigma^2}}{a} = \hat{\sigma_B^2}$. So $\hat{\sigma_B^2} = \frac{MS(Block) - \hat{\sigma^2}}{a}$ and we can use our anova table to help get these values. A reminder of our anova table is below.

```
problem1_anova
```

```
## Analysis of Variance Table
##
## Response: response
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## subject   11 10.2113 0.92830  7.2186 0.001377 **
## time       1  0.0704 0.07042  0.5476 0.474810
## Residuals 11  1.4146 0.12860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from this table, we can find that $MS(Block) = MS(Subject) = 0.9282955$. We also know that $\hat{\sigma^2} = MS(E) = 0.1285985$. Lastly, a is the number of levels in our other variable (in this case `time`) which is 2. So $a = 2$. Therefore, $\hat{\sigma_B^2} = \frac{MS(Block) - \hat{\sigma^2}}{a}$ = (1/2)*(0.9282955 - 0.1285985) = 0.3998485. Therefore, $\hat{\sigma_B^2}$ = 0.3998485.

1.4.3 Part iii (Told in Discussion Board we do not have to do)

Report an estimate of the intra-subject correlation. Is the scatterplot above consistent with this estimate?

Based on the discussion board post, we are told to solve this by using the formula $\frac{\sigma_S^2}{\sigma_S^2 + \sigma^2}$. We know that $\sigma_S^2 = \sigma_B^2 = 0.3998485$ and $\sigma^2 = 0.1285985$. Therefore, the correlation is 0.3998485 / (0.3998485 + 0.1285985) = 0.7566483, which is a fairly strong correlation. When comparing to the scatterplot, there seems to be weak correlation on this so the results from the estimate and the scatterplot do not seem to match up.

2 Problem 2

Fuel efficiency of four blends of gasoline is measured in MPG. There is considerable variability due to driver. Another source of variability is model of car. An experiment randomizes four models of car and gasoline blends (A,B,C,D) to drivers according to the design below:

Driver	Model 1	Model 2	Model 3	Model 4
1	15.5(A)	33.8(B)	13.7(C)	29.2(D)
2	16.3(B)	26.4(C)	19.1(D)	22.5(A)
3	10.5(C)	31.5(D)	17.5(A)	30.1(B)
4	14.0(D)	34.5(A)	19.7(B)	21.6(C)

2.1 Part A

Assuming normally distributed data, propose a model in which the effects of model, driver and blend are additive on the mean.

We can make a model that is found from what we have learned for Latin Squares (as this experiment is designed in this way) as $Y_{ij} = \mu + \rho_i + \kappa_j + \tau_k + E_{ij}$ where $i = 1, 2, 3, 4; j = 1, 2, 3, 4; E_{ij} \sim^{iid} N(0, \sigma^2)$ and k is determined by the design.

Now we are going to put this model into R. First we are going to store the data and then we are going to make the appropriate model.

```
# Make vector of rows
model <- rep(1:4, 4)
driver <- c(rep(1, 4), rep(2, 4), rep(3, 4), rep(4, 4))
gas_blend <- c(LETTERS[1:4], LETTERS[c(2:4, 1)], LETTERS[c(3:4, 1:2)], LETTERS[c(4, 1:3)])
mpg <- c(15.5, 33.8, 13.7, 29.2,
         16.3, 26.4, 19.1, 22.5,
         10.5, 31.5, 17.5, 30.1,
         14.0, 34.5, 19.7, 21.6)

# Make data frame
mpg_df <- data.frame(driver, model, gas_blend, mpg) %>%
  mutate(driver = as.factor(driver),
         model = as.factor(model),
         gas_blend = as.factor(gas_blend))

# Make model
mpg_model <- lm(mpg ~ driver + model + gas_blend, mpg_df)
```

2.2 Part B

Obtain an ANOVA table for this model.

Here we are going to show the ANOVA table below for our model

```
(problem2_anova <- anova(mpg_model))
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value     Pr(>F)
## driver     3   8.33   2.777  0.6443    0.61428
## model      3 755.37 251.791 58.4116 0.00007838 ***
## gas_blend  3 106.27  35.424  8.2178    0.01515 *
## Residuals  6  25.86   4.311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Something to note is that it seems to be that the driver is said to not be significantly different so it is plausible to think that driver could potentially not make a difference.

