Quiz 1 Bonus Assignment

Thomas E. Love

due at noon Monday 2018-03-26. Version 2018-03-10

Contents

Instructions

1. You will submit your R Markdown and HTML result for this assignment to Canvas.
2. Your completed work is due at noon on Monday 2018-03-26, although you are welcome to submit it earlier.
3. If you scored 75 or higher on Quiz 1, you are not required to do this bonus work. If you scored below 75, then you are required to do it.
4. A passing grade on this assignment is 110/130. If you score 110 or more points on this assignment, I will increase your grade on Quiz 1 half of the distance between your current grade and a score of 85.
5. You will get two chances at this if you make the deadline. If you submit the assignment on time and score below 110, we will return it to you indicating the questions where you lost credit, and let you redo it once (with a 48-hour turnaround.) If you score **115** or more points in this revision, then I will increase your grade on Quiz 1 as indicated above.
6. You can discuss this assignment with the teaching assistants and with Dr. Love, and no one else, up through the Friday before it is due. You are permitted to ask questions through 431-help or in person.

Question 1. (5 points)

Questions 1-5 involve data from a randomized controlled trial compares four potential treatments for the repair of a valve. You will find a copy of the data in the `dat01.csv` file. The data set includes a variable called `treat` which contains the values 0, 1, 2, 3, and which needs to be recoded as a factor in R. Provide the code to import the `dat01.csv` file into a tibble called `dat01` in R, and then create a factor within `dat01` named `treatment` that has the following properties:

1. `treat` = 0 refers to `treatment` = UC
2. `treat` = 1 is captured as `treatment` = A1
3. `treat` = 2 refers to `treatment` = A2
4. `treat` = 3 refers to `treatment` = B
5. the factor `treatment` is ordered A1, A2, B, UC.

Hint: You will know you have completed the task successfully when you obtain the following output:

```
dat01 %>% count(treatment)
```

```
# A tibble: 4 x 2
  treatment     n
  <fct>     <int>
1 A1           30
2 A2           30
3 B            30
4 UC           30
```

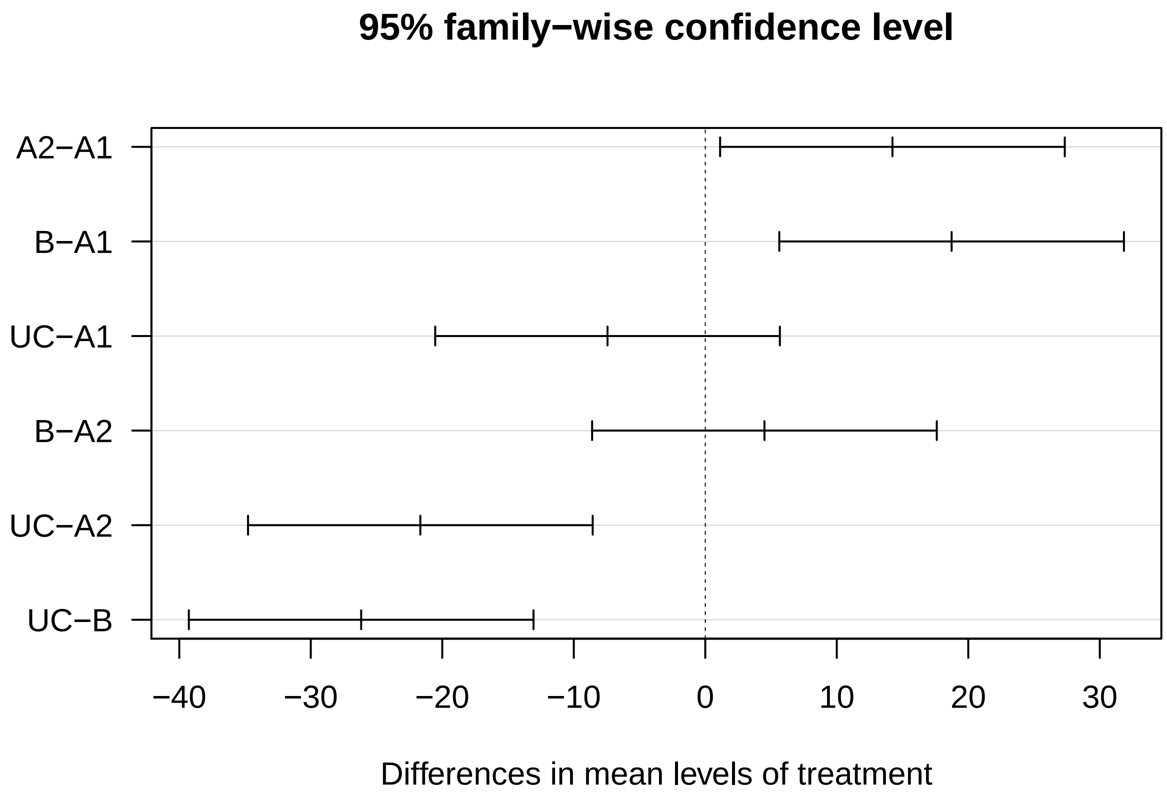
Question 2. (5 points)

