Lab 7: Analysis of Variance

Complete all of the following questions, adding your inputs as code chunks (enclose within triple accent marks) within Rmarkdown.

The exercises are not marked and will not be factored into your course grade, but it is important to complete them to make sure you have the skills to answer assessment questions. You may consult any resource, including other students and the instructor. Please Knit this document to a PDF and upload your work via Canvas at the end of the session. Solutions will be posted for you to check your own answers.

Simple ANOVA

A baking company is conducting a consumer trial of five competing biscuit recipes. Passersby in a supermarket are given a biscuit and asked to score it on a scale of 1 (worst) to 10 (best).
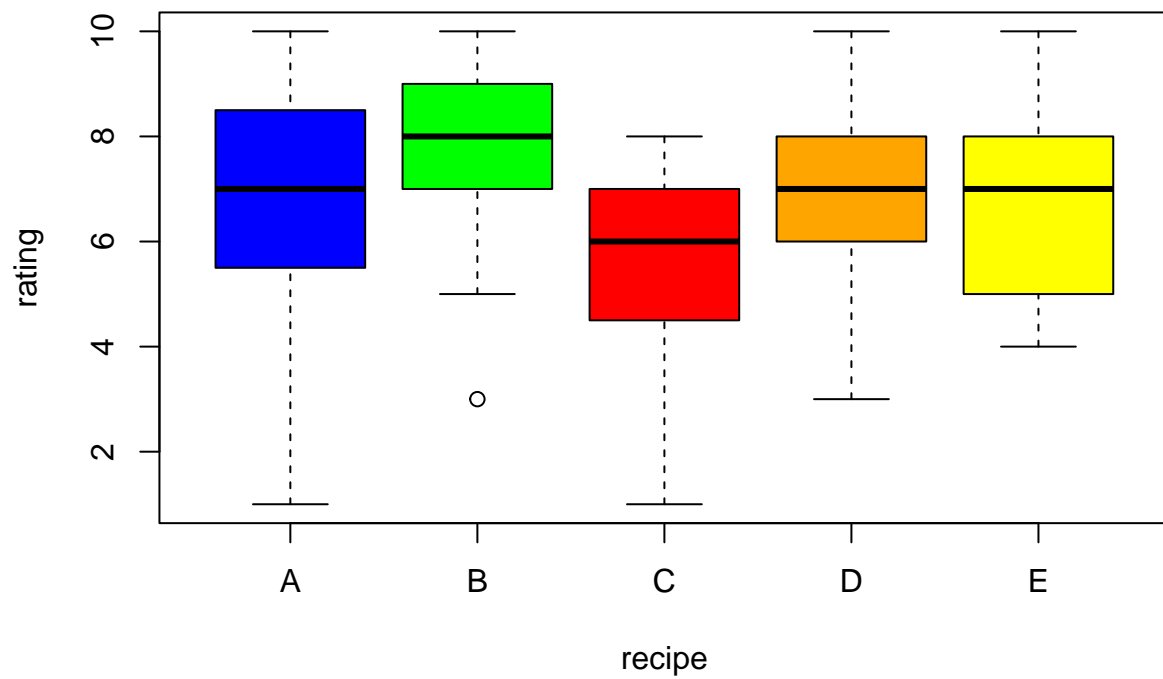
The data are provided in biscuits.csv.

