Answer Sketch for 432 Quiz 1 Honors Opportunity

Thomas E. Love

Version 2019-04-02

```
library(janitor)
library(leaps)
library(rms)
library(broom)
library(tidyverse)
```

Setup for Questions 1-3

The data in the `honors.csv` file contain information for 255 subjects on:

- a binary `outcome` (Good or Bad),
- a `size` (quantitative, between 60 and 200, in millimeters),
- an indicator of whether a `treatment` was used (1 = treatment was used or 0 = 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 `honors` 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. Questions 1-3 use a **complete case** analysis. One such analysis yields these results.

```
honors <- read_csv("honors.csv") %>% clean_names() %>%
  mutate(goodoutcome = ifelse(outcome == "Good", 1, 0),
         ses_group = factor(ses_group))
```

Here is the actual `m1` model that was fit:

```
m1 <- glm(goodoutcome ~ size + treatment + ses_group, data = honors, family = binomial)
```

```
tidy(m1, conf.int = TRUE, conf.level = 0.95,
     exponentiate = TRUE) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.155	0.822	-2.267	0.023	0.030	0.751
size	1.010	0.006	1.677	0.093	0.998	1.022
treatment	0.557	0.291	-2.010	0.044	0.314	0.985
ses_group2	1.226	0.601	0.339	0.735	0.378	4.087
ses_group3	1.384	0.523	0.622	0.534	0.510	4.051
ses_group4	1.110	0.558	0.188	0.851	0.378	3.442
ses_group5	1.416	0.511	0.680	0.496	0.535	4.068

The output below comes from another approach to fitting the identical logistic regression model that we saw previously, still using only the complete cases. I'll call this model `m1a`, to emphasize that it contains the same outcome and predictors, put together in the same way. Here is what was fit in `m1a`:

```
d <- datadist(honors)
options(datadist = "d")
```

```
m1a <- lrm(goodoutcome ~ size + treatment + ses_group, data = honors)

m1a
```

```
Frequencies of Missing Values Due to Each Variable
goodoutcome        size   treatment   ses_group
          0           5           4          14

Logistic Regression Model

 lrm(formula = goodoutcome ~ size + treatment + ses_group, data = honors)


                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
 Obs           233    LR chi2      8.34    R2        0.049    C       0.614
  0            160    d.f.            6    g         0.479    Dxy     0.228
  1             73    Pr(> chi2) 0.2143    gr        1.615    gamma   0.229
 max |deriv| 3e-11                         gp        0.100    tau-a   0.099
                                           Brier     0.208

             Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept   -1.8628 0.8216 -2.27  0.0234
 size         0.0099 0.0059  1.68  0.0935
 treatment   -0.5852 0.2911 -2.01  0.0444
 ses_group=2  0.2035 0.6007  0.34  0.7348
 ses_group=3  0.3251 0.5226  0.62  0.5339
 ses_group=4  0.1047 0.5576  0.19  0.8511
 ses_group=5  0.3479 0.5115  0.68  0.4964
```

