432 Quiz 1 for Spring 2023

Thomas E. Love, Ph.D.

2023-02-22

Table of contents

Links

We'll post all Quiz 1 links at https://thomaselove.github.io/432-2023/quiz1.html no later than 5 PM on Thursday 2023-02-23.

This will include links to:

- the Main Document (this PDF document) containing the instructions and questions
- the Google Form Answer Sheet, and
- the **five** data sets we're providing

Instructions

This PDF document is **26** pages long. There are **28** questions on this Quiz. It is to your advantage to answer all of the Questions. Your score is based on the number of correct responses, so there's no chance a blank response will be correct, and a guess might be, so you should definitely answer all of the questions.

0.1 The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, and we will provide a link to that sheet on 2023-02-23. All of your answers must be submitted through the Google Form by **9 PM** on Monday 2023-02-27, without exception. The form will close at 9:30 PM, and no extensions beyond that time will be made available, so please do not wait until Monday evening to submit. We will not accept any responses except through the Google Form.

The Google Form contains places to provide your responses to each question, and a final affirmation where you'll type in your name to tell us that you followed the rules for the Quiz. You must complete that affirmation and then submit your results. When you submit your results (in the same way you submit a Minute Paper) you will receive an email copy of your submission, with a link that will allow you to edit your results.The Answer Sheet works like a Minute Paper, in that you must be logged into Google via CWRU to access it.

If you wish to work on some of the quiz and then return later, you can do this by [1] completing the final question (the affirmation) which asks you to type in your full name, and then [2] submitting the quiz. You will then receive a link at your CWRU email which will allow you to return to the Quiz Answer Sheet as often as you like without losing your progress.

0.2 The Data Sets

I have provided **five** data sets (called `dat1.Rds`, `dat10.Rds`, `dat13.csv`, `dat16.csv` and `datC.csv`) that are mentioned in the Quiz. Each of these may (or may not) be helpful to you.

0.3 Getting Help

This is an open book, open notes quiz. You are welcome to consult the materials provided on the course website and that we've been reading in the class, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants. You will be required to complete a short affirmation that you have obeyed these rules as part of submitting the Quiz.

If you need clarification on a Quiz question, you have exactly two ways of getting help:

1. You can ask your question in a **private** post on Campuswire to the instructors and TAs using the "Quiz 1" label.
2. You can ask your question via email to **431-help at case dot edu**.

During the Quiz period (2023-02-23 through 2023-02-27) we will not answer questions about the Quiz except through the two approaches listed above. We promise to respond to all questions received before 5 PM on 2023-02-27 in a timely fashion.

A few cautions:

- Specific questions are more likely to get helpful answers.
- We will not review your code or your English for you.
- We will not tell you if your answer is correct, or if it is complete.
- We will post to Campuswire in the "Quiz 1" folder and to the Quiz 1 page if we find an error in the Quiz that needs fixing.

0.4 When Should I Ask for Help?

We recommend the following process.

- If you encounter a tough question, skip it, and build up your confidence by tackling other questions.
- When you return to the tough question, spend no more than 10-15 minutes on it. If you still don't have it, take a break (not just to do other questions) but an actual break.
- When you return to the question, it may be much clearer to you. If so, great. If not, spend 5-10 minutes on it, at most, and if you are still stuck, ask us for help.
- This is not to say that you cannot ask us sooner than this, but you should **never, ever** spend more than 20 minutes on any question without asking for help.

0.5 Scoring and Timing

All questions are worth 2, 3, 4, or 5 points, as indicated, adding to a total of 100 points. The questions are not in any particular order, and range in difficulty from "things Dr. Love expects everyone to get right" to "things that are deliberately tricky". Some questions will take more time than others to answer.

The Quiz is meant to take 4-5 hours to complete. I expect most students will take 3-6 hours, and some will take as little as 2 or as many as 8. Again, it is **not** a good idea to spend a long time on any one question.

Dr. Love will grade the Quiz, and results (including an answer sketch) will be available by class time on Thursday 2023-03-02.

0.6 What does the Quiz cover?

Quiz A includes material from the first 10 classes in 432, as well as:

- Chapters 5, 8-17 and 19-22 of the 432 course notes, and
- all of Jeff Leek's *How to be a Modern Scientist*.

0.7 Writing Code into the Answer Sheet

Occasionally, we ask you to provide R code in your response. You need not include the `library` command at any time for any of your code. Assume in all questions that all relevant packages have been loaded in R.

0.8 R Packages and Settings

This is the set of packages I drew from in creating this test and the sketch.