2.3 Part C

Report the average fuel efficiency for each blend and the lowest level of significance α at which these averages can be said to differ significantly.

First we are going to find the average fuel efficiency for each blend.

```
(
  mpg_blend_avgs <- mpg_df %>%
    group_by(gas_blend) %>%
    summarize(avg_mpg = mean(mpg))
)
```

```
## # A tibble: 4 x 2
##   gas_blend avg_mpg
##   <fct>       <dbl>
## 1 A            22.5
## 2 B            25.0
## 3 C            18.0
## 4 D            23.4
```

We can see the average mpg for each gas blend with gas blend **B** having the highest and gas blend **C** having the lowest. From our anova table we can see that the `gas_blend` variable p-value is 0.0151489 which by definition is also the lowest level of significance α at which these averages can be said to differ significantly. Therefore, the lowest level of significance α at which these averages can be said to differ significantly is 0.0151489.

3 Problem 3

For each of several species, an Ecology researcher ran an exposure assay in which groups of $n = 50$ ants are measured for mortality after exposure to one of three different bacteria. For each species, ants are randomized to the three bacteria treatments within each of 15 colonies. That is, a randomized complete block design is used for each species, with colonies serving as complete blocks. Find the results for the DB species on moodle as **"DBdat.mtx"**.

Here we are going to read in the data.

```
(
  db <- read_csv("DBdat.mtx", skip = 1) %>%
    mutate(Colony = as.factor(Colony))
)
```

```
## # A tibble: 15 x 4
##    Colony Control S.epid E.coli
##    <fct>    <dbl>  <dbl>  <dbl>
##  1 1         0.16   0.12   0.86
##  2 2         0.14   0      0.6
##  3 3         0      0      0.36
##  4 4         0.02   0.02   0.54
##  5 5         0.1    0.08   0.38
##  6 6         0.06   0.2    0.56
##  7 7         0      0.04   0.62
```

```
##  8 8         0.06   0      0.16
##  9 9         0      0      0.3
## 10 10        0.06   0.02   0.14
## 11 11        0.02   0.02   0.18
## 12 12        0.14   0.06   0.32
## 13 13        0.04   0.04   0.08
## 14 14        0.08   0      0.24
## 15 15        0.04   0.04   0.26
```

3.1 Part A

Watch the video on the `hiddenf` package. What is the name of the function in R that will create a directory system containing all of the help files that a developer can complete to create documentation for a package? `____-skeleton`.

The function is called `package.skeleton`.

3.2 Part B

Obtain an interaction plot with proportion of ants dying out of $n = 50$ on the vertical axis, bacteria treatment on the horizontal axis and lines connecting mortality rates from the same colony.

Here we are going to make an interaction plot as described above. We need to pivot our data to a longer format do make this work. We will do this first.