```
summary(m1a) %>% knitr::kable(digits = 3)
```

	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95	Type
size	100.525	135.5	34.975	0.346	0.207	-0.058	0.751	1
Odds Ratio	100.525	135.5	34.975	1.414	NA	0.943	2.120	2
treatment	0.000	1.0	1.000	-0.585	0.291	-1.156	-0.015	1
Odds Ratio	0.000	1.0	1.000	0.557	NA	0.315	0.986	2
ses_group - 1:5	5.000	1.0	NA	-0.348	0.511	-1.350	0.655	1
Odds Ratio	5.000	1.0	NA	0.706	NA	0.259	1.924	2
ses_group - 2:5	5.000	2.0	NA	-0.144	0.481	-1.086	0.798	1
Odds Ratio	5.000	2.0	NA	0.866	NA	0.337	2.220	2
ses_group - 3:5	5.000	3.0	NA	-0.023	0.379	-0.765	0.720	1
Odds Ratio	5.000	3.0	NA	0.978	NA	0.465	2.054	2
ses_group - 4:5	5.000	4.0	NA	-0.243	0.425	-1.075	0.589	1
Odds Ratio	5.000	4.0	NA	0.784	NA	0.341	1.802	2

0.1 Note on the importance of using `goodoutcome` rather than `outcome`

- Note that if you incorrectly used `outcome` in the `glm` fit, rather than `goodoutcome`, you get an error message, since the values aren't 0 or 1.
- But if you used `outcome` in the `lrm` fit, rather than `goodoutcome`, the machine assumes that what you want to look at is the factor version of `outcome`, and in this case, because `Good` comes alphabetically

after `Bad`, it fits the same model (`m1a`) as above.

1 Question 1 (1 point)

What do you conclude from the `m1a` summary about the odds ratio and confidence interval associated with the `treatment` variable? To answer this question, provide a complete description (in complete English sentences) of the odds ratio effect associated with `treatment` in the `summary(m1a)` output. This should require two or three sentences.

1.1 Answer for Question 1

The odds of a good outcome are estimated to be 0.557 times as large for a subject receiving the `treatment` than they are for a subject of the same `size` and the same `ses_group` who is not receiving the `treatment`. It wasn't necessary here to mention the 95% confidence interval for this odds ratio, which is (0.315, 0.986), which indicates that this effect is large enough to reach our usual standard to declare the effect of `treatment` to be statistically significant. Apparently, adjusting for `size` and `ses_group`, treatment is associated with a higher chance of a bad outcome in this model.

- English can be tricky. It is reasonable to write 0.557 times as high, but it isn't reasonable to write 0.557 times *higher*. That (higher) would only work if the value of the odds ratio was larger than 1. So I'd stick with "XXX times as large" or "XXX times as high".

2 Question 2 (1 point)

Why is the odds ratio shown in the `m1a` output for `size` different from that shown in the earlier presentation using `tidy` for the `m1` model? Keep your answer to two or three sentences.

2.1 Answer for Question 2

In `m1a`, the default choices for `summary` (in this `lrm` model) describe the impact of moving from a size at the 25th percentile of the data (100.525 mm) to a size at the 75th percentile of the data (135.5 mm). In `m1`, the default choice of `summary` (in this `glm` fit) describe the impact of moving 1 mm (for example, from 100.525 mm to 101.525 mm). Hence, the estimated effect of such a change on the odds of a good outcome appears much larger in the `m1a` output.

Note that while it is also true that the baseline category for `ses_group` changes from category 1 (in `m1`) to category 5 (in `m1a`) this has no impact on the odds ratio for `size` in the `summary(m1a)` output. Suppose, for instance that we reran `m1` but now forcing `ses_group = 5` to be the baseline category. The odds ratios estimated for `size` would not change, but everything else would match up perfectly with the `m1a` output. We would get:

```
honors_rev <- honors %>%
    mutate(ses_group = fct_relevel(ses_group, "5", "1", "2", "3", "4"))

m1_rev <- glm(goodoutcome ~ size + treatment + ses_group, data = honors_rev, family = binomial)

tidy(m1_rev, conf.int = TRUE, conf.level = 0.95,
     exponentiate = TRUE) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.220	0.752	-2.013	0.044	0.049	0.945
size	1.010	0.006	1.677	0.093	0.998	1.022
treatment	0.557	0.291	-2.010	0.044	0.314	0.985
ses_group1	0.706	0.511	-0.680	0.496	0.246	1.869

term	estimate	std.error	statistic	p.value	conf.low	conf.high
ses_group2	0.866	0.481	-0.300	0.764	0.327	2.184
ses_group3	0.978	0.379	-0.060	0.952	0.463	2.054
ses_group4	0.784	0.425	-0.573	0.567	0.335	1.786