Question 2 involves the results of a randomized controlled trial comparing four potential treatments for the repair of a valve, which in question 1 you labeled as `A1`, `A2`, `B` and `UC`. `A1` and `A2` are two versions of the "A" approach, `B` is a separate approach, and `UC` describes the standard approach (or "usual care".) The outcome of interest is `time_to_repair` = survival time before another repair of the same valve is required, which we'd like to be as large as possible.

Your job is to use the data provided to you to produce the Target Plot shown below, which describes the results of Tukey's HSD comparison, with a 95% confidence level. Your response should include the code to produce the Tukey HSD plot, working from the data frame you developed in Question 1.

Note: to get the labels to show up on the y axis in this fashion rather than flipped 90 degrees, you'll have to add the command `las = 1` to your plot.

The Target Plot for Question 2 is shown below

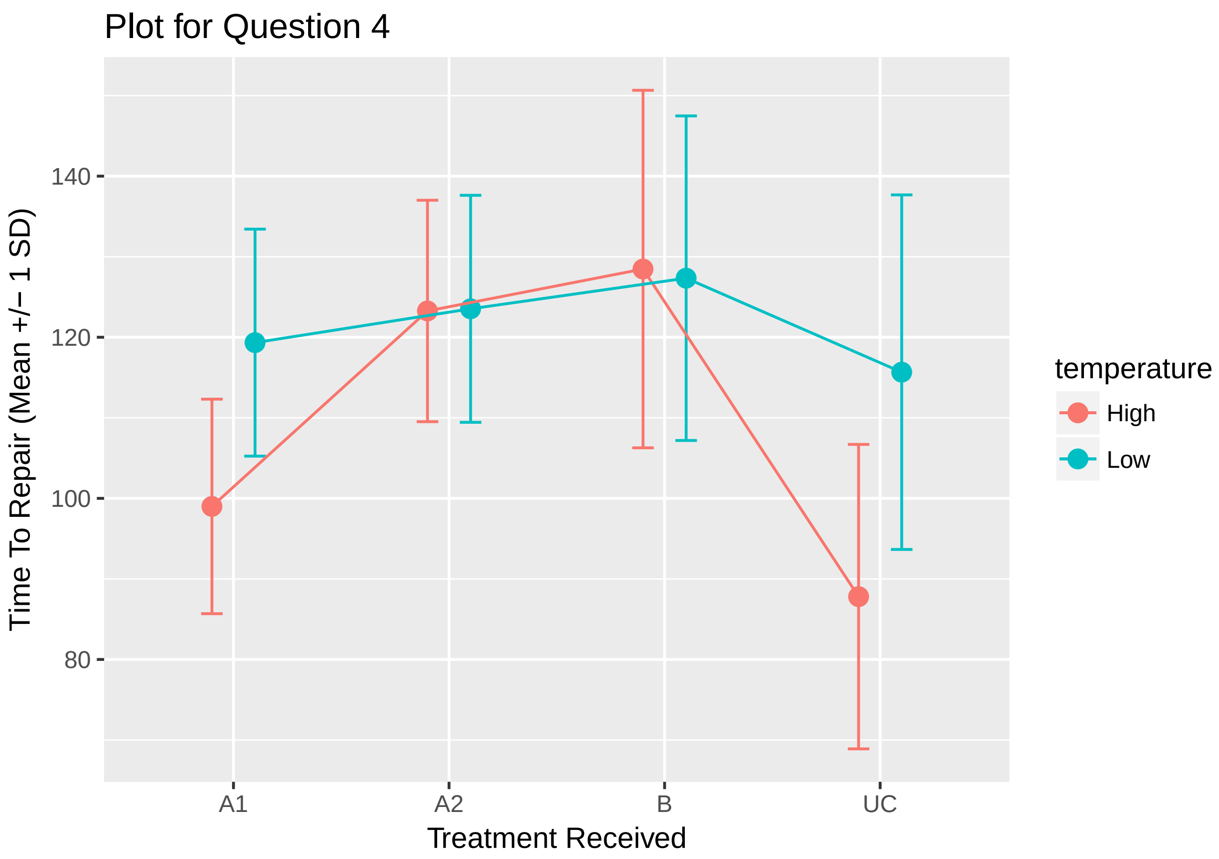


Question 3. (10 points)