1. Load the data in from disk. Make a box plot, showing the medians and quartiles for each group.

```
data=read.csv('biscuits.csv')
head(data)
```
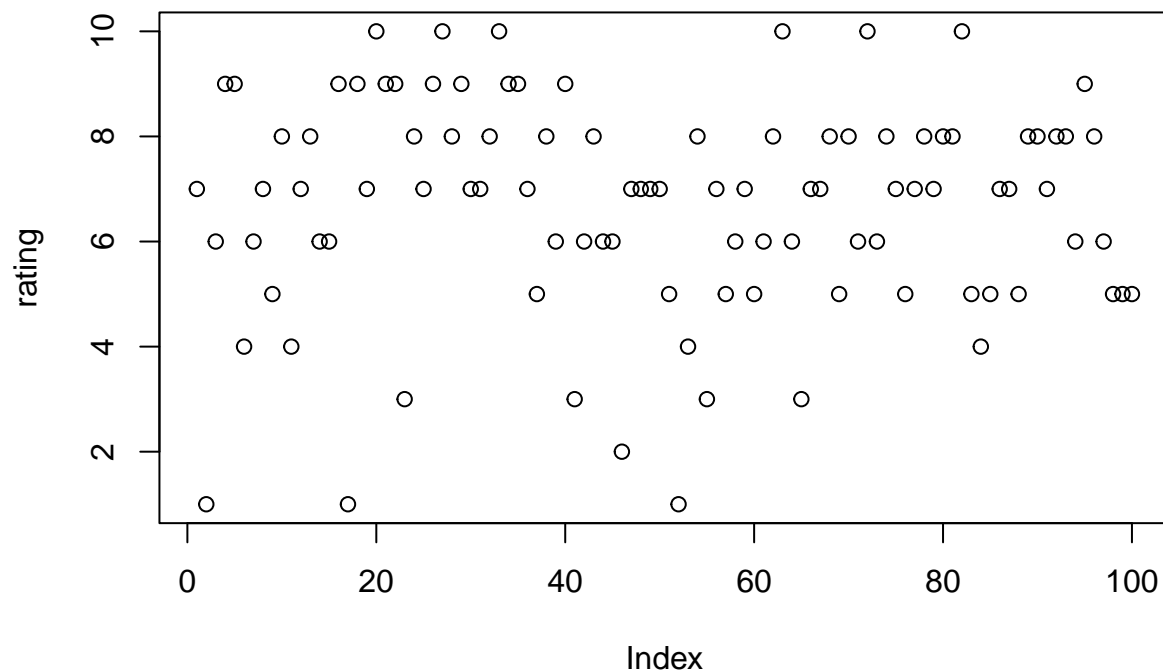
```
##   recipe rating
## 1      A      7
## 2      A      1
## 3      A      6
## 4      A      9
## 5      A      9
## 6      A      4
```

```
attach(data)
boxplot(rating~recipe, col=c('blue','green','red', 'orange', 'yellow'))
```

2. Make a simple scatter plot of all the ratings (no x-axis variable other than the index number). Colour-code the individual points by recipe.

```
plot(rating, pch=21, bg=c('blue','green','red', 'orange', 'yellow')[recipe], ylab='rating')
```

3. Calculate the mean and SD (of ratings) for the whole data set (the "overall" mean and overall SD, sometimes also called the grand mean and grand SD).

```
g.mean = mean(rating)
g.sd=sd(rating)
g.mean
```

```
## [1] 6.7
```

```
g.sd
```

```
## [1] 2.042478
```

4. Calculate all group means, SD's, sizes, and standard errors (use `tapply`).

    a. Which recipe produced the highest mean score?

    b. Which produced the lowest mean score?

```
means = tapply(rating, recipe, mean)
SD = tapply(rating, recipe, sd)

means;SD
```

3

```
##    A    B    C    D    E
## 6.45 7.85 5.50 7.00 6.70
```

```
##        A        B        C        D        E
## 2.502104 1.725200 1.986799 1.654340 1.657519
```

Recipe B produce hoights mean which is 7 and recipe C produced lowest mean whioch is 5.5.

5. Examine the group SD's above, and compare these values to the "overall" SD from #3. Does grouping the data by recipe seem to reduce the variation by much?

```
g.sd
```

```
## [1] 2.042478
```

```
SD
```

```
##        A        B        C        D        E
## 2.502104 1.725200 1.986799 1.654340 1.657519
```

ANOVA formalizes this procedure of comparing the SD (or variance) of all the data together (#3) versus the SD (or variance) of the data grouped by category (#5). It is worked most easily by calculating sums of squares, degrees of freedom, then dividing and performing and F-test.

6. Calculate SSY, the sum of squares of: all data points minus the overall mean. (Use a vector operation.)

```
rate=c(rating)
SSY= sum((rate-g.mean)^2)
```

```
SSY
```

```
## [1] 413
```

7. Calculate SSM, the sum of squares of (replicated) means minus the overall mean. Do this by creating a vector of length equal to the number of data points, then setting all elements in the vector matching the 'A' recipe to the mean of A, then setting all elements matching the 'B' recipe to the mean of B, etc. (You can do this explicitly for each group one by one, or you could use a loop or the rep command.) Then calculate the sum of squares of: this vector minus the grand mean.

```
m = rep(0,nrow(data))
#for (i in 1:length(means)) m[biscuits$recipe==names(means)[i]] = means[i]
#Alternative slower way
m[recipe=='A'] = means['A']
m[recipe=='B'] = means['B']
m[recipe=='C'] = means['C']
m[recipe=='D'] = means['D']
m[recipe=='E'] = means['E']
SSM = sum((m-g.mean)^2)
SSM
```

```
## [1] 58.3
```