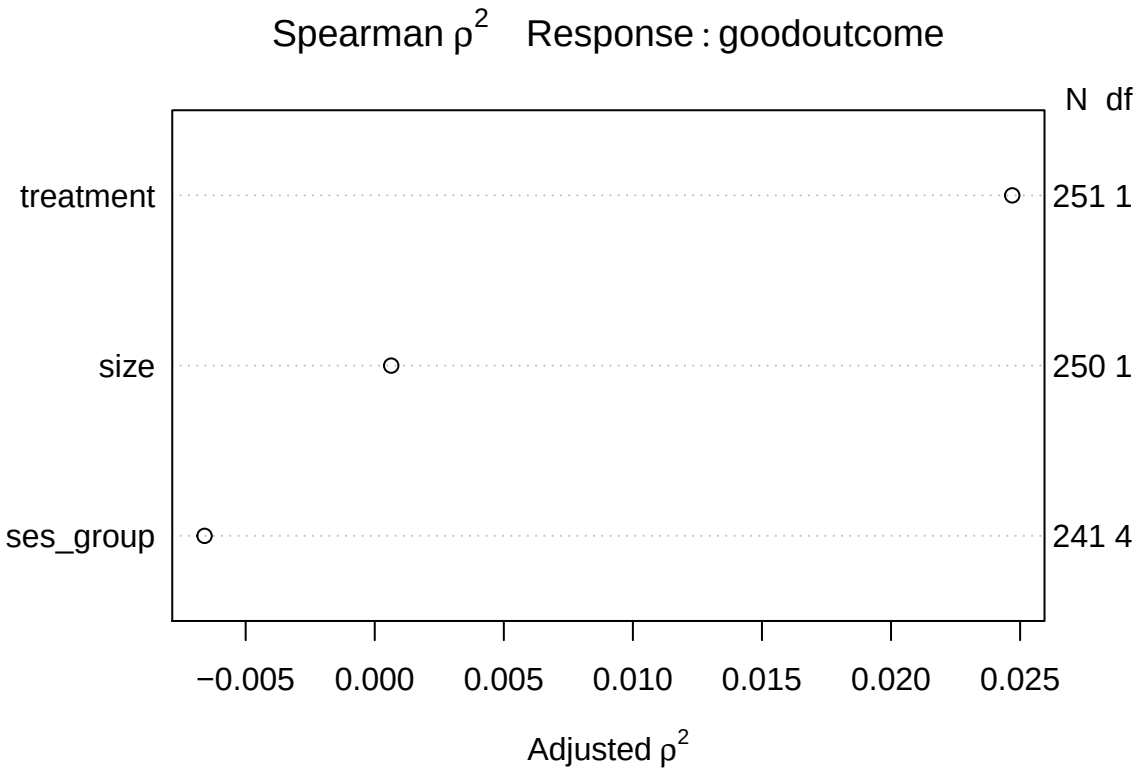
3 Question 3 (1 point)

Using the `honors` data (again without imputing any missing values), 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? This should be explained in one or more complete English sentences.

3.1 Answer for Question 3

The Spearman ρ^2 plot without doing any imputation is:

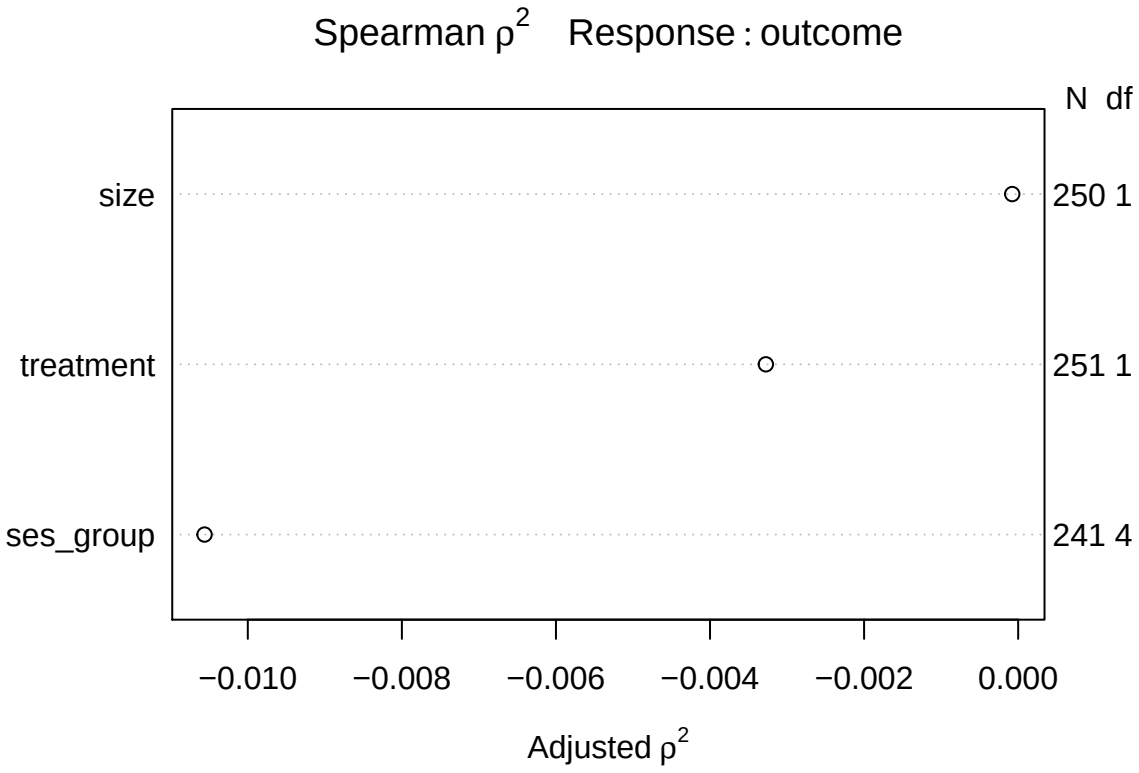
```
plot(spearman2(goodoutcome ~ size + treatment + ses_group, data = honors))
```



