MATH4561 Mastery VI

Melissa Van Bussel

April 20, 2019

Problem 1

A study was conducted to determine the effects of individual bathers on the fecal and total coliform bacterial populations in water. The variables of interest were the time since the subject's last bath, the vigor of the subject's activity in the water, and the subject's sex. The experiments were perfomed in a 100-gallon polyethylene tub using dechlorinated tap water at 38 degrees Celsius. The bacterial contribution of each bather was determined by subtracting the bacterial concentration measured at 15 and 30 minutes from that measured initially. A replicated 2^3 factorial design was used for this experiment. (Note: Because the measurement of bacterial populations in water involves a dilution technique, the experimental errors do not have constant variance. Rather, the variation increases with the value of the mean.) Perform analysis using a logarithmic transfomation of the data.

The variables of interest are:

- x_1: time since last bath, with levels 1-hour and 24-hours
- x_2: vigor of bathing activity, with levels Lethargic and Vigorous
- x_3: sex of bather, with levels Female and Male

The response variable is:

- y_3: Total coliform contribution after 15 minutes (organisms/100mL)

(a) Briefly explain why this is a 2^3 factorial design.

(b) Calculate the factorial effects (main and interaction effects) on total coliform populations after 15 minutes. Interpret the main effect of x_1, and the interaction between x_1 and x_3. Ensure you model the logarithm of y_3.

```
prb0506 <- read.table(file = "prb0506.dat",header = T)
```

Problem 1: Solution

1 (a)

This is referred to as a 2^3 factorial design because we have 3 factors, each of which have two levels. The first factor is time since last bath (levels: 1 hour or 24 hours), the second factor is vigor of bathing activity (levels: lethargic or vigorous), and the third factor is sex of bather (levels: male or female). The number of treatment combinations is $2^3 = 8$, which is where the name comes from.

1 (b)

To estimate the effects, we can use the `lm` function.