8. Calculate SSE, the sum of squares of: all data points minus the appropriate group mean (model point). Use the model vector you made for #7 above.

```
SSE = sum((rating-m)^2)
SSE
```

```
## [1] 354.7
```

9. Confirm that SSM+SSE=SSY.

```
SSM+SSE
```

```
## [1] 413
```

```
SSY
```

```
## [1] 413
```

10. Determine the number of degrees of freedom for SSM and SSE and store these in variables dofM and dofE. The degrees of freedom for SSM is the number of groups minus one. The degrees of freedom for SSE is the number of data points minus the number of groups.

```
dofM = 5-1
dofE = length(rating)-5
dofM
```

```
## [1] 4
```

```
dofE
```

```
## [1] 95
```

11. Calculate the F-ratio, which is (SSM/dofM)/(SSE/dofE). The degrees of freedom are dofM = M-1 and dofE = N-M, where we have M=5 groups and N=100 data points. Then, perform a one-tailed F-test to calculate a p-value. Does the specific recipe have an effect on consumer taste ratings?

```
F = (SSM/dofM)/(SSE/dofE)
F
```

```
## [1] 3.903651
```

```
1-pf(F,dofM,dofE)
```

```
## [1] 0.00559195
```

12. Perform an ANOVA in R using `aov()`. Use the `summary()` command and confirm your results from #7, #8, #10, and #11.

```
m=aov(rating~recipe)
m
```

```
## Call:
##    aov(formula = rating ~ recipe)
##
## Terms:
##                  recipe Residuals
## Sum of Squares     58.3     354.7
## Deg. of Freedom       4        95
##
## Residual standard error: 1.932274
## Estimated effects may be unbalanced
```

13. Check the diagnostic summary plots from your ANOVA. Are there any serious problems?

```
summary(m)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## recipe       4   58.3  14.575   3.904 0.00559 **
## Residuals   95  354.7   3.734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(m)
```
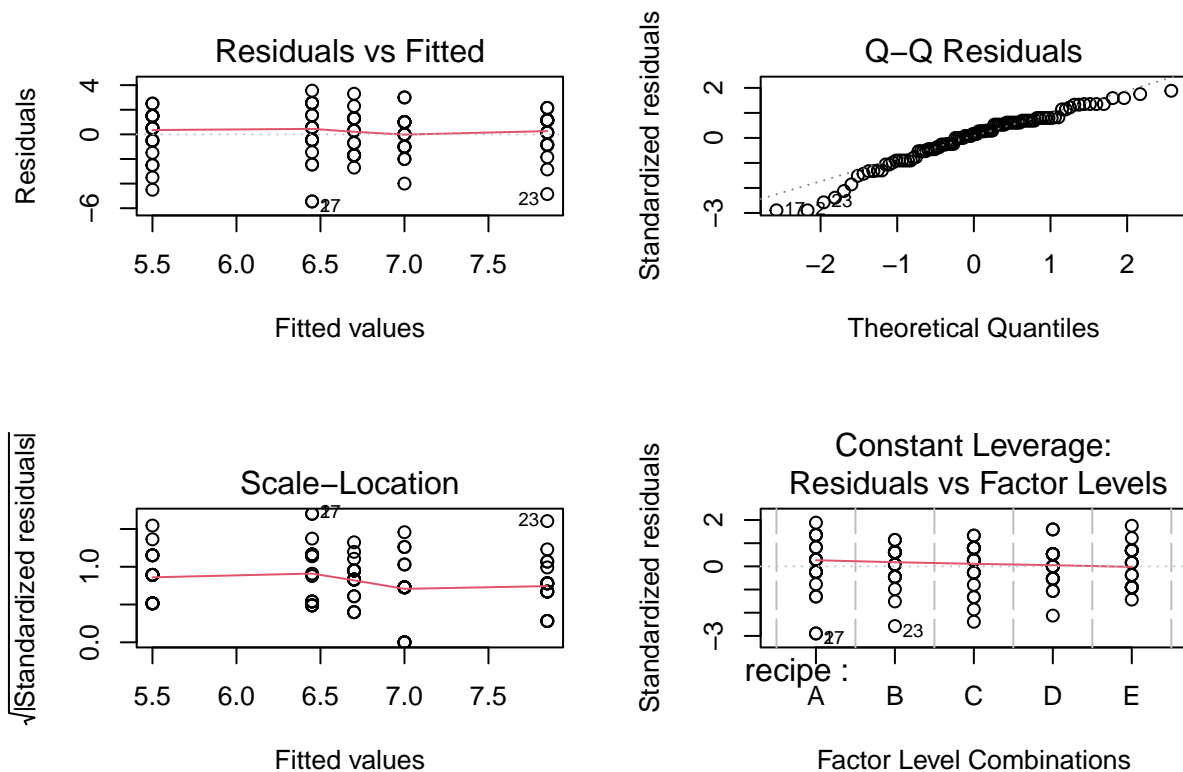
14. Examine R's linear model using `summary.lm()`. Make sure you understand the meaning of all the terms of the table. Which model terms are significant based on the p-values? Does this really mean that there is a significant difference at that level of confidence?

```
summary.lm(m)
```

```
##
## Call:
## aov(formula = rating ~ recipe)
##
## Residuals:
##    Min    1Q Median    3Q    Max
## -5.45  -1.00   0.30  1.30   3.55
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.4500     0.4321  14.928   <2e-16 ***
## recipeB        1.4000     0.6110   2.291   0.0242 *
## recipeC       -0.9500     0.6110  -1.555   0.1233
## recipeD        0.5500     0.6110   0.900   0.3703
## recipeE        0.2500     0.6110   0.409   0.6834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.932 on 95 degrees of freedom
## Multiple R-squared:  0.1412, Adjusted R-squared:  0.105
## F-statistic: 3.904 on 4 and 95 DF,  p-value: 0.005592
```