`treatment` is the most promising variable, and is binary. `size` is next, which is quantitative, so it looks like adding a `treatment-size` interaction is the most promising way to spend a single additional degree of freedom.

- Note the importance of using `goodoutcome` as the outcome of interest here. It would be equally reasonable to use `outcome = "good"`, but not to just use `outcome`, because `outcome` sets up different results for the squared Spearman correlation.
- The plot below demonstrates the **wrong** approach...

```
plot(spearman2(outcome ~ size + treatment + ses_group, data = honors))
```



Setup for Questions 4 and 5

Using the `honors` data, fit the model you specified in Question 3 (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, with 20 imputations. Be sure to show the code you used to fit your imputation model and your outcome model in your HTML file. Call the imputation model `model_imp` and the outcome model `m2`.

3.2 Models Fit by Dr. Love in Developing this Sketch

Here's a count by variable of the missingness in the data:

```
map_df(honors, function(x) sum(is.na(x)))
```

```
# A tibble: 1 x 6
  subject outcome  size treatment ses_group goodoutcome
    <int>   <int> <int>     <int>     <int>       <int>
1       0       0     5         4        14           0
```

So we're missing nothing in our `goodoutcome` variable, but are missing some `size`, `treatment` and `ses_group` information.

Here is the imputation model I fit:

```
set.seed(432432)
d <- datadist(honors)
```

```
options(datadist = "d")

model_imp <- aregImpute(~ goodoutcome + size +
                         treatment + ses_group,
                        nk = c(0, 3),
                        tlinear = TRUE, data = honors,
                        n.impute = 20, pr = FALSE)
```

- Note that I included `goodoutcome` in this model, and that I didn't include any interaction term.