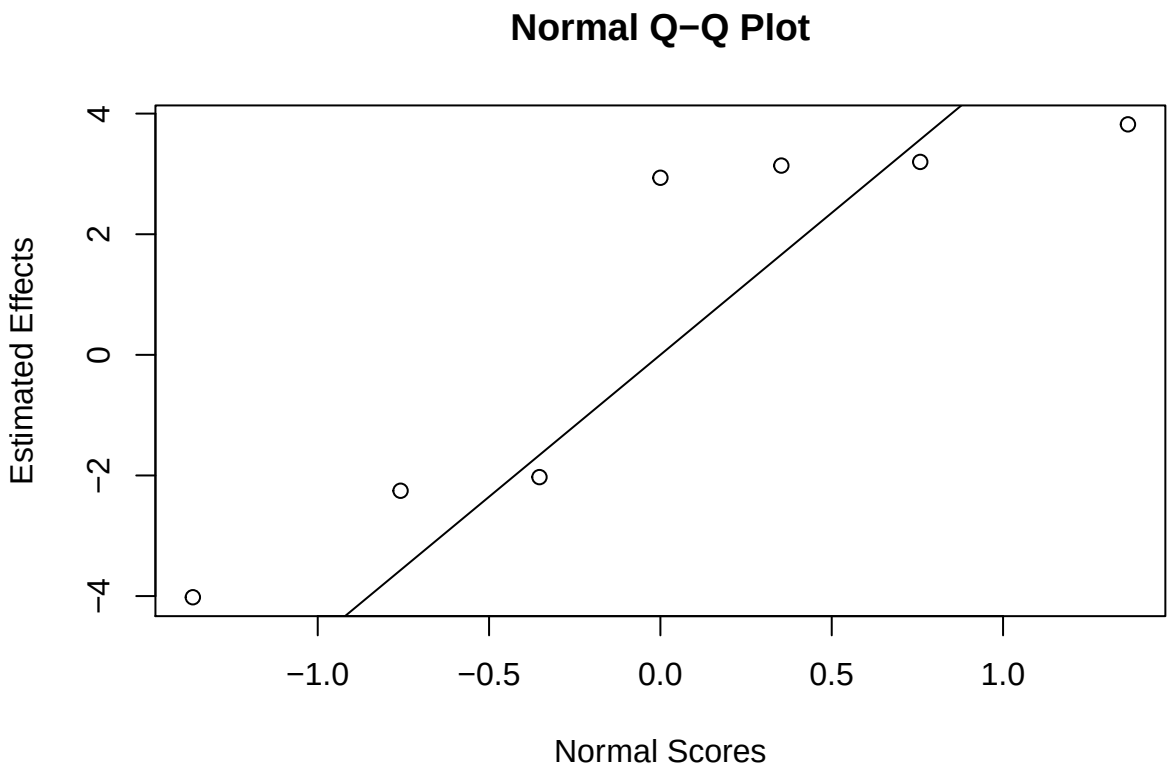
```
bath <- as.factor(prb0506$x1)
vigor <- as.factor(prb0506$x2)
sex <- as.factor(prb0506$x3)
response <- log(prb0506$y3)
```

```
coliform <- data.frame(bath, vigor, sex, response)
mod <- lm(response ~ bath*vigor*sex, data = coliform)
summary(mod)
```

```
##
## Call:
## lm(formula = response ~ bath * vigor * sex, data = coliform)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4534 -0.6505  0.0000  0.6505  1.4534
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.7006     0.8278   2.054   0.0740 .
## bath1               3.1383     1.1707   2.681   0.0279 *
## vigor1              2.9365     1.1707   2.508   0.0365 *
## sex1                3.8218     1.1707   3.265   0.0114 *
## bath1:vigor1       -4.0194     1.6556  -2.428   0.0413 *
## bath1:sex1         -2.2535     1.6556  -1.361   0.2106
## vigor1:sex1        -2.0275     1.6556  -1.225   0.2555
## bath1:vigor1:sex1   3.1990     2.3413   1.366   0.2090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.171 on 8 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.5823
## F-statistic: 3.987 on 7 and 8 DF,  p-value: 0.03551
```

The "effects" are simply the coefficients of the terms in the model. From the summary output, we can see which effects are significant. However, a more visual representation can also help with the interpretation. Recall that if all coefficients are insignificant, then they should come from a Normal distribution with mean 0 due to the Central Limit Theorem. Therefore, by creating a Normal Q-Q plot, we can see which effects and interactions are significant. Coefficients which fall along the straight line are insignificant (come from a Normal Distribution), and outliers are significant (do not come from a Normal Distribution).