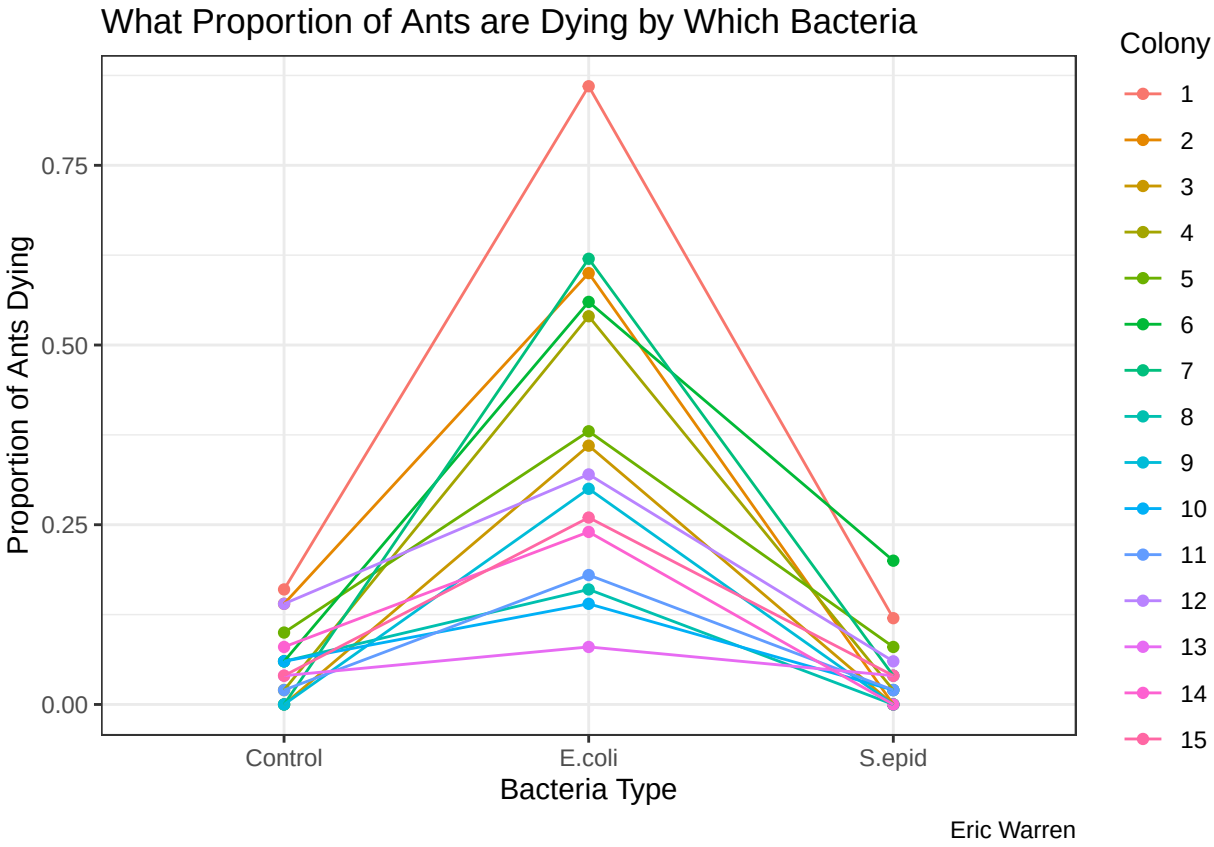
```
(
  db_longer <- db %>%
    pivot_longer(cols = Control:E.coli,
                 names_to = "bacteria",
                 values_to = "rates")
)
```

```
## # A tibble: 45 x 3
##    Colony bacteria rates
##    <fct>  <chr>    <dbl>
##  1 1      Control   0.16
##  2 1      S.epid    0.12
##  3 1      E.coli    0.86
##  4 2      Control   0.14
##  5 2      S.epid    0
##  6 2      E.coli    0.6
##  7 3      Control   0
##  8 3      S.epid    0
##  9 3      E.coli    0.36
## 10 4      Control   0.02
## # i 35 more rows
```

Now we are going to make the interaction plot as described in the directions.

```
db_longer %>%
  ggplot(aes(x = bacteria, y = rates)) +
  geom_point(aes(color = Colony)) +
```

```
  geom_line(aes(group = Colony, color = Colony)) +
  labs(x = "Bacteria Type",
       y = "Proportion of Ants Dying",
       color = "Colony",
       title = "What Proportion of Ants are Dying by Which Bacteria",
       caption = "Eric Warren") +
  theme_bw()
```



It seems that **E.coli** has a spike in the proportion of ants dying in each colony.

3.3 Part C

Either by using the `hiddenf` package in R, or by assigning colonies to groups, obtain an ANOVA table with the following sources of variability: treatments, groups, group-by-treatment interaction, colony within group.