15. Perform Tukey's Honestly Significant Difference test with `TukeyHSD()`. What pairs of recipes are "definitely" significantly different?

```
TukeyHSD(m)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rating ~ recipe)
##
## $recipe
##       diff        lwr        upr     p adj
## B-A   1.40 -0.2992162  3.0992162 0.1568021
## C-A  -0.95 -2.6492162  0.7492162 0.5299339
## D-A   0.55 -1.1492162  2.2492162 0.8961773
## E-A   0.25 -1.4492162  1.9492162 0.9940211
## C-B  -2.35 -4.0492162 -0.6507838 0.0019871
## D-B  -0.85 -2.5492162  0.8492162 0.6348610
## E-B  -1.15 -2.8492162  0.5492162 0.3340465
## D-C   1.50 -0.1992162  3.1992162 0.1099831
## E-C   1.20 -0.4992162  2.8992162 0.2916446
## E-D  -0.30 -1.9992162  1.3992162 0.9880307
```

16. Suppose that A is the "current" recipe on the market and the company is considering switching the recipe to B, C, D, or E. Apply Bonferroni's correction to the p-values in #14 to assess if any of these four recipes are significantly better than the current one.

```
pairwise.t.test(rating, recipe, p.adjust.method="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  rating and recipe
##
##   A      B      C      D
## B 0.2416 -      -      -
## C 1.0000 0.0022 -      -
## D 1.0000 1.0000 0.1591 -
## E 1.0000 0.6289 0.5247 1.0000
##
## P value adjustment method: bonferroni
```

Two-way ANOVA

"Stereotype threat" is a concept in the psychological literature which states that expectations about one's group can play a significant role in performance, particularly on examinations or other formal tasks. According to stereotype threat, groups which are socially perceived to be weaker on certain tasks actually perform worse as a direct result of latent anxiety about conforming with those stereotypes. It has been suggested to be a possible explanation for performance gaps between genders and races on standardized examinations.

The data in stereotype.csv represents a simulated version of a classical type of stereotype threat experiment. Students are randomly assigned to four classrooms, and then given a mathematics exam. At the start, an announcement is made:

Control classroom: Only instructions about the exam.
Classroom A: Students are told that men and women tend to do equally well on the exam.
Classroom B: Students are told that men tend to do better than women on the exam.
Classroom C: Students are told that women tend to do better than men on the exam.