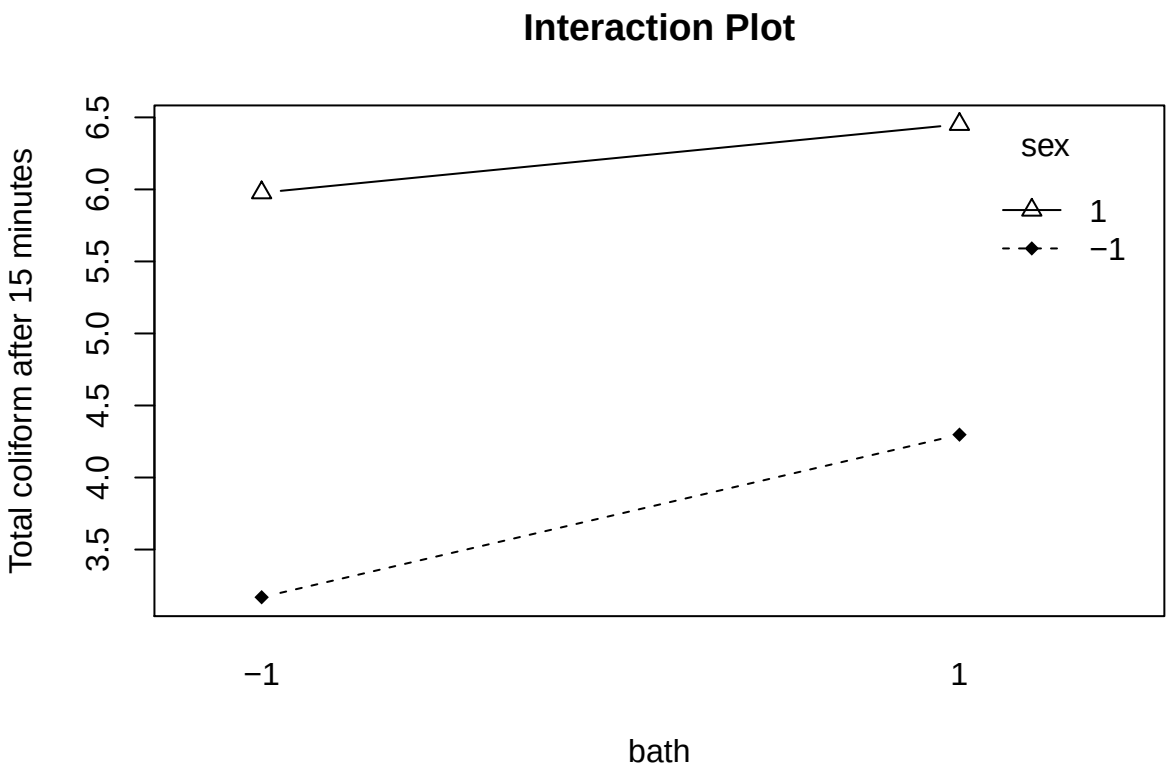
```
library(daewr)
fullnormal(coef(mod)[-1], alpha=.025)
```



We can see from the plot that many of the points deviate from the line, indicating significant effects. The only terms which are insignifcant are the interaction terms `bath*sex`, `vigor*sex` and `bath*vigor*sex`.

We can further visualize the interaction effects by using interaction plots. For this question, we are interested in the interaction between `bath` and `sex`:

```
with(coliform, (interaction.plot(bath, sex, response, type = "b", pch =
c(18,24), main = "Interaction Plot",
xlab = "bath", ylab = "Total coliform after 15 minutes")))
```




```
## NULL
```

From this interaction plot, we observe that there is no significant interaction, since the lines are approximately parallel. This tells us that, on average, changing the experimental unit from a man to a woman has a similar effect on the response (total coliform after 15 minutes) when the experimental unit has bathed an hour ago compared to when they bathed 24 hours ago. Changing the level of the `sex` variable does not change the effect that the `bath` variable has on the response, and vice versa.

However, since we are also interested in the `bath` variable for this question, we need to consider any terms in the model which involve this variable, including interaction terms. If there are significant interaction terms which involve the `bath` variable, then the coefficient on the `bath` variable cannot be interpreted separately / on its own.

Recall that the `bath*vigor` interaction term was significant in our model. Thus, we cannot directly interpret the `bath` variable's coefficient without considering other terms. We can, however, take a look at the coefficient of the `bath` variable and make a more general statement about it. First, we need to know which level of this factor had a higher average response.

```
bath_neg <- mean(subset(coliform, bath == -1, select = response)$response)
bath_pos <- mean(subset(coliform, bath == 1, select = response)$response)
bath_neg
```

```
## [1] 4.572881
```

```
bath_pos
```

```
## [1] 5.374509
```

We can see that the mean coliform level is higher for `bath = 1`. We aren't told whether `bath = 1` corresponds to 1 hour or 24 hours, but it seems reasonable to assume that bathing more recently would correspond to less

bacteria, so let's make the assumption (for the purpose of answering this question) that `bath = 1` corresponds to having taken a bath 24 hours ago, while `bath = -1` corresponds to having taken a bath 1 hour ago.