We are going to use the `hiddenf` package in R to help us get this anova table. We need to make sure our data is a matrix as this package only takes from those values. We also need to make sure the matrix does not include any of our factor variables (for example we need to make sure `colony` is not included in the package). We are going to do this below. Eventually when our data is the correct form we can get our ANOVA table.

```
library(hiddenf)

# Make data into matrix with only bacteria columns
db_mtx <- as.matrix(db[ , 2:4])
```

```r
# Now put it in the hiddenf format
db_hiddenf <- HiddenF(db_mtx)

# Get anova table
anova(db_hiddenf)
```

```
## The ACMIF test for the hidden additivity form of interaction
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value             Pr(>F)
## group      1 0.24128 0.24128  54.4314 0.0000000779924377
## col        2 1.03516 0.51758 116.7614 0.0000000000001024
## row       13 0.11415 0.00878   1.9808            0.06693
## group:col  2 0.28612 0.14306  32.2724 0.0000000902458525
## Residuals 26 0.11525 0.00443
## C.Total   44 1.79196
## (Pvalues in ANOVA table are NOT corrected for multiplicity.)
```

This anova table matches up with SAS's table by doing the `PROC GLM` in its platform. I have attached a chunk of code and the output below.

```
 * Get a set of the data
 for grouping
 ;
data DB1;
     set DB1;
     group = 2 -(colony in (1 ,2 ,4 ,6 ,7));
 run;

 * Get the values for the
 terms for the hw problem
 ;
proc glm;
 class treatment group colony ;
 model y= treatment|group colony(group);
 lsmeans treatment|group ;
 run;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treatment	2	1.03516444	0.51758222	116.76	<.0001
group	1	0.24128444	0.24128444	54.43	<.0001
treatment*group	2	0.28611556	0.14305778	32.27	<.0001
colony(group)	13	0.11414667	0.00878051	1.98	0.0669

Note for purpose of comparing R to SAS output the following terms match up:

- `group` (in R) with `group` (in SAS)
- `col` (in R) with `treatment` (in SAS)
- `row` (in R) with `colony(group)` (in SAS)
- `group:col` (in R) with `treatment*group` (in SAS)

After finding those matching values we can see a consistent anova table between both platforms and use these values we found with the `hiddenf` package.

3.4 Part D

Obtain the p-value for the test of group-by-treatment interaction after and report it after multiplying by $2^{15-1} - 1$. Is there evidence that the resistance to bacteria varies across colonies?

There are multiple ways we can do this. The first is using the `hiddenf` package anova table p-value and multiply by $2^{15-1} - 1$ to get the real p-value. In this case, I can get 0.0000000902458525 as our p-value from our table and then multiplying it by $2^{15-1} - 1$ gets 0.0014785.

The second option is to use the summary function with our `hiddenf` object.

```
summary(db_hiddenf)
```

```
## Number of configurations: 16383
## Minimum adjusted pvalue: 0.001478498
##
## Rows in group 1: 1 2 4 6 7
## Rows in group 2: 3 5 8 9 10 11 12 13 14 15
##
## Column means for grp 1: 0.076 0.076 0.636
## Column means for grp 2: 0.054 0.026 0.242