1. This doesn't mean you need to use all of these packages.
2. This doesn't mean that I used all of these packages in building my sketch and the test itself.
3. This also doesn't mean that you are prevented from using other packages we've discussed in class to complete the Quiz, but it is definitely possible to complete all tasks using only these packages.

```
knitr::opts_chunk$set(comment = NA)

library(broom)
library(car)
library(caret)
library(equatiomatic)
library(GGally)
library(glue)
library(gt)
library(Hmisc)
library(janitor)
library(knitr)
library(MASS)
library(mosaic)
library(naniar)
library(patchwork)
library(pROC)
library(rms)
library(ROCR)
library(rsample)
library(simputation)
library(tidyverse)

# Note that all data files were downloaded onto
# my machine into a sub-folder called data below
# my main R Project directory for Quiz 1.

theme_set(theme_bw())
```

Setup for Questions 1-8

The `dat1.Rds` data file will be used for Questions 1-8. The `dat1` tibble contains three variables.

- `subject` is an identifying code, called the subject number,
- `calories` expended is a quantitative variable, and
- `satisfaction` is a quantitative rating scored on a 0-10 point scale, including one decimal place.

1 Question 1 (2 points)

Which subject in `dat1` expended the largest number of `calories` among the subjects in `dat1`? Your response should be the subject's identifying code (their `subject` value.)

2 Question 2 (2 points)

How many of the subjects in `dat1` have `satisfaction` values which are less than 5?

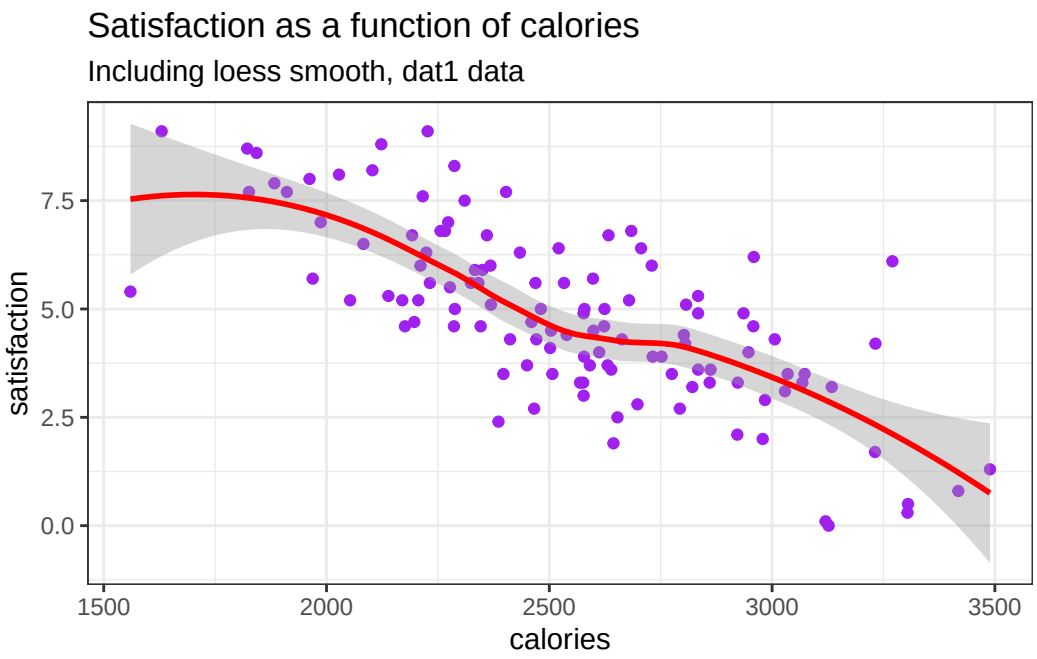
3 Question 3 (2 points)

How many of the subjects in `dat1` have both `calories` above 2400 **and** `satisfaction` below 5?

4 Question 4 (5 points)

Using the `dat1` tibble, a student attempted unsuccessfully to generate the Q4 Target Plot shown below, in R, developing the code shown in the Q4 Code Attempt also shown below. Explain, in a couple of sentences, how you would FIX the code in the Q4 Code Attempt to generate the Q4 Target Plot. Be specific about the changes you would make. Note the colors in the Q4 Target Plot are "purple" for the points and "red" for the smooth fit.

4.1 Q4 Target Plot



4.2 Q4 Code Attempt

```
ggplot(dat1, aes(x = calories, y = satisfaction)) +
    geom_point() +
    geom_smooth(formula = y ~ x, method = "lm") +
    theme_bw() +
    labs(title = "Satisfaction as a function of calories",
         subtitle = "Including loess smooth, dat1 data")
```

5 Question 5 (3 points)

Using the `dat1` tibble, specify the code required to fit (using `lm`) a model called `mod5` that predicts the `satisfaction` score across these subjects using a restricted cubic spline with 4 knots in the `calories` variable.