If we pretend for a moment that there was no significant interaction term involving the `bath` variable, we would interpret this coefficient in the following way: if the other factors are kept constant (vigor and sex), then changing the `bath` variable from `-1` to `1` will increase the log total coliform contribution after 15 minutes by an average of 3.1383217. For example, if we were to take a measurement of the total coliform after 15 minutes from a man doing vigorous exercise who bathed an hour ago (and take the log of the measurement), and then repeat the experiment but have the man not shower for 24 hours beforehand, then we would expect the measurement of total coliform after 15 minutes to increase by roughly $e^{3.138322}$. However, this is a very small increase, and is essentially negligible, as this factor was found to be insignificant. While this factor does have *some* effect on the total coliform measurement (since the coefficient is not 0), the effect is so small that it is not worth including in our final model. We always desire the simplest model possible (the parsimonious model), so we do not include variables which are found to be insignificant.

However, there is a significant interaction term involving the `bath` variable, so we cannot interpret the coefficient directly like that. Rather, we can interpret the coefficient as being significant, meaning that, on average, experimental units who bathe 24 hours ago will affect / increase the response more than those who bathed an hour ago. We can't directly interpret the value of the coefficient itself, though, because we'd need to take other terms in the model into consideration.

Problem 2

Le Riche and Csima (1964) evaluated four hypnotic drugs and a placebo to determine their effect on the quality of sleep in elderly patients. The treatment levels were labeled (A = Placebo, B = Ethchlorvynol, C = Glutethimide, D = Chloral hydrate, and E = Secobarbitol sodium). Elderly patients were given one of the capsules for five nights in succession and their quality of sleep was rated by a trained nurse on a 4-point scale (0 = poor to 3 = excellent) each night. An average score was calculated for each patient over the five nights in a week. Each patient received all five treatments in successive weeks. A Latin-square design was used to account for patient-to-patient differences and week-to-week effects. The design and the response (mean quality of sleep rating) are shown in the table attached.

Patient	1	2	3	4	5
1	B (2.92)	E (2.43)	A (2.19)	C (2.71)	D (2.71)
2	D (2.86)	A (1.64)	E (3.02)	B (3.03)	C (3.03)
3	E (1.97)	B (2.50)	C (2.47)	D (2.65)	A (1.89)
4	A (1.99)	C (2.39)	D (2.37)	E (2.33)	B (2.71)
5	C (2.64)	D (2.31)	B (2.44)	A (1.89)	E (2.78)

(a) What is the appropriate model for this data?

(b) Complete the ANOVA and determine if there are any significant differences among the treatments.

(c) Use an appropriate method to determine if there is a significant difference between the placebo and the average of the other drugs, and if there are significant differences among the four drugs.

Problem 2: Solution

2 (a)

In this situation, we are treating each patient as a block, and we are using a Latin Square Design to randomly assign the treatment order of the 5 treatments in each "block". Our column blocking factor is time, while

our row blocking factor is the patient. Our 5 treatment levels are A = Placebo, B = Ethchlorvynol, C = Glutethimide, D = Chloral hydrate, and E = Secobarbitol sodium. Therefore, our model is:

$$y_{ijk} = r_i + c_j + \tau_k + \epsilon_{ijk}$$

where $r_i,\ i = 1,2,3,4,5$ is the row blocking effect, $c_j,\ j = 1,2,3,4,5$ is the column-blocking effect, $\tau_k,\ k = 1,2,3,4,5$ is the treatment effect, and ϵ_{ijk} is the experimental error term.

2 (b)

First, we import the data.