## $group1
## [1] 1 2 4 6 7
##
## $group2
##  [1]  3  5  8  9 10 11 12 13 14 15
##
## $grp1means
```

```
## [1] 0.076 0.076 0.636
##
## $grp2means
## [1] 0.054 0.026 0.242
```

We see that our minimum adjusted p-value is 0.001478498 from this output.

The third method is using the SAS anova table to get the F-value for the interaction term being 32.27 with degrees of freedom 2, 26 (this is found as our error degrees of freedom which is just the total observations – 45 – minus the degrees of freedom from the other terms added up – $1 + 2 + 13 + 2 = 18$ minus one more additional degree of freedom). So we can get our error degrees of freedom of $45 - 18 - 1 = 26$. Anyways, we can get the p-value of a F-value of 32.27 with degrees of freedom 2, 26 using the `pf(F-value, df1, df2, lower.tail = F)` to get `pf(32.27, 2, 26, lower.tail = F)` or a corresponding p-value of 0.0000001. Now we can multiply this by $2^{15-1} - 1$ to get a real p-value which is 0.0014795 which is also roughly the same p-value.

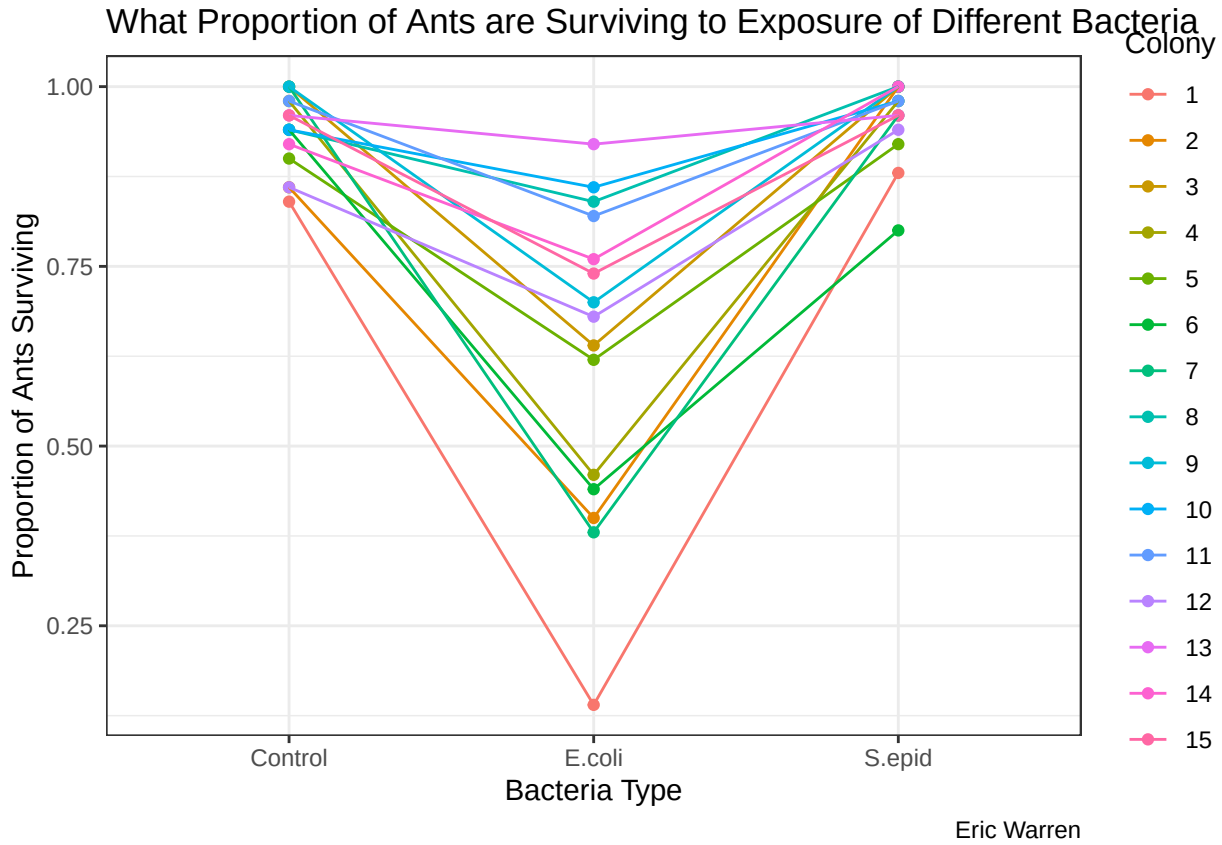
If we had a test of significance that was $\alpha = .05$ as most are, since our p-value has been shown to be less than that we can say that there is statistically significant evidence to say that the resistance to bacteria varies across colonies.

3.5 Part E

Obtain another plot of survival versus bacteria with different lines for colonies. Color the colony lines according to group.

Now we are going to make the survival plot as described in the directions. Note we are given the mortality rates and the survival rates are $1 - mortality$. Thus, we are going to make our value for graphing being `1 - rates` in our code and use the tall formatted data frame we made before which is `db_longer`.