Here is the outcome model, including the interaction term:

```
d <- datadist(honors)
options(datadist = "d")

m2 <- fit.mult.impute(goodoutcome ~
                        size + treatment + ses_group +
                        size*treatment,
                      fitter = lrm, xtrans = model_imp,
                      data = honors, x = TRUE, y = TRUE)
```

```
Variance Inflation Factors Due to Imputation:

       Intercept             size        treatment      ses_group=2
            1.03             1.02             1.03             1.06
     ses_group=3      ses_group=4      ses_group=5 size * treatment
            1.03             1.08             1.07             1.03

Rate of Missing Information:

       Intercept             size        treatment      ses_group=2
            0.03             0.02             0.03             0.06
     ses_group=3      ses_group=4      ses_group=5 size * treatment
            0.03             0.07             0.07             0.02

d.f. for t-distribution for Tests of Single Coefficients:

       Intercept             size        treatment      ses_group=2
        27411.89         65215.54         23496.55          5272.14
     ses_group=3      ses_group=4      ses_group=5 size * treatment
        23016.94          3627.65          4371.00         30883.55

The following fit components were averaged over the 20 model fits:

  stats linear.predictors
```

```
m2
```

```
Logistic Regression Model

 fit.mult.impute(formula = goodoutcome ~ size + treatment + ses_group +
     size * treatment, fitter = lrm, xtrans = model_imp, data = honors,
     x = TRUE, y = TRUE)

                      Model Likelihood     Discrimination    Rank Discrim.
```

```
                       Ratio Test           Indexes          Indexes
Obs           255    LR chi2     12.85    R2       0.069    C       0.633
 0            173    d.f.            7    g        0.579    Dxy     0.266
 1             82    Pr(> chi2) 0.0809    gr       1.786    gamma   0.267
max |deriv| 2e-08                        gp       0.118    tau-a   0.117
                                         Brier    0.208

                 Coef    S.E.   Wald Z Pr(>|Z|)
Intercept        -0.6330 1.0801 -0.59  0.5578
size             -0.0007 0.0083 -0.08  0.9356
treatment        -2.3813 1.3930 -1.71  0.0874
ses_group=2       0.2779 0.5988  0.46  0.6425
ses_group=3       0.5416 0.5043  1.07  0.2828
ses_group=4       0.2280 0.5557  0.41  0.6817
ses_group=5       0.4747 0.5033  0.94  0.3456
size * treatment  0.0140 0.0115  1.22  0.2213
```

4 Question 4 (1 point)

If Harry was size 100 mm and fell into group 4 in socio-economic status and Sally was size 120 mm and fell into group 3 in socio-economic status, and both Harry and Sally received the treatment, which of the two would have a larger probability of a Good outcome according to your model? How do you know? Your answer should be given in complete English sentences.

4.1 Answer for Question 4

The direct predictions in terms of probabilities can be made from the `lrm` model as follows:

```
newdat <- data.frame(name = c("Harry", "Sally"),
                     treatment = c(TRUE, TRUE),
                     ses_group = c(4, 3),
                     size = c(100, 120))

predict(m2, newdat, type = "fitted")
```

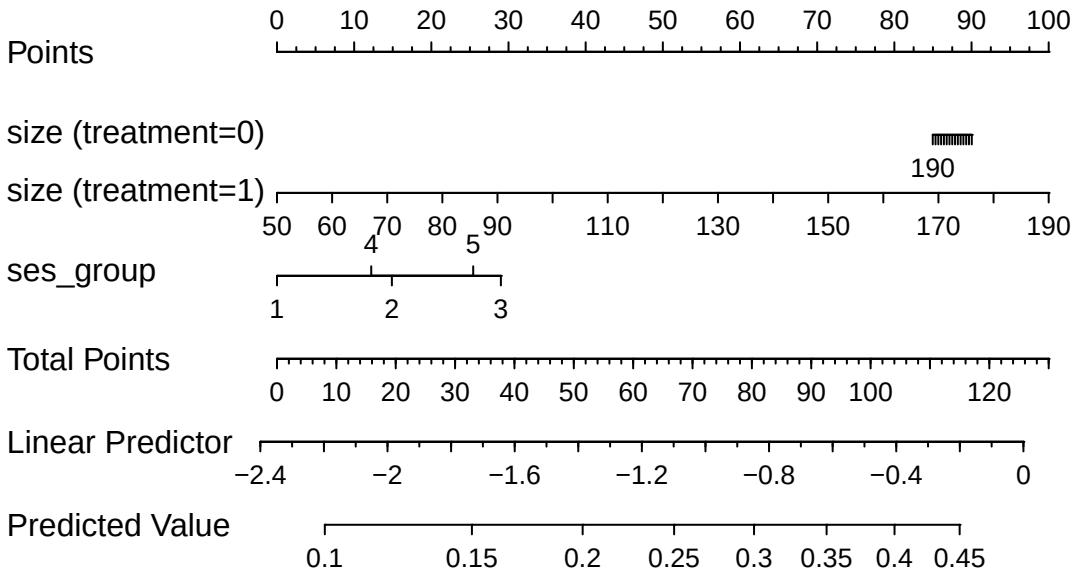
```
        1         2
0.1895625 0.2947395
```

So Sally's probability is definitely larger than Harry's.