Patient	1	2	3	4	5
1	B (2.92)	E (2.43)	A (2.19)	C (2.71)	D (2.71)
2	D (2.86)	A (1.64)	E (3.02)	B (3.03)	C (3.03)
3	E (1.97)	B (2.50)	C (2.47)	D (2.65)	A (1.89)
4	A (1.99)	C (2.39)	D (2.37)	E (2.33)	B (2.71)
5	C (2.64)	D (2.31)	B (2.44)	A (1.89)	E (2.78)

```
patient <- as.factor(rep(1:5, each = 5))
week <- as.factor(rep(1:5, 5))
p1t <- c("B", "E", "A", "C", "D") # patient 1 treatment order
p2t <- c("D", "A", "E", "B", "C")
p3t <- c("E", "B", "C", "D", "A")
p4t <- c("A", "C", "D", "E", "B")
p5t <- c("C", "D", "B", "A", "E")
treatment <- as.factor(c(p1t, p2t, p3t, p4t, p5t))
p1r <- c(2.92, 2.43, 2.19, 2.71, 2.71) # patient 1 response
p2r <- c(2.86, 1.64, 3.02, 3.03, 3.03)
p3r <- c(1.97, 2.50, 2.47, 2.65, 1.89)
p4r <- c(1.99, 2.39, 2.37, 2.33, 2.71)
p5r <- c(2.64, 2.31, 2.44, 1.89, 2.78)
response <- c(p1r, p2r, p3r, p4r, p5r)
sleep <- data.frame(patient, week, treatment, response)
head(sleep)
```

```
##   patient week treatment response
## 1       1    1         B     2.92
## 2       1    2         E     2.43
## 3       1    3         A     2.19
## 4       1    4         C     2.71
## 5       1    5         D     2.71
## 6       2    1         D     2.86
```

Now, we can get the ANOVA table:

```
mod <- aov(response ~ patient + week + treatment)
summary(mod)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## patient      4 0.6073  0.1518   3.178 0.05358 .
## week         4 0.3689  0.0922   1.930 0.17000
## treatment    4 2.0498  0.5125  10.725 0.00062 ***
```

```
## Residuals   12 0.5734  0.0478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the summary of the model that the `treatment` variable is significant at significance level $\alpha = 0.001$. Therefore, we have strong evidence that there are some differences between the different treatment levels.

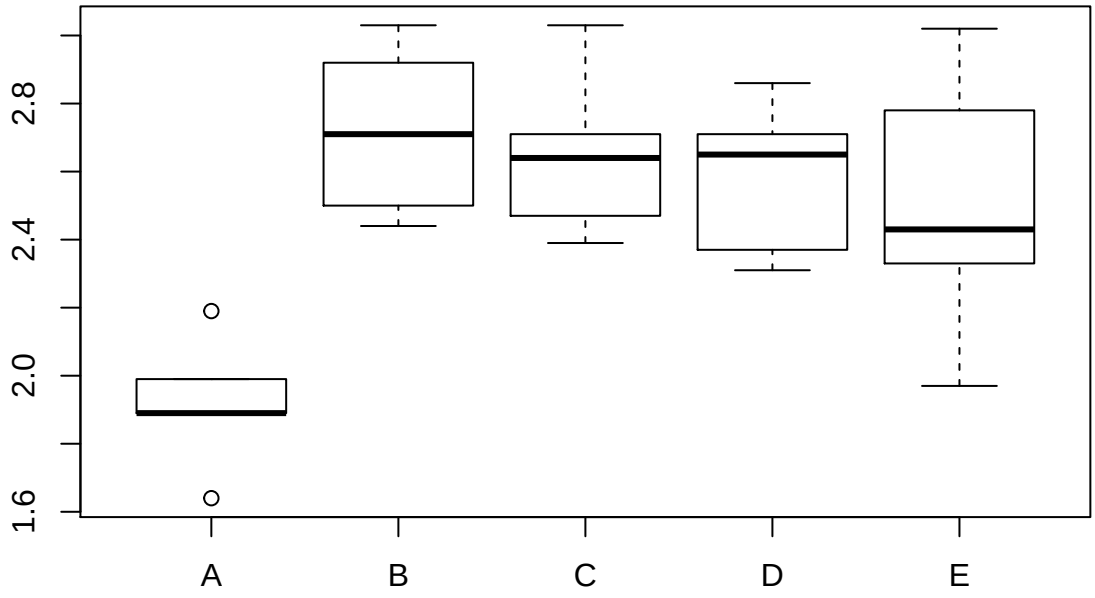
2 (c)

To determine if there is a significant difference between the placebo and the average of the other drugs, we can use the following contrast: $\tau_1 - \frac{1}{4}(\tau_2 + \tau_3 + \tau_4 + \tau_5)$. We can also use boxplots to visualize the data.