17. Load the data in from disk. Plot a histogram of the distribution of all scores (no grouping) and/or make a QQ normal plot to confirm that the score distribution is roughly normal in appearance.

```
stero=read.csv('stereotype.csv')
hist(stero$score, col='lightblue', main='Scores')
```

Scores
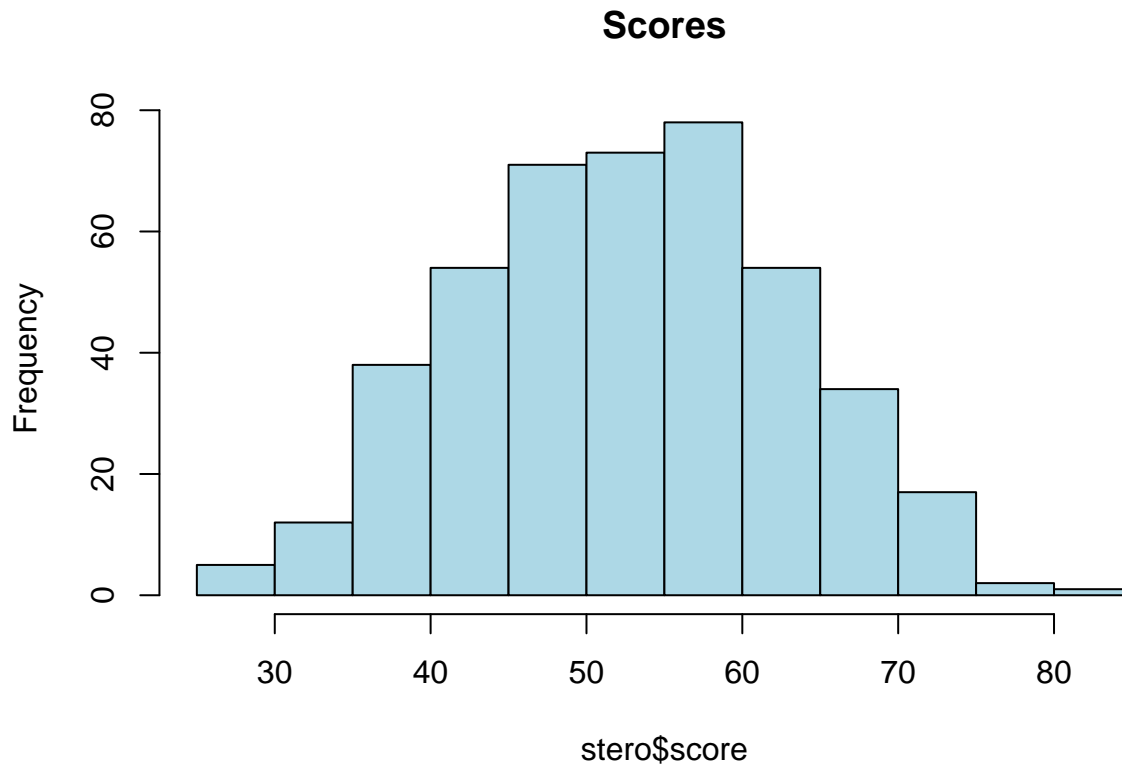
18. Calculate the means and standard errors for all eight groups (four announcement factor levels × two gender factor levels). You can save yourself a lot of hassle by using the 2D version of `tapply()` to calculate the means, lengths, and SD's of each bivariate group and using matrix operations on the resulting tables to obtain the SE.

```
summary(stero)
```

```
##   announcement         gender             score
##   Length:439       Length:439        Min.   :27.00
##   Class :character   Class :character   1st Qu.:46.00
##   Mode  :character   Mode  :character   Median :53.00
##                                         Mean   :53.01
##                                         3rd Qu.:60.00
##                                         Max.   :85.00
```

```
means=tapply(stero$score, list(stero$gender, stero$announcement), mean)
```

```
means
```

```
##          A        B        C     none
## F 52.58824 48.24138 54.77083 49.33962
## M 53.94915 55.82353 55.01493 54.48077
```

```r
sds=tapply(stero$score, list(stero$gender, stero$announcement), sd)

sds
```

```
##           A         B         C      none
## F 11.246647  9.942484 11.231111  9.042514
## M  9.337399 10.487528  9.279323 10.104339
```

```r
n= tapply(stero$score,list(stero$gender,stero$announcement),length)
n
```

```
##     A  B  C none
## F 51 58 48   53
## M 59 51 67   52
```
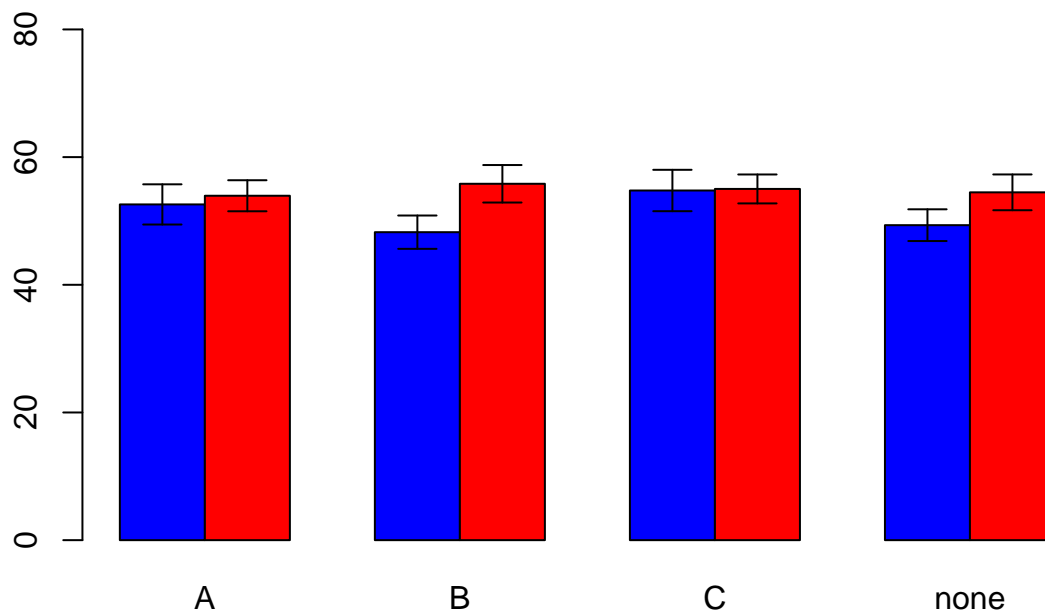
```r
se=sds/sqrt(n)
se
```

```
##          A        B        C     none
## F 1.574846 1.305512 1.621071 1.242085
## M 1.215626 1.468548 1.133650 1.401220
```