```
db_longer %>%
  ggplot(aes(x = bacteria, y = 1 - rates)) +
  geom_point(aes(color = Colony)) +
  geom_line(aes(group = Colony, color = Colony)) +
  labs(x = "Bacteria Type",
       y = "Proportion of Ants Surviving",
       color = "Colony",
       title = "What Proportion of Ants are Surviving to Exposure of Different Bacteria",
       caption = "Eric Warren") +
  theme_bw()
```



Again we can see that it seems there is a proportion of less ants are surviving (or a proportion of more ants are dying) when exposed to E.coli than the other two bacterias.

4 Problem 4

An experiment investigates the growth of oysters. Four bags with ten oysters each are randomly placed at four underwater stations next to a power plant:

- Trt1: At the bottom of a discharge canal
- Trt2: At the top of a discharge canal
- Trt3: At the bottom of an intake canal
- Trt4: At the top of an intake canal

Average initial weight x and final weight y are measured for each of the 16 bags. (Bags serve as the experimental units.) Let $z = x - \bar{x}$ denote the difference from average of the initial weights. SAS code and output to fit an ANCOVA model appear at the end of the exam.

First we will read in the data to make sure we can do our analysis later.

```
(oysters <- read_table("oysters.dat") %>%
   mutate(trt = as.factor(trt)))
```

```
## # A tibble: 16 x 4
##    trt   initial     z final
```

```
##    <fct>   <dbl> <dbl> <dbl>
##  1 1        27.2   0.2  32.6
##  2 1        32     5    36.6
##  3 1        33     6    37.7
##  4 1        26.8  -0.2  31
##  5 2        28.6   1.6  33.8
##  6 2        26.8  -0.2  31.7
##  7 2        26.5  -0.5  30.7
##  8 2        26.8  -0.2  30.4
##  9 3        28.6   1.6  35.2
## 10 3        22.4  -4.6  29.1
## 11 3        23.2  -3.8  28.9
## 12 3        24.4  -2.6  30.2
## 13 4        29.3   2.3  35
## 14 4        21.8  -5.2  27
## 15 4        30.3   3.3  36.4
## 16 4        24.3  -2.7  30.5
```

4.1 Part A

Obtain the F-ratio for a test of equal final weights in a one-way ANOVA where initial weight z is ignored.

If we ignore the initial weight then we are only fitting a model with `final` being our response variable and `trt` being our only predictor. Thus, we can make the model below with the appropriate anova table

```
# Make the model
ignore_z_model <- lm(final ~ trt, oysters)

# Show anova table
(anova_ignore_z <- anova(ignore_z_model))
```

```
## Analysis of Variance Table
##
## Response: final
##           Df  Sum Sq Mean Sq F value Pr(>F)
## trt        3  29.045  9.6817  0.9745 0.4369
## Residuals 12 119.215  9.9346
```

As we can see in our one way anova table, we can that the F-value for our test of equal weights is 0.9745418 with a degrees of freedom of 3, 12. This gets us a corresponding p-value of 0.4368591, which gives us a conclusion that we do not have statistically significant evidence to say that the final weights of each treatment are different. Therefore, we can say that the final weights of each bag are plausibly the same.

To answer the question, we found the F-ratio is 0.9745418 with a degrees of freedom of 3, 12.

4.2 Part B

Obtain the F-ratio for a test of equal final weights in a one-way ANOVA after controlling for initial weight z.

Here when controlling for the initial weight we are looking at two models to compare our anova table on. The first model is the full model of $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_z x_i$ and the reduced model we are comparing to is $Y_i = \beta_0 + \beta_z x_i$. We can do this test below where our $H_0 : \beta_1 = \beta_2 = \beta 3 = 0$ and $H_A :\geq 1 \beta_i \neq 0$. We are going to do this test in our code chunk below.