6 Question 6 (3 points)

Summarize the `mod5` model you built in Question 5. What is the observed (unadjusted) R^2 value for your model `mod5`, expressed as a proportion, and rounded to three decimal places?

7 Question 7 (3 points)

What does the `mod5` model predict will be the `satisfaction` value for a new subject who expends 2000 calories, rounding your answer to one decimal place? A point estimate is what we are looking for.

8 Question 8 (3 points)

Suppose we fit a new model (which I'll call `mod8`) to the `dat1` data, now using a restricted cubic spline with 3 knots in our `calories` variable. The `glance` function applied to `mod8` shows an AIC of 412.1 and a BIC of 423.2. Compare these results to those you obtain for `mod5`.

Which of the following conclusions is most appropriate based on these results?

a. The addition of the fourth knot in Model `mod5` is not helpful according to either AIC or BIC.
b. The addition of the fourth knot in Model `mod5` is helpful according to exactly one of AIC or BIC.
c. The addition of the fourth knot in Model `mod5` is helpful according to both AIC and BIC.
d. None of these conclusions are appropriate.

9 Question 9 (4 points)

In addition to the raw data, which of the following should be part of the "data package" that you share, according to Jeff Leek in *How to be a Modern Scientist*, when you are trying to maximize speed in the analysis of the data?

[**CHECK ALL RESPONSES THAT APPLY**]

a. A research question.
b. A tidy data set.
c. The results of an exploratory data analysis of the outcome of interest.
d. A code book describing each variable and its values.
e. An explicit recipe describing how you went from the raw data to the tidy data set and code book.
f. A substantial bribe.

10 Question 10 (4 points)

In Question 10, we look at a new model, built using the `dat10.Rds` data I have made available.

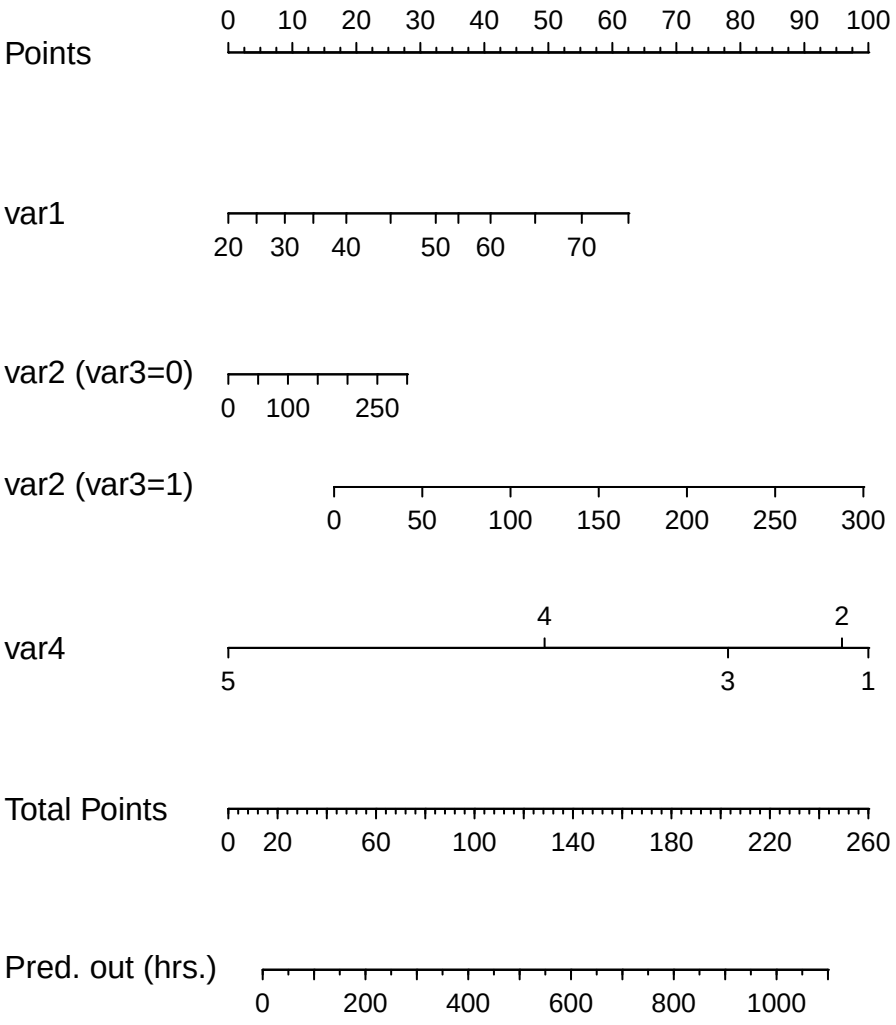
Use the nomogram shown on the next page as **Q10 Nomogram** to make a prediction about the outcome variable (called `out`, which is measured in hours) for two new subjects, Noah and Sophia.