19. Summarize the above as a plot with error bars. Use 2*SE for the error bars in each direction (approximately 95% confidence interval).

```r
loc=barplot(means, beside = TRUE, ylim=c(0,80), col = c("blue", "red"))
arrows(loc,means-se*2,loc,means+se*2,angle=90,code=3,length=0.1)
```

20. Now run a two-way ANOVA on the data, including interactions.

```
m= aov(stero$score~stero$gender*stero$announcement)
```

21. Examine the output from `summary()`. Are there significant interactions? What does an interaction mean in the context of this experiment?

```
summary(m)
```

```
##                                Df Sum Sq Mean Sq F value   Pr(>F)
## stero$gender                    1   1526  1526.1  15.091 0.000119 ***
## stero$announcement              3    558   186.0   1.839 0.139351
## stero$gender:stero$announcement 3    944   314.8   3.113 0.026152 *
## Residuals                     431  43586   101.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

22. Does the fact that the "announcement" term is not significant in the ANOVA mean that the announcement had no significant effect?

23. Now examine the `summary.lm()` output of the ANOVA. Try to make sure you understand what all of the numbers in the primary table mean in context of the experiment ("Estimate", "Std. Error", and the associated P-value). Do you understand why the t-test significance of the specific interaction terms is different from the F-test significance from `summary.aov()` of the interaction in general?

11