```
full_model <- lm(final ~ trt + z, oysters)
reduced_model <- lm(final ~ z, oysters)

(control_z_anova <- anova(reduced_model, full_model))
```

```
## Analysis of Variance Table
##
## Model 1: final ~ z
## Model 2: final ~ trt + z
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     14 11.1787
## 2     11  3.5046  3    7.6741 8.0289 0.004101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see in our one way anova table, we can that the F-value for our test of equal weights after controlling for initial weight z is NA with a degrees of freedom of 3, 11. This gets us a corresponding p-value of 0.0041005, which gives us a conclusion that we have statistically significant evidence to say that the final weights of each treatment are different after controlling for initial weight z.

To answer the question, we found the F-ratio is is NA with a degrees of freedom of 3, 11.

4.3 Part C

Use the output to report the unadjusted means for treatments 1 and 4. (Use the 2nd column in the table below.)

Treatment	Unadjusted Mean	Adjusted Mean	Standard Error
1			
4			

We can get the unadjusted means by taking the average of each treatment's response (which comes from the column `final`). We are going to take the means of each treatment below.

```
(
  oyster_unadjusted_means <- oysters %>%
    group_by(trt) %>%
    summarize(avg = mean(final))
)
```

```
## # A tibble: 4 x 2
##   trt     avg
##   <fct> <dbl>
## 1 1      34.5
## 2 2      31.6
## 3 3      30.8
## 4 4      32.2
```

Here we can see that the unadjusted average for treatment 1 is 34.475 and the unadjusted average for treatment 4 is 32.225. Now our chart looks like this:

Treatment	Unadjusted Mean	Adjusted Mean	Standard Error
1	34.475		
4	32.225		

Note this matches the output we are given in the homework problem, but it is good to know the steps if we had to code this ourselves from scratch.

4.4 Part D

For bag i, let $x_{i1}, ..., x_{i4}$ denote indicator variables for treatments 1-4, respectively. Consider the analysis of covariance model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_z x_i + E_i$

4.4.1 Part i

Use the fitted model to report the mean final weight, after adjustment to the average initial weight $\bar{x}$, for treatments 1 and 4. Fill in the table above, show any work below.

Now remember our table from **Part C** which is below:

Treatment	Unadjusted Mean	Adjusted Mean	Standard Error
1	34.475		
4	32.225		

We are now going to find the adjusted means for the treatments. Here we are going to group the treatments and find the mean of the `z` variable. We do this below.

```
(
  z_averages <- oysters %>%
    group_by(trt) %>%
    summarize(avg = mean(z))
)
```

```
## # A tibble: 4 x 2
##   trt      avg
##   <fct>  <dbl>
## 1 1      2.75
## 2 2      0.175
## 3 3     -2.35
## 4 4     -0.575
```

Note this matches the output we are given in the homework problem, but it is good to know the steps if we had to code this ourselves from scratch.

Now we are going to find the $\bar{z}$ or the average of the covariates. In this case we know that $\bar{z} = \frac{\bar{z_1}+\bar{z_2}+\bar{z_3}+\bar{z_4}}{4} =$ (2.75 + 0.175 + -2.35 + -0.575) / 4 = 0. Therefore, we can use this by saying that $\bar{z} = 0$. This value makes sense since z is found by subtracting from a mean of another value by the original value so we should get the mean minus the mean (on average) which gives the value of 0. Again for practice, it is good to know the math behind it for in the future when this is not the case.

Now we can find the adjusted means for treatments 1 and 4 below using the formula to find the adjusted mean of treatment i of $\bar{y_{i,a}} = \hat{\beta}_0 + \hat{\beta}_i + \hat{\beta}_z * \bar{z}$. Note to get the exact values to match up with the correct

answers we are going to use the SAS output on the last page of the homework to get the "correct" values for each β_i. We are first going to list what SAS has for each β_i value which is slightly different that how I have found it in R.