```
con1 <- c(1, -1/4, -1/4, -1/4, -1/4)
library(gmodels)
fit.contrast(mod, "treatment", con1)
```

```
##                                           Estimate Std. Error   t value
## treatment c=( 1 -0.25 -0.25 -0.25 -0.25 )  -0.6935  0.1092924 -6.345362
##                                              Pr(>|t|)
## treatment c=( 1 -0.25 -0.25 -0.25 -0.25 ) 3.689679e-05
## attr(,"class")
## [1] "fit_contrast"
```

```
boxplot(response ~ treatment, data = sleep)
```



From both the `fit.contrast` output and the boxplots, we can see that there is a significant difference in

sleep quality between the placebo and the average of the other treatments. In fact, the other treatments significantly increase the patient's quality of sleep.

To determine if there are significant differences between the four drugs, we will have to use the Tukey Honestly Significant Differences methods. If we were only comparing a couple of the drugs, we would be able to continue using contrats; however, if we were to do all of the pairwise combinations, we would quickly run into orthogonality issues. This is demonstrated below:

```
con2 <- c(0, 1, -1, 0, 0)
con3 <- c(0, 1, 0, -1, 0)
con4 <- c(0, 1, 0, 0, -1)
con5 <- c(0, 0, 1, -1, 0)
con6 <- c(0, 0, 1, 0, -1)
con7 <- c(0, 0, 0, 1, -1)
(con1 %*% con2)[1, 1]
```

```
## [1] 0
```

```
(con1 %*% con3)[1, 1]
```

```
## [1] 0
```

```
(con1 %*% con4)[1, 1]
```

```
## [1] 0
```

```
(con1 %*% con5)[1, 1]
```

```
## [1] 0
```

```
(con1 %*% con6)[1, 1]
```

```
## [1] 0
```

```
(con1 %*% con7)[1, 1]
```

```
## [1] 0
```

```
(con2 %*% con3)[1, 1]
```

```
## [1] 1
```

```
(con2 %*% con4)[1, 1]
```

```
## [1] 1
```

```
(con2 %*% con5)[1, 1]
```

```
## [1] -1
```

```
(con2 %*% con6)[1, 1]
```

```
## [1] -1
```

```
(con2 %*% con7)[1, 1]
```

```
## [1] 0
```

```
(con3 %*% con4)[1, 1]
```

```
## [1] 1
```

```
(con3 %*% con5)[1, 1]
```

```
## [1] 1
```

```
(con3 %*% con6)[1, 1]
```

```
## [1] 0
```

```
(con3 %*% con6)[1, 1]
```

```
## [1] 0
```

```
(con4 %*% con5)[1, 1]
```

```
## [1] 0
```

```
(con4 %*% con6)[1, 1]
```

```
## [1] 1
```

```
(con4 %*% con7)[1, 1]
```

```
## [1] 1
```

```
(con5 %*% con6)[1, 1]
```

```
## [1] 1
```

```
(con5 %*% con7)[1, 1]
```

```
## [1] -1
```

```
(con6 %*% con7)[1, 1]
```

```
## [1] 1
```

So, instead, we use the `TukeyHSD` function.