```
summary.lm(m)
```

```
##
## Call:
## aov(formula = stero$score ~ stero$gender * stero$announcement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.015  -7.481   0.051   6.905  32.412
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           52.588      1.408  37.346   <2e-16 ***
## stero$genderM                          1.361      1.923   0.708   0.4794
## stero$announcementB                   -4.347      1.930  -2.252   0.0248 *
## stero$announcementC                    2.183      2.022   1.079   0.2811
## stero$announcementnone                -3.249      1.973  -1.647   0.1003
## stero$genderM:stero$announcementB      6.221      2.725   2.283   0.0229 *
## stero$genderM:stero$announcementC     -1.117      2.704  -0.413   0.6798
## stero$genderM:stero$announcementnone   3.780      2.748   1.376   0.1696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 431 degrees of freedom
## Multiple R-squared:  0.06497,    Adjusted R-squared:  0.04978
## F-statistic: 4.278 on 7 and 431 DF,  p-value: 0.0001377
```

24. Interpreting the model terms above is challenging due to R's alphabetical choice of 'first' factor levels. It would be more useful to calculate differences of the experimental groups against the control, and (in the context of stereotype threat theory) to calculate differences of women versus men. Reorder the factor levels to account for this: set the order of the announcement factors to none,A,B,C, and set the order of the gender factors to M,F. Check to confirm that the re-ordering was successful.

```
styp=stero
styp$announcement= factor(styp$announcement, levels = c('none', 'A', 'B', 'C'))
styp$gender=factor(styp$gender, levels = c('M', 'F'))
levels(styp$announcement )
```

```
## [1] "none" "A"    "B"    "C"
```

```
levels(styp$gender)
```

```
## [1] "M" "F"
```

25. Rerun the ANOVA on the re-ordered data. Confirm that the `summary.aov()` output (F-test significances) does not change, then confirm that the linear-model terms from `summary.lm()` do change. Also inspect the model check plots with `plot()`.

```
m1=aov(score~gender*announcement, styp)
summary.aov(m1)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## gender              1   1526  1526.1  15.091 0.000119 ***
## announcement        3    558   186.0   1.839 0.139351
## gender:announcement 3    944   314.8   3.113 0.026152 *
## Residuals         431  43586   101.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(m1)
```

```
##
## Call:
## aov(formula = score ~ gender * announcement, data = styp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.015  -7.481   0.051   6.905  32.412
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           54.4808     1.3945  39.067  < 2e-16 ***
## genderF               -5.1411     1.9629  -2.619  0.00912 **
## announcementA         -0.5316     1.9128  -0.278  0.78120
## announcementB          1.3428     1.9818   0.678  0.49843
## announcementC          0.5342     1.8585   0.287  0.77394
## genderF:announcementA  3.7802     2.7477   1.376  0.16960
## genderF:announcementB -2.4410     2.7530  -0.887  0.37576
## genderF:announcementC  4.8971     2.7329   1.792  0.07386 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 431 degrees of freedom
## Multiple R-squared:  0.06497,    Adjusted R-squared:  0.04978
## F-statistic: 4.278 on 7 and 431 DF,  p-value: 0.0001377
```