Refer to the Target Plot from Question 2 and describe, in complete English sentences, the pairwise comparisons of the four potential treatments. What conclusions can you draw, in terms of statistical significance? Don't forget to clearly indicate for each comparison whether one treatment appears to perform better or worse than another in terms of time to repair.

Question 4. (5 points)

The `dat04.csv` file provided to you specifies the subjects (that we have been studying in Questions 1-3) as well as their temperature (High or Low) at the time of the valve repair. I merged those data with the data in my `dat01` data frame to produce the plot for Question 4 shown below.



In a few sentences, describe what this plot tells you about the relationships between time to repair and the treatment received and the temperature. In particular, what does this plot tell you about how to fit an analysis of variance model?

Question 5. (10 points)

Merge together the `dat01` data frame you've developed through Questions 1-3 with this new `temperature` information from Question 4.

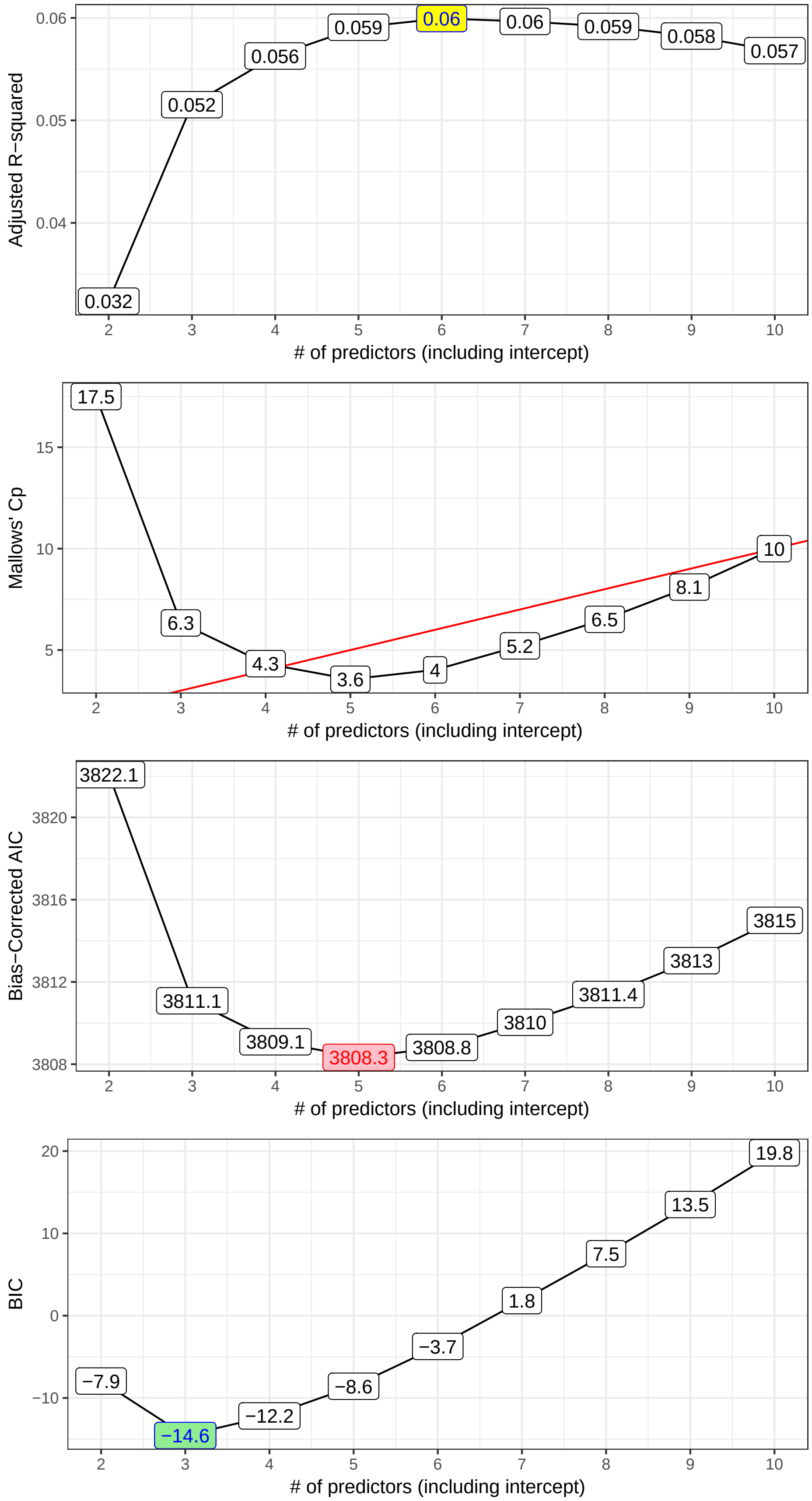
Hint: import the `dat04` data as a tibble then use the `left_join` command to merge the two tibbles you have (`dat01` and `dat04`) into one.