```
TukeyHSD(mod, which = "treatment")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = response ~ patient + week + treatment)
##
## $treatment
##       diff        lwr       upr     p adj
## B-A  0.800  0.3593528 1.2406472 0.0006716
## C-A  0.728  0.2873528 1.1686472 0.0015170
## D-A  0.660  0.2193528 1.1006472 0.0033718
## E-A  0.586  0.1453528 1.0266472 0.0082663
## C-B -0.072 -0.5126472 0.3686472 0.9835041
## D-B -0.140 -0.5806472 0.3006472 0.8447948
## E-B -0.214 -0.6546472 0.2266472 0.5537116
## D-C -0.068 -0.5086472 0.3726472 0.9866453
## E-C -0.142 -0.5826472 0.2986472 0.8382286
## E-D -0.074 -0.5146472 0.3666472 0.9817589
```

From this, we can see that there is a significant difference in sleep quality between the placebo and each of the other treatments. There is no significant difference in sleep quality between any pairs of the drugs. Thus, so long as a patient takes one of the hypnotic drugs (doesn't matter which one), they should, on average, see an increase in their sleep quality.

Problem 3

A wooden catapult can be used to flip a foam ball. The catapult has three factors that can be adjusted: the start angle, the stop angle, and the pivot height. The distance the ball travels can be measured with a tape measure.

(a) If experiments were to be conducted with the catapult by flipping the ball and measuring the distance, what would the experimental unit be?

(b) Using the numbers 1, 2, and 3 to represent the levels of start angle and stop angle, and holding the pivot height constant at its high level, make a randomized list of experiments for a 3 X 3 factorial experiment with $r = 2$ replicates per cell.

Problem 3: Solution

3 (a)

Recall, from Lawson: *Experimental Unit* is the item under study upon which something is changed.

In our case, the item under study / the experimental unit is the distance of the catapulted ball, given the start angle, stop angle, and pivot height.

3 (b)

We are told to keep the pivot height constant, so we are only really varying two factors, each with 3 levels. Thus, we have a 3 x 3 factorial design. We need to use the `expand.grid` function on two vectors which each contain 3 levels, and then we can combine these results with a constant pivot height and a randomized order to obtain our final randomization plan.