```
par(mfrow=c(2,2))
plot(m1)
```
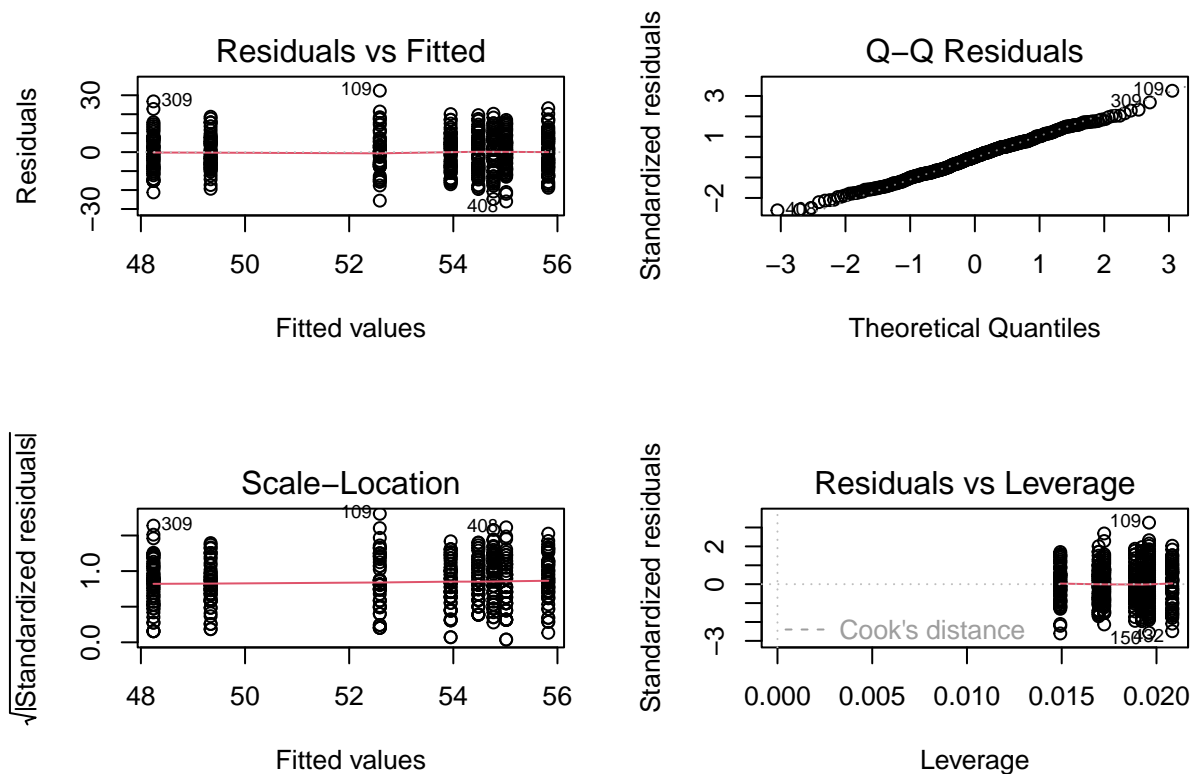
26. Do these data support the notion of stereotype threat? Are they able to constrain what fraction of the gender gap on this exam might be due to stereotype threat? *(Note: the data is simulated; conclusions about the real world should not be drawn from this exercise!)*

ANOVA with structured error

The data in bloodpressure.csv come from a hypothetical simulation of a medical trial for a new blood pressure drug. In the trial, 16 patients with high blood pressure were given the new drug. Of these, 8 patients were already on another blood pressure medication and 8 patients were not on another medication. The blood pressure of each patient was measured at the start of the trial, at 1 month, and again at 2 months. The variables in the data frame are:

id: Patient ID number
time: time elapsed in the trial (start, 1 month, or 2 months)
othermed: Whether this patient was on any other blood pressure medication.
pressure: Blood pressure measurement

27. Load in the data and run a "naive" ANOVA with no knowledge of the error structure. Would you conclude that the blood pressure medication is effective (does it cause a decrease in blood pressure over time)? Would you conclude that its effectiveness depends on whether the patient was on medication already?

```
df=read.csv('bloodpressure.csv', as.is = FALSE)
head(df)
```

```
##   id pressure    time othermed
## 1  1 139.7353   start      yes
## 2  1 134.6761  1month      yes
## 3  1 140.7460 2months      yes
## 4  2 141.8482   start      yes
## 5  2 144.3599  1month      yes
## 6  2 142.8481 2months      yes
```

```
naive=aov(df$pressure~df$time*df$othermed, data=df)
naive
```

```
## Call:
##    aov(formula = df$pressure ~ df$time * df$othermed, data = df)
##
## Terms:
##                  df$time df$othermed df$time:df$othermed Residuals
## Sum of Squares   360.0126    281.3201            153.4013 2858.3585
## Deg. of Freedom         2           1                   2        42
##
## Residual standard error: 8.249615
## Estimated effects may be unbalanced
```

28. Now run a new ANOVA, accounting for the error structure. (Hint: Which variable is between subjects, and which is within subjects? Also, be sure you convert the ID number variable to a factor!) How does this affect your conclusions about the effectiveness and medication-dependence of the new drug?

```
df$id= as.factor(df$id)
new_anova=aov(pressure~othermed*time+Error(id/time),data=df)

summary(new_anova)
```

```
##
## Error: id
##             Df Sum Sq Mean Sq F value Pr(>F)
## othermed     1  281.3   281.3   1.601  0.226
## Residuals   14 2460.3   175.7
##
## Error: id:time
##               Df Sum Sq Mean Sq F value   Pr(>F)
## time           2  360.0  180.01  12.662 0.000121 ***
## othermed:time  2  153.4   76.70   5.395 0.010424 *
## Residuals     28  398.0   14.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

29. Make a summary plot using `interaction.plot()` to visualise the effects. Make time the x-axis variable (you may want to re-order the factor levels) and the medication style the line-style variable. Is the new drug more effective for people already taking medication, or for people not already taking medication?

```
time_order= factor(df$time, levels = c('start', '1month', '2months'))
interaction.plot(time_order, df$othermed, df$pressure)
```