Next, obtain an appropriate two-way analysis of variance evaluating the impact of treatment and temperature on time. There are two possible two-way ANOVA models you could fit, one with interaction and one without. Display the ANOVA table for the model you select, specify why you chose that model, and then specify the conclusions of that table in English sentences.

Question 6. (10 points)

The `dat06.csv` data set contains information on 800 observations of an outcome variable, `y`, and nine predictors, labeled `x1` through `x9`, plus a `sample` variable, which takes the value "development" for 600 observations and the value "testing" for the remaining 200 observations.

The four plots and the table below were obtained by running a "best subsets" procedure to identify models from the complete set of available predictors (the `x`s) to predict `y`, but including only the data in `dat06` in the “development” `sample`.


Adjusted R−squared
0.06
0.05
0.04
0.032
0.052
0.056
0.059
0.06
0.06
0.059
0.058
0.057
2
3
4
5
6
7
8
9
10
# of predictors (including intercept)
Mallows' Cp
15
10
5
17.5
6.3
4.3
3.6
4
5.2
6.5
8.1
10
2
3
4
5
6
7
8
9
10
# of predictors (including intercept)
Bias−Corrected AIC
3820
3816
3812
3808
3822.1
3811.1
3809.1
3808.3
3808.8
3810
3811.4
3813
3815
2
3
4
5
6
7
8
9
10
# of predictors (including intercept)
BIC
20
10
0
−10
−7.9
−14.6
−12.2
−8.6
−3.7
1.8
7.5
13.5
19.8
2
3
4
5
6
7
8
9
10
# of predictors (including intercept)

```r
rs$which
```

```
  (Intercept)    x1    x2    x3    x4    x5    x6    x7    x8    x9
1        TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
2        TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
3        TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
4        TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE
5        TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE
6        TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
7        TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
8        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
9        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Specify the models suggested by each of the four plots, by identifying the predictors included in those models. Be sure to indicate which plot specifies which candidate model clearly, and use the data in the "development" sample to specify the prediction equation obtained from least squares regression for each model. If two plots specify the same model, indicate that.

Question 7. (15 points)

Now that you have fit the models suggested by each of the four plots within the "development" `sample` in Question 6, in Question 7 you will now use those models to make predictions for every value in the "testing" `sample`. Compare the four models in terms of their root mean squared prediction error and their mean absolute prediction error. Write a sentence or two to explain the results and pick a "winning" model.

Question 8. (5 points for each part, so 10 points total)

Fit the model that performed best in Question 7 to the full data set (all 800 observations.)

a. What is the percentage of variation in `y` accounted for by this model in these 800 observations?
b. Estimate the percentage of variation in `y` that will be accounted for by this model when applied to a new, similar data set. Show what you did to obtain this estimate.

Question 9. (10 points)

Assess the regression assumptions of the model that you fit in Question 8. Show the usual four residual and diagnostic plots of the regression model after fitting it to the entire sample of 800 observations, and interpret the results.

Question 10. (5 points for each part, so 15 points total)

The data in the `dat10.csv` file contain information for 250 subjects on a binary `outcome` (Good or Bad), a `size` (quantitative, between 60 and 200), a logical indicator of whether a `treatment` was used (TRUE = treatment was used or FALSE = treatment was not used), and a specification as to which of five ordered groups (1 = lowest, 5 = highest) by socio-economic status (`ses_group`) the subject falls in, along with a subject ID.

Import the data into the `dat10` frame, and then fit a logistic regression model to predict the log odds of a Good `outcome` using the subject's `size`, `treatment` status and `ses_group`, treating the `ses_group` as a categorical variable. For Questions 10 and 11, use a complete case analysis. A complete case analyses yields the results specified below.