- Note: In the Friday piece, I used a tibble here, rather than `data.frame` but the tibble forces you to say something about treating the `ses_group` as a factor.

Another way to assess this would be with a nomogram. Sally's larger size (since they each were treated) and group 3 status (as opposed to Harry's group 4) will yield a clearly larger probability of a good outcome for Sally.

```
plot(nomogram(m2, fun = plogis))
```



- Note the need for `plogis` in the nomogram call in order to make predictions in terms of probabilities.

4.1.1 What if we used a different imputation method?

Other reasonable imputation models would have been:

```
set.seed(432432)
d <- datadist(honors)
options(datadist = "d")

model_impZ <- aregImpute(~ goodoutcome + size +
                          treatment + ses_group,
                         data = honors,
                         n.impute = 20, pr = FALSE)

m2Z <- fit.mult.impute(goodoutcome ~
                         size + treatment + ses_group +
                         size*treatment,
                       fitter = lrm, xtrans = model_impZ,
                       data = honors, x = TRUE, y = TRUE)
```

```
Variance Inflation Factors Due to Imputation:

       Intercept             size        treatment      ses_group=2
```

```
           1.02            1.02            1.03            1.11
     ses_group=3     ses_group=4     ses_group=5 size * treatment
           1.09            1.08            1.06            1.03

Rate of Missing Information:

       Intercept            size       treatment     ses_group=2
            0.02            0.02            0.03            0.10
     ses_group=3     ses_group=4     ses_group=5 size * treatment
            0.08            0.08            0.06            0.02

d.f. for t-distribution for Tests of Single Coefficients:

       Intercept            size       treatment     ses_group=2
         42713.59        70102.29        28019.38         1832.94
     ses_group=3     ses_group=4     ses_group=5 size * treatment
          3094.34         3110.13         5342.33        30918.64

The following fit components were averaged over the 20 model fits:

  stats linear.predictors
```

```
predict(m2Z, newdat, type = "fitted")
```

```
        1         2
0.1904870 0.2959125
```

which yields predictions of essentially 0.19 for Harry and 0.30 for Sally,

and

```
set.seed(432432)
d <- datadist(honors)
options(datadist = "d")

model_impY <- aregImpute(~ goodoutcome + size +
                          treatment + ses_group,
                         nk = c(0, 3:5),
                         tlinear = FALSE, data = honors,
                         n.impute = 20, pr = FALSE)

m2Y <- fit.mult.impute(goodoutcome ~
                         size + treatment + ses_group +
                         size*treatment,
                       fitter = lrm, xtrans = model_impY,
                       data = honors, x = TRUE, y = TRUE)
```

```
Variance Inflation Factors Due to Imputation:

       Intercept            size       treatment     ses_group=2
            1.03            1.02            1.03            1.08
     ses_group=3     ses_group=4     ses_group=5 size * treatment
            1.04            1.09            1.07            1.02

Rate of Missing Information:
```

```
       Intercept             size        treatment      ses_group=2
            0.03             0.02             0.03             0.07
     ses_group=3      ses_group=4      ses_group=5 size * treatment
            0.04             0.08             0.07             0.02

d.f. for t-distribution for Tests of Single Coefficients:

       Intercept             size        treatment      ses_group=2
         28265.31         72518.76         27581.92          3786.92
     ses_group=3      ses_group=4      ses_group=5 size * treatment
         13779.68          2649.40          4169.24         36559.09

The following fit components were averaged over the 20 model fits:

  stats linear.predictors
```

```
predict(m2Y, newdat, type = "fitted")
```

```
        1         2
0.1901077 0.2957259
```

which yields predictions of essentially 0.19 for Harry and 0.30 for Sally, as well.

5 Question 5 (1 point)

Write a few English sentences describing how the addition of imputation and a non-linear term changes (or doesn't change) the conclusions that you draw in `m2` from what you saw in the `m1` (or, equivalently, the `m1a`) model examined earlier.

5.1 Answer for Question 5

The main things I was looking for:

- The appropriate conclusion is that the addition of imputation and an interaction effect have had at most a modest impact on the conclusions of the model. It remains a weak model.
- In the model post-imputation, (`m2`) it looks like the impact of size for those without the treatment is much more modest than the impact of size when the treatment is received, according to the nomogram for model `m2` shown earlier.