- Noah and Sophia each have `var4` = 3, but …
  - Noah has `var1` = 45, `var2` = 150 and `var3` = 0.
  - Sophia, on the other hand, has `var1` = 30, `var2` = 200 and `var3` = 1.

Which of the following conclusions is most appropriate?

a. Noah and Sophia will have the same predicted outcome.
b. Noah's predicted outcome is longer than Sophia's, but by 100 hours or fewer.
c. Noah's predicted outcome is longer than Sophia's, and by more than 100 hours.
d. Noah's predicted outcome is shorter than Sophia's, but by 100 hours or fewer.
e. Noah's predicted outcome is shorter than Sophia's, and by more than 100 hours.
f. It is impossible to tell from the information provided.

10.1 Q10 Nomogram



11 Question 11 (4 points)

In attempting to measure complex relationships between four potential treatments and primary insurance on a health measure obtained after treatment among Northeast Ohio residents, two linear models were developed, called `model1` and `model2`. The 179 subjects each received one of four Treatments (Treatments W and X were more common than Y or Z), and the sample included similar numbers of Medicare, Medicaid and Commercially insured subjects.

Consider the Q11 Output below. What was included in `model2` but not included in `model1`?

11.1 Q11 Output

```
favstats(health ~ treatment + insurance, data = dat11) |> kable(dig = 1)
```

treatment.insurance	min	Q1	median	Q3	max	mean	sd	n	missing
W.Commercial	136.8	169.5	185.5	194.7	226.9	182.2	22.0	20	0
X.Commercial	169.8	184.9	200.8	218.1	251.1	202.3	21.7	20	0
Y.Commercial	163.8	194.2	197.0	217.4	233.8	203.4	21.3	10	0
Z.Commercial	148.8	166.9	174.6	177.9	185.4	169.8	12.7	9	0
W.Medicaid	163.1	173.0	185.9	195.0	214.2	185.1	14.0	20	0
X.Medicaid	124.4	152.7	159.6	175.4	181.9	161.0	16.7	20	0
Y.Medicaid	158.8	188.9	195.4	208.8	227.2	196.2	19.0	10	0
Z.Medicaid	147.6	153.4	166.3	176.7	183.4	165.3	13.3	10	0
W.Medicare	161.9	179.8	195.4	205.2	213.6	192.1	15.4	20	0
X.Medicare	132.6	162.7	174.9	186.7	208.4	174.1	18.5	20	0
Y.Medicare	184.4	207.6	223.4	225.5	264.0	219.3	21.9	10	0
Z.Medicare	146.2	161.5	175.3	182.1	197.0	172.1	16.6	10	0

```
anova(model1, model2)
```

```
Analysis of Variance Table

Model 1: health ~ treatment + insurance
Model 2: health ~ treatment * insurance
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    173 69757
2    167 55203  6     14554 7.338 5.622e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

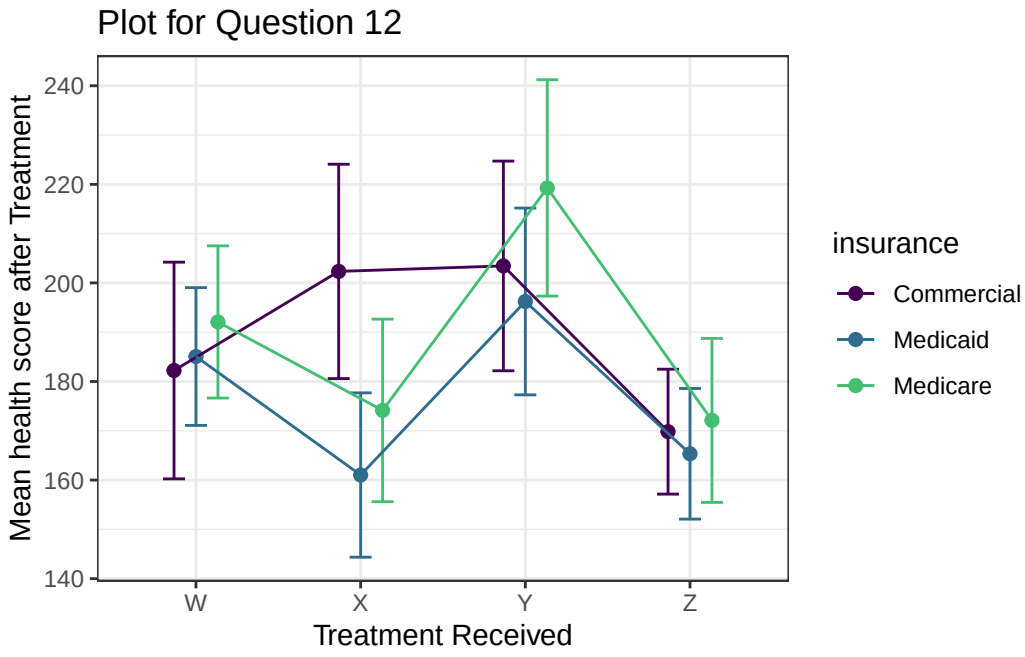
12 Question 12 (3 points)

Consider again the situation described in Question 11. In the Q12 Output below, we built an additional plot to help us study our data and choose between `model1` and `model2`. Specifically, the plot shows group means with intervals indicating one standard deviation in either direction.

What does the Q12 output suggest about the best choice of model, comparing `model1` to `model2`?

a. `model1` seems like the better choice.
b. `model2` seems like the better choice.
c. This plot does not help us make the decision.

12.1 Q12 Output



13 Question 13 (3 points)

In Question 13, we consider four potential models for an `outcome`, using various combinations of seven available predictors, which are labeled `a`, `b`, `c`, `d`, `e`, `f` and `g`. I have provided you with the data I used, in the `dat13.csv` file.

Consider the validation summaries provided for the four potential models shown in the Q13 Output below. Which of the models shown in the Q13 Output displays the strongest R^2 after bootstrap validation **and** the best mean squared error results after bootstrap validation?

a. The model that uses two of the seven predictors (`c` and `d`).
b. The model that uses three of the predictors (`c`, `d` and `g`).
c. The model that uses four of the predictors (`c`, `d`, `e` and `f`).
d. The model that uses all seven predictors (`a` through `g`).
e. None of these models.
f. It is impossible to tell from this output.

13.1 Q13 Output (start)