```
dat10 <- read.csv("dat10.csv") %>% tbl_df

dat10 <- dat10 %>%
    mutate(goodoutcome = ifelse(outcome == "Good", 1, 0))
```

Note that the fitting of the actual m1 is not shown.

```
exp(coef(m1))
```

```
       (Intercept)               size      treatmentTRUE 
         0.1585338          1.0090549          0.5728293 
factor(ses_group)2 factor(ses_group)3 factor(ses_group)4 
         1.3569411          1.4220772          1.2805544 
factor(ses_group)5 
         1.4752464 
```

```
exp(confint(m1))
```

```
                        2.5 %    97.5 %
(Intercept)        0.03042517 0.7635905
size               0.99754375 1.0209696
treatmentTRUE      0.32260266 1.0142938
factor(ses_group)2 0.42226765 4.4921913
factor(ses_group)3 0.52570487 4.1528045
factor(ses_group)4 0.43702393 3.9643440
factor(ses_group)5 0.56280890 4.2102848
```

a. Specify the code used to fit the m1 model in this setting. Use the data to verify that your answer is correct.

b. Interpret the treatmentTRUE value of 0.573 specified in the output above, in a complete English sentence.

c. What does the confidence interval (0.323, 1.014) for treatmentTRUE tell you about the treatment variable? Why?

Question 11. (5 points for each part, so 10 points total)

The output below comes from another approach to fitting the identical logistic regression model that we saw in Question 10, still using only the complete cases. I'll call this model m1L, to emphasize that it contains the same outcome and predictors, put together in the same way.

```
summary(m1L)
```

```
             Effects              Response : goodoutcome 

 Factor          Low   High  Diff. Effect    S.E.    Lower 0.95 Upper 0.95
 size            100.5 135.5 35     0.315490 0.20642 -0.089079  0.720070  
  Odds Ratio     100.5 135.5 35     1.370900      NA  0.914770  2.054600  
 treatment         0.0   1.0  1    -0.557170 0.29154 -1.128600  0.014247  
  Odds Ratio       0.0   1.0  1     0.572830      NA  0.323490  1.014300  
 ses_group - 1:3   3.0   1.0 NA    -0.352120 0.52128 -1.373800  0.669570  
  Odds Ratio       3.0   1.0 NA     0.703200      NA  0.253140  1.953400  
 ses_group - 2:3   3.0   2.0 NA    -0.046886 0.49511 -1.017300  0.923510  
```

```
  Odds Ratio       3.0   2.0 NA      0.954200      NA  0.361580  2.518100
 ses_group - 4:3   3.0   4.0 NA     -0.104830 0.44313 -0.973350  0.763700
  Odds Ratio       3.0   4.0 NA      0.900480      NA  0.377810  2.146200
 ses_group - 5:3   3.0   5.0 NA      0.036706 0.38085 -0.709750  0.783170
  Odds Ratio       3.0   5.0 NA      1.037400      NA  0.491770  2.188400
```

a. What do you conclude from this summary about the odds ratio and confidence interval associated with the `size` variable?

b. Why is the odds ratio shown in this output for `size` different from that shown in the earlier presentation of `exp(coef(m1))` for the same model?

Question 12. (5 points)

Using the data from Question 10, obtain a Spearman ρ^2 plot and use it to identify a good way to add a single additional non-linear term to this model (you may spend only a single additional degree of freedom). What addition would you make?

Question 13. (5 points)

Using the data from Question 10, identify how many subjects are missing data in at least one variable. Are there any missing outcome data? How do you know?

Question 14. (15 points)

Using the data from Question 10, fit the model you specified in Question 12 (including the non-linear term), while also accounting for missing data using **multiple imputation**. Set your seed to be `432432`, and impute the predictors that need imputation using all available observations on all available variables.

Specify the code you used to fit your imputation model and your outcome model, specify the resulting fitted outcome model in terms of odds ratio estimates and a nomogram, and then write a few English sentences describing how the addition of imputation and a non-linear term changes (or doesn't change) the conclusions from what we saw in the `m1` model in Questions 10 and 11.