- $\beta_0 = 32.82685402$
- $\beta_1 = -1.23028630$
- $\beta_2 = -1.36002698$
- $\beta_3 = 0.48289720$
- $\beta_4 = 0$
- $\beta_z = 1.04670265$

We can plug this formula $\bar{y_{i,a}} = \hat{\beta}_0 + \hat{\beta}_i + \hat{\beta}_z \bar{z}$ in below with our respective treatments to find:

- Treatment 1: $\bar{y_{1,a}} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_z * \bar{z} = 32.82685402 + (-1.23028630) + 1.04670265 \text{ x } 0 = 31.5965677$
- Treatment 4: $\bar{y_{4,a}} = \hat{\beta}_0 + \hat{\beta}_4 + \hat{\beta}_z \bar{z} = 32.82685402 + 0 + 1.04670265 \text{ x } 0 = 32.826854$. This makes sense that this value is just the `intercept` value we get in our output since our output is done with the parameter estimates corresponding to the parameterization with treatment 4 set to 0.

Now we can update our table with the following values by plugging in our adjusted means.

Treatment	Unadjusted Mean	Adjusted Mean	Standard Error
1	34.475	31.5965677	
4	32.225	32.826854	

4.4.2 Part ii

Report the standard error for the adjusted means for locations 1 and 4. (Fill in the table, writing "NP" if it is not possible to give a number based on the provided output.)

The formula to find the standard error is $\hat{SE} = \sqrt{c'(X'X)^{-1}cMS(E)}$. Given we do not have $(X'X)^{-1}$ from the output, this would make it very difficult to solve. The good news is that we can do this in SAS and get these estimates. I have attached images of the code and output below.

```
* Get the adjusted means
and standard errors
;
PROC GLM DATA=oysters;
CLASS trt;
MODEL final = trt z;
means trt;
lsmeans trt/stderr;
RUN;
```

The GLM Procedure
Least Squares Means

trt	final LSMEAN	Standard Error	Pr > |t|
1	31.5965677	0.3201005	<.0001
2	31.4668270	0.2823885	<.0001
3	33.3097512	0.3103391	<.0001
4	32.8268540	0.2839864	<.0001

We can also do this in R by using the `emmeans::emmeans()` function. We will show this below.

```
(problem4_adjusted_means_se <- emmeans::emmeans(full_model, ~trt))
```

```
##  trt emmean    SE df lower.CL upper.CL
##  1     31.6 0.320 11     30.9     32.3
##  2     31.5 0.282 11     30.8     32.1
##  3     33.3 0.310 11     32.6     34.0
##  4     32.8 0.284 11     32.2     33.5
##
## Confidence level used: 0.95
```

This output matches what we found in SAS so we should be very confident we found the stanard errors. Also not the adjusted means also match from what we got before. As we can see the standard error for treatment 1 is 0.3201005 and the standard error for treatment 4 is 0.2839864 (which also seems to match the intercept standard error from the output). We can now fill out our table completely.

Treatment	Unadjusted Mean	Adjusted Mean	Standard Error
1	34.475	31.5965677	0.3201005
4	32.225	32.826854	0.2839864

4.5 Part E

Consider the difference between mean final weights under treatments 1 and 4. Estimate this difference after controlling for initial weight. Report a standard error and a p-value for a test of no difference.

So when we want to look at the mean final weights under treatments 1 and 4 after controlling for initial weight, we are just looking at our adjusted mean values. Since we are parameterizing on treatment 4, as described before, this esimated difference between the two treatments is just $\hat{\beta}_1 = -1.23028630$ since $\hat{\beta}_1 =$ adjusted mean from treatment 1 - adjusted mean from treatment 4. To continue our standard error is just the $SE(\hat{\beta}_1) = 0.43892225$. Now our hypothesis test to see if there is any difference between the two treatments' adjusted means is just testing with $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$. The good news is that our output gives us a t-value of 2.80 (could also be solved by doing $(1.23028630 - 0)/0.43892225$) with 11 degrees of freedom

(from the error degrees of freedom in our output) and this gives us a two sided test p-value of about 0.0172, which can be found from output or by doing `2 * pt((1.23028630 - 0)/0.43892225, 11, lower.tail = F)` in R. This test would tell us that there is a statistically significant difference in mean final weights under treatments 1 and 4 after controlling for initial weight.

So in conclusion we found the estimated difference to be $\hat{\beta}_1 = -1.23028630$, the standard error to be $SE(\hat{\beta}_1) = 0.43892225$, and the corresponding p-value to be about 0.0172.