There were lots of ways to get to those conclusions.

```
m1a
```

```
Frequencies of Missing Values Due to Each Variable
goodoutcome        size   treatment   ses_group
          0           5           4          14

Logistic Regression Model

 lrm(formula = goodoutcome ~ size + treatment + ses_group, data = honors)


                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes           Indexes
 Obs           233    LR chi2      8.34    R2        0.049    C       0.614
  0            160    d.f.            6    g         0.479    Dxy     0.228
```

```
 1             73    Pr(> chi2) 0.2143    gr       1.615    gamma   0.229
max |deriv| 3e-11                          gp       0.100    tau-a   0.099
                                           Brier    0.208

            Coef    S.E.   Wald Z Pr(>|Z|)
Intercept   -1.8628 0.8216 -2.27  0.0234
size         0.0099 0.0059  1.68  0.0935
treatment   -0.5852 0.2911 -2.01  0.0444
ses_group=2  0.2035 0.6007  0.34  0.7348
ses_group=3  0.3251 0.5226  0.62  0.5339
ses_group=4  0.1047 0.5576  0.19  0.8511
ses_group=5  0.3479 0.5115  0.68  0.4964
```

```
m2
```

```
Logistic Regression Model

 fit.mult.impute(formula = goodoutcome ~ size + treatment + ses_group +
     size * treatment, fitter = lrm, xtrans = model_imp, data = honors,
     x = TRUE, y = TRUE)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes           Indexes
 Obs           255    LR chi2     12.85    R2       0.069    C       0.633
  0            173    d.f.            7    g        0.579    Dxy     0.266
  1             82    Pr(> chi2) 0.0809    gr       1.786    gamma   0.267
 max |deriv| 2e-08                          gp       0.118    tau-a   0.117
                                            Brier    0.208

                  Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept        -0.6330 1.0801 -0.59  0.5578
 size             -0.0007 0.0083 -0.08  0.9356
 treatment        -2.3813 1.3930 -1.71  0.0874
 ses_group=2       0.2779 0.5988  0.46  0.6425
 ses_group=3       0.5416 0.5043  1.07  0.2828
 ses_group=4       0.2280 0.5557  0.41  0.6817
 ses_group=5       0.4747 0.5033  0.94  0.3456
 size * treatment  0.0140 0.0115  1.22  0.2213
```

In the `m1` models, the effect of `treatment` is statistically significant at the 5% level after accounting for `size` and `ses_group`, but in the `m2` model, after the inclusion of the interaction term, this no longer appears to be the case. The interaction term, like the rest of `m2`, carries no statistical significance by Wald tests in `m2`, whereas in `m1a`, the `treatment` effect appeared to be just under our usual standard for statistical significance. In ANOVA testing for `m2`, shown below, we can see the combined impact of `treatment` (main effect + interaction) does still exhibit a p value below 0.05.

```
anova(m2)
```

```
               Wald Statistics          Response: goodoutcome

 Factor                                          Chi-Square d.f. P
 size  (Factor+Higher Order Factors)              2.91      2    0.2339
  All Interactions                                1.50      1    0.2213
 treatment  (Factor+Higher Order Factors)         7.97      2    0.0186
```

```
 All Interactions                                1.50      1    0.2213
ses_group                                        1.55      4    0.8173
size * treatment  (Factor+Higher Order Factors)  1.50      1    0.2213
TOTAL                                           11.21      7    0.1296
```

- In the model post-imputation, (`m2`) it looks like the impact of size for those without the treatment is much more modest than the impact of size when the treatment is received, according to the nomogram we saw earlier.
- In neither model does the effect of `size` or `ses_group` appear to meet the standard for statistical significance.
- The model discrimination is a bit better in the model with imputation and the interaction term. Specifically, the C statistic for `m2`, with imputation and interaction, is 0.633, as compared to 0.614 for the original model, but both are still weak.
- The Nagelkerke R-squared for `m2` is 0.069, a bit larger than the 0.049 for the original model, though still weak.
- The appropriate conclusion, then, is that the addition of imputation and an interaction effect have had at most a modest impact on the conclusions of the model. It remains a weak model.