```
d <- datadist(dat13)
options(datadist = "d")

m13w <- ols(outcome ~ c + d,
            data = dat13, x = TRUE, y = TRUE)
m13x <- ols(outcome ~ c + d + e + f,
            data = dat13, x = TRUE, y = TRUE)
m13y <- ols(outcome ~ a + b + c + d + e + f + g,
            data = dat13, x = TRUE, y = TRUE)
m13z <- ols(outcome ~ c + d + g,
             data = dat13, x = TRUE, y = TRUE)
```

```
set.seed(43201); validate(m13w)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5549   0.5582  0.5525   0.0057          0.5492 40
MSE          27.7986  27.7392 27.9510  -0.2118         28.0104 40
g             6.6831   6.6972  6.6666   0.0306          6.6525 40
Intercept     0.0000   0.0000  0.3091  -0.3091          0.3091 40
Slope         1.0000   1.0000  0.9936   0.0064          0.9936 40
```

The **Q13 Output** continues on the next page.

13.2 Q13 Output (continued)

```
set.seed(43202); validate(m13x)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5590   0.5630  0.5548   0.0082          0.5508 40
MSE          27.5439  27.0075 27.8088  -0.8013         28.3452 40
g             6.7246   6.7127  6.7157  -0.0030          6.7276 40
Intercept     0.0000   0.0000  0.0556  -0.0556          0.0556 40
Slope         1.0000   1.0000  0.9995   0.0005          0.9995 40
```

```
set.seed(43203); validate(m13y)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5634   0.5759  0.5563   0.0196          0.5438 40
MSE          27.2691  26.9538 27.7136  -0.7598         28.0289 40
g             6.7604   6.8856  6.7261   0.1595          6.6009 40
Intercept     0.0000   0.0000  1.4345  -1.4345          1.4345 40
Slope         1.0000   1.0000  0.9768   0.0232          0.9768 40
```

```
set.seed(43204); validate(m13z)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5590   0.5588  0.5553   0.0035          0.5555 40
MSE          27.5441  27.6752 27.7737  -0.0985         27.6426 40
g             6.7239   6.7242  6.7035   0.0206          6.7033 40
Intercept     0.0000   0.0000  0.4654  -0.4654          0.4654 40
Slope         1.0000   1.0000  0.9926   0.0074          0.9926 40
```

14 Question 14 (3 points)

You are building a linear regression model for a quantitative outcome called `out` with only a limited number of observations, and need to include four predictors: `age` (in years), `prior` (1 = had prior surgery, 0 = no prior surgery), `severity` (three categories: High, Medium, Low) and `length` (in centimeters). Note that I have not provided you with the data set for this Question.

Suppose you are permitted to spend an additional four degrees of freedom beyond the five accounted for by the intercept term and the main effects of these four predictors. Based on the Q14 Spearman ρ^2 Plot shown on the next page, which of these models best does this additional spending?

Model	Specification
A	`out ~ age*severity + prior*length`
B	`out ~ rcs(age, 3) + rcs(length, 3) + prior + severity + severity %ia% prior`
C	`out ~ rcs(age, 4) + length + rcs(severity, 3) + prior`
D	`out ~ rcs(age, 4) + length + severity + prior + severity %ia% age`

Note that each specification listed above is just a part of the full specification. Each specification would be preceded by an appropriate `datadist` setup, and then the actual model fit would start with `ols(` and would end with `, data = dat14, x = TRUE, y = TRUE)`.

So the actual specification for Model A, for example, would be

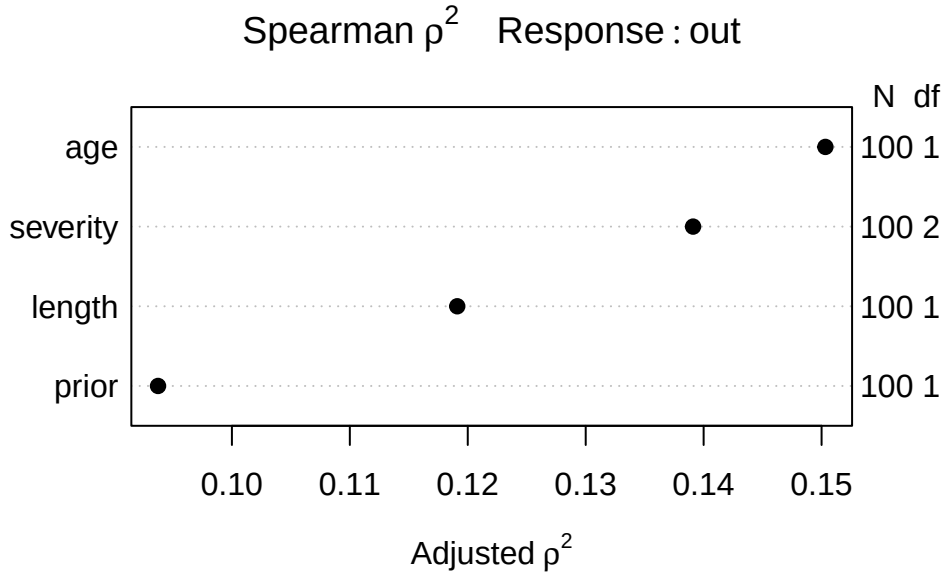
```
dd <- datadist(dat14); options(datadist = "dd")
modelA <- ols(out ~ age*prior + severity*length,
              data = dat14, x = TRUE, y = TRUE)
```

Now, which of the models does the best job of meeting our requirements?

a. Model A
b. Model B
c. Model C
d. Model D
e. None of these models are appropriate.

Note: The Q14 Spearman ρ^2 Plot is on the next page.

14.1 Q14 Spearman ρ^2 Plot



15 Question 15 (4 points)

Suppose you are reviewing an academic paper and you have the four options listed below. In "How to be a Modern Scientist", Jeff Leek suggests that there is a #1 way to be a jerk reviewer. Which of the following recommendation decisions could be made by someone who was actively TRYING TO BE a jerk reviewer?

[**CHECK ALL RESPONSES THAT APPLY**]

a. Reject
b. Major revisions
c. Minor revisions
d. Accept

16 Question 16 (5 points)

The `dat16.csv` file provided to you contains insurance data on thousands of subjects, each of whom is classified as falling into one of four different insurance categories, specifically Medicare, Commercial, Medicaid, and Uninsured. Some of the subjects (less than 5%) have missing data on this `insurance` variable.

Assume that all of the packages I listed at the top of this Quiz have been loaded in R, and that the data have been ingested into a tibble called `dat16`.

Suppose you now want to create a variable called `gov_ins` within the `dat16` tibble that (a) is a factor, and (b) which takes the value Yes if the subject's insurance is provided by the government (Medicare or Medicaid) but No otherwise, while (c) retaining NA for the missing values. Your first attempt is as shown below in the Q16 Code Attempt.

Fix the call to the `mutate` function in the Q16 Code Attempt so that your resulting code will actually do what is required. Your response should begin with `mutate(gov_ins =` on the answer sheet.

16.1 Q16 Code Attempt

```
dat16 <- dat16 |>
    type.convert(as.is = FALSE) |>
    mutate(gov_ins = factor(insurance,
                            Medicare or Medicaid = Yes,
                            Commercial or Uninsured = No))
```

Setup for Questions 17-27

The data in the `datC.csv` file contain information for 232 subjects on

- a binary `outcome` (Good or Bad),
- a `size` (quantitative, between 10 and 140, measured in centimeters),
- an indicator of whether a `treatment` was used (Yes = treatment was used or No = 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 code.

Import the data into a tibble called `datC` and use that tibble to respond to Questions 17-27.

17 Question 17 (4 points)

Using your `datC` tibble, 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 through the creation of a new variable called `ses_grpf`.

Ignore the missing values for now, so that you generate a complete-case analysis, so that some values are deleted due to missingness. We will deal with the missing values starting with Question 22.

The Q17 Output provided below will guide you as to what we're looking for.

You will have to create appropriate additional code in order to fit this `modC1` model (including the creation of the `ses_grpf` variable.) Note that you should then use the data and the output to verify that your code produces results that match those presented below.

Once you have accomplished that, we ask that you find the value of Akaike's Information Criterion (AIC) for your `modC1` model, by running `summary(modC1)`. Some of the output from my version of that summary appears below.

Your task on the answer sheet for Question 17 is to specify that AIC value (rounded to zero decimal places.)

17.1 Q17 Output

An appropriate analysis yields the following results. Note that you do not need to filter for complete cases in Questions 17-21.

```
datC <- read_csv("data/datC.csv", show_col_types = FALSE) |>
    mutate(goodout = ifelse(outcome == "Good", 1, 0),
           subject = as.character(subject))
```

Note that the fitting of the actual `modC1` and the creation of `ses_grpf` are not shown here.

```
exp(coef(modC1))
```

```
(Intercept)         size treatmentYes    ses_grpf2    ses_grpf3    ses_grpf4
   0.035236     1.016995     2.243937     2.525699     2.110962     2.903183
  ses_grpf5
   2.709872
```

```
exp(confint(modC1))
```

```
Waiting for profiling to be done...
```

```
                   2.5 %   97.5 %
(Intercept)  0.007857692 0.131734
size         1.004331658 1.030396
treatmentYes 1.222946023 4.166074
ses_grpf2    0.698235166 9.876384
ses_grpf3    0.711071534 7.191661
ses_grpf4    1.020048790 9.641941
ses_grpf5    0.854207790 9.698395
```

Here is a partial listing of the summary of the fitted `modC1` I created.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3563  -0.8844  -0.6293   1.2056   2.2171

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 265.16  on 214  degrees of freedom
Residual deviance: 244.64  on 208  degrees of freedom
  (17 observations deleted due to missingness)
```

18 Question 18 (5 points)

Interpret the `treatmentYes` value of 2.24 specified in the Q17 Output shown above, using a complete English sentence or two.

19 Question 19 (5 points)

What does the provided 95% confidence interval (1.22, 4.17) for `treatmentYes` in the Q17 Output tell you about the `treatment` variable? Provide your response in the form of 1-2 complete English sentences.

20 Question 20 (5 points)

Consider the Q20 Output provided below. Why is the odds ratio shown in the Q20 Output referring to `size` different from that shown in the earlier presentation (in the Q17 Output) of `exp(coef(modC1))` for the `size` variable in the same model? Again provide your response in the form of 1-2 complete English sentences.

20.1 Q20 Output

The output below comes from another approach to fitting the identical logistic regression model that we saw in Question 17, still using only the complete cases. I'll call this model `modC1L`, to emphasize that it contains the same outcome and predictors as were used in `modC1`.

```
summary(modC1L)
```

```
            Effects              Response : goodout

Factor             Low    High  Diff.  Effect    S.E.    Lower 0.95 Upper 0.95
size               60.575 96.65 36.075  0.607940 0.23481  0.14773   1.068100
 Odds Ratio        60.575 96.65 36.075  1.836600      NA  1.15920   2.910000
treatment - Yes:No  1.000  2.00     NA  0.808230 0.31180  0.19711   1.419400
 Odds Ratio         1.000  2.00     NA  2.243900      NA  1.21790   4.134400
ses_grpf - 1:4      4.000  1.00     NA -1.065800 0.56325 -2.16980   0.038141
 Odds Ratio         4.000  1.00     NA  0.344450      NA  0.11421   1.038900
ses_grpf - 2:4      4.000  2.00     NA -0.139290 0.52301 -1.16440   0.885780
 Odds Ratio         4.000  2.00     NA  0.869980      NA  0.31212   2.424900
ses_grpf - 3:4      4.000  3.00     NA -0.318660 0.40628 -1.11500   0.477640
 Odds Ratio         4.000  3.00     NA  0.727120      NA  0.32793   1.612300
ses_grpf - 5:4      4.000  5.00     NA -0.068906 0.44912 -0.94917   0.811360
 Odds Ratio         4.000  5.00     NA  0.933410      NA  0.38706   2.251000
```

21 Question 21 (4 points)

Again ignoring missingness in the `datC` tibble, obtain a Spearman ρ^2 plot and use it to identify a good way to add **ONE** non-linear term to this model (you may spend up to four additional degrees of freedom beyond the main effects model, but you can spend less if that makes sense.) Which of the following additions does the Spearman plot suggest?

a. A restricted cubic spline with 5 knots in size.
b. A restricted cubic spline with 5 knots in SES grouping.
c. A restricted cubic spline with 5 knots in treatment.
d. An interaction term between treatment and size.
e. An interaction term between treatment and SES grouping.
f. An interaction term between SES grouping and size.

22 Question 22 (3 points)

How many subjects in the `datC` tibble are missing data in at least one variable?

23 Question 23 (3 points)

How many missing observations are there on the outcome for your logistic regression models (i.e. the `goodout` variable) in the `datC` tibble?

Setting Up Questions 24-27

Note that in Questions 24-27, you will again be using the `datC` data, and you will fit a new model (which we'll call `modC2`) adding in the non-linear component that you specified in Question 21 to what was fit in `modC1` and `modC1L`, while also accounting for missing data using **multiple imputation**.

24 Question 24 (4 points)

The code listed below uses the `aregImpute()` function to fit a multiple imputation model, using `set.seed(2023)`.

```
set.seed(2023)
datC_imp <- aregImpute(~ goodout + treatment + ses_grpf + size,
                       nk = 0, data = datC, B = 10,
                       n.impute = 15, x = TRUE, pr = FALSE)
```

Run the code above, to complete the imputation process, and then consider the results by looking at what's stored in `datC_imp`.

Which of the variables has the largest observed R^2 value for predicting its non-missing values based on the last imputations completed by this approach?

a. the variable describing SES group
b. the treatment variable
c. the size variable
d. the goodout variable
e. It is impossible to tell.

25 Question 25 (5 points)

Fit the outcome model called `modC2` using `fit.mult.impute()`. Your `modC2` model should incorporate the multiple imputations from Question 24 that you stored in `datC_imp` and the outcome model you develop should include each of the original set of predictors of `goodout` augmented by the non-linear component you selected in Question 21. Your fit of model `modC2` should also save the important features of the design matrix to allow for subsequent assessment of calibration and discrimination.

Specify the code you used to fit model `modC2`. In the Answer Sheet, your code should begin with

```
modC2 <- fit.mult.impute(
```

26 Question 26 (3 points)

What is the in-sample estimated area under the ROC curve for your `modC2`, rounded to three decimal places?

27 Question 27 (3 points)

Validate your results in `modC2` using bootstrap validation with the default number of bootstrap replications. Set your seed as `2023` immediately before executing the relevant R function. What is the bootstrap-validated estimate of the area under the ROC curve for your `modC2`, rounded to three decimal places?

28 Question 28 (3 points)

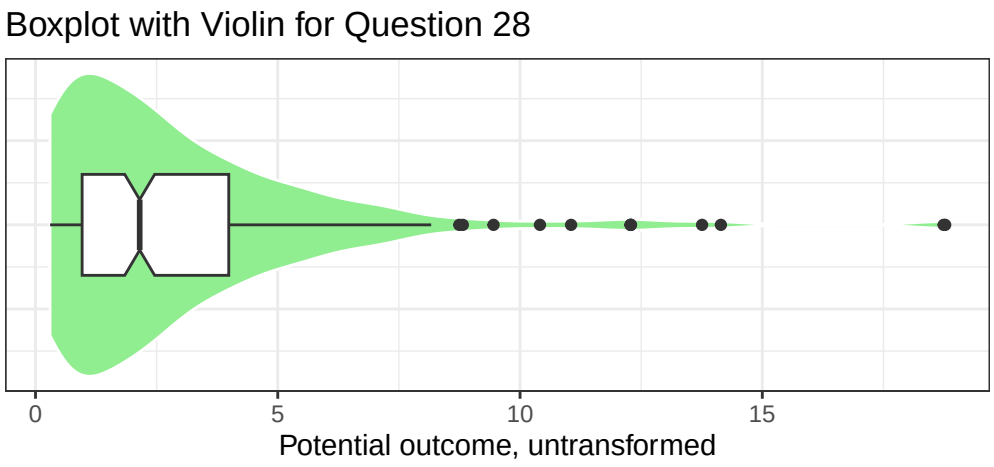
Consider the information provided below (in the Q28 Output) on the distribution of a potential outcome variable in a linear regression model to be built using the `dat28` tibble. Note that I have deliberately not provided you with these data.

Based on the three pieces of output provided, which of the following transformations of the `outcome` data would be most appropriate?

a. No transformation is needed. Fit the model to the raw outcome.
b. A logarithmic transformation is likely to be helpful.
c. Squaring the data would be helpful.
d. We should use a restricted cubic spline.
e. We should center the data.
f. It is impossible to tell from the information provided.

28.1 Q28 Output (start)

Here's a boxplot of the outcome data for Question 28.