```
start <- 1:3
stop <- 1:3
combos <- expand.grid(start, stop)
start <- as.factor(rep(combos$Var1, 2))
stop <- as.factor(rep(combos$Var2, 2))
replicate <- rep(1:2, each = 9)
pivot <- as.factor(rep(3, 18))
order <- sample(1:18, 18, replace = FALSE)
catapult <- data.frame(start, stop, pivot, replicate, order)
kable(catapult, "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = "striped")
```

start	stop	pivot	replicate	order
1	1	3	1	13
2	1	3	1	17
3	1	3	1	1
1	2	3	1	5
2	2	3	1	18
3	2	3	1	2
1	3	3	1	15
2	3	3	1	7
3	3	3	1	12
1	1	3	2	9
2	1	3	2	11
3	1	3	2	16
1	2	3	2	10
2	2	3	2	6
3	2	3	2	8
1	3	3	2	3
2	3	3	2	4
3	3	3	2	14

Problem 6

In an experiment to maximize the Y = resolution of a peak on a gas chromatograph, a significant interaction between A = column temperature and C = gas flow rate was found. The table below shows the mean resolution in each combination of column temperature and gas flow rate.

Column Temp	Flow Low	Flow High
120	10	13
180	12	18

(a) Construct an interaction graph.

(b) Write a sentence or two to interpret this interaction.

Problem 6: Solution

6 (a)

To be honest, the above table isn't set up very well – I can't really tell if those numbers are all supposed to be mean resolutions, or if the "column temp" column is showing the two levels of column temperature. I'm going to assume it's the latter, because it doesn't really make sense otherwise. To make my assumption more clear, this is how I'm interpreting the above table:

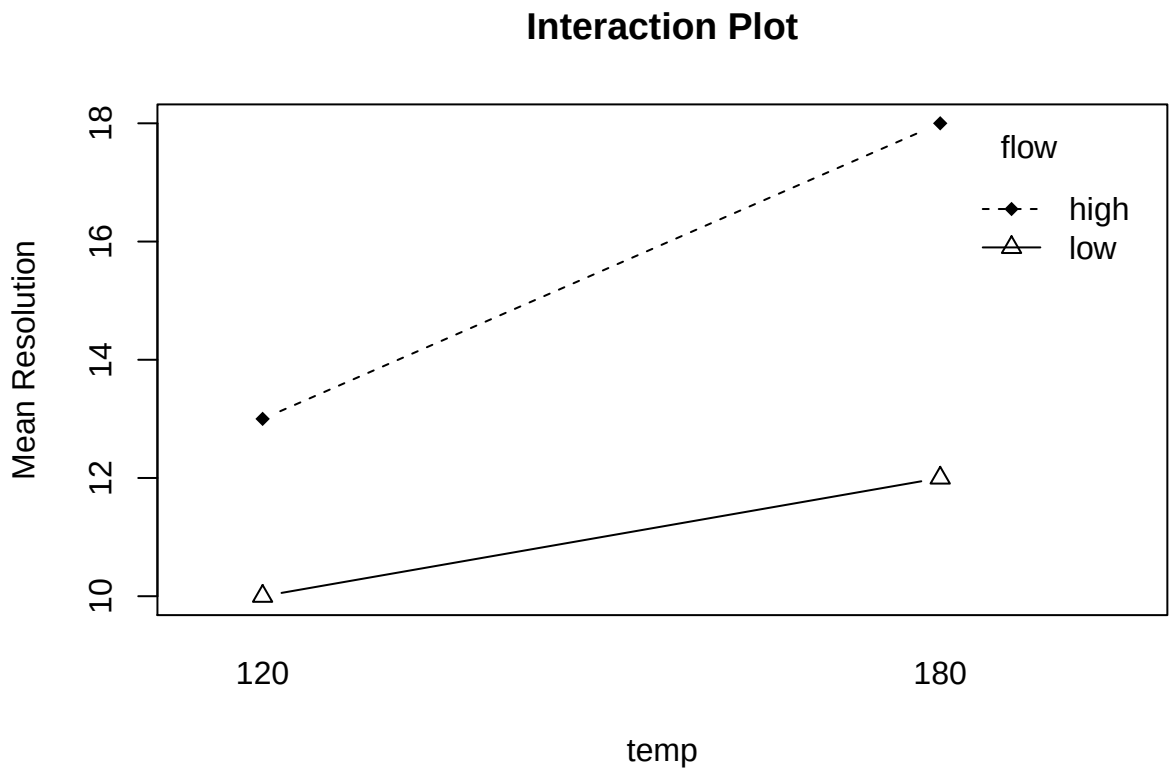
```
temp <- as.factor(c(120, 180, 120, 180))
flow <- as.factor(c(rep("low", 2), rep("high", 2)))
resolution <- c(10, 12, 13, 18)
chromatograph <- data.frame(temp, flow, resolution)
```

```
kable(chromatograph, "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = "striped")
```

temp	flow	resolution
120	low	10
180	low	12
120	high	13
180	high	18

Now we can create the interaction plot:

```
with(chromatograph, (interaction.plot(temp, flow, resolution, type = "b", pch =
c(18,24), main = "Interaction Plot",
xlab = "temp", ylab = "Mean Resolution")))
```




```
## NULL
```

6 (b)

We can see that there is a significant interaction between the `temp` and `flow` variables, since the slopes are not parallel. From the interaction plot, it can be observed that when the gas flow rate is low, changing the column temperature from 120 to 180 causes a small increase (2 units) in the mean resolution of the peak; however, when the gas flow rate is high, changing the column temperature from 120 to 180 causes a much larger increase (5 units) in the mean resolution of the peak on the chromatograph.

I don't really remember everything from physics, but I'm pretty sure that having higher resolution is a good thing, because it's easier to detect where the actual peak is (?). If that's the case, then we can use this interaction plot to recommend that the scientists use a high gas flow rate and high column temperature (if these were the only two variables at play). A quick Google search reveals that it is indeed recommended to use a high gas flow rate / high temperature to improve resolution in gas chromatography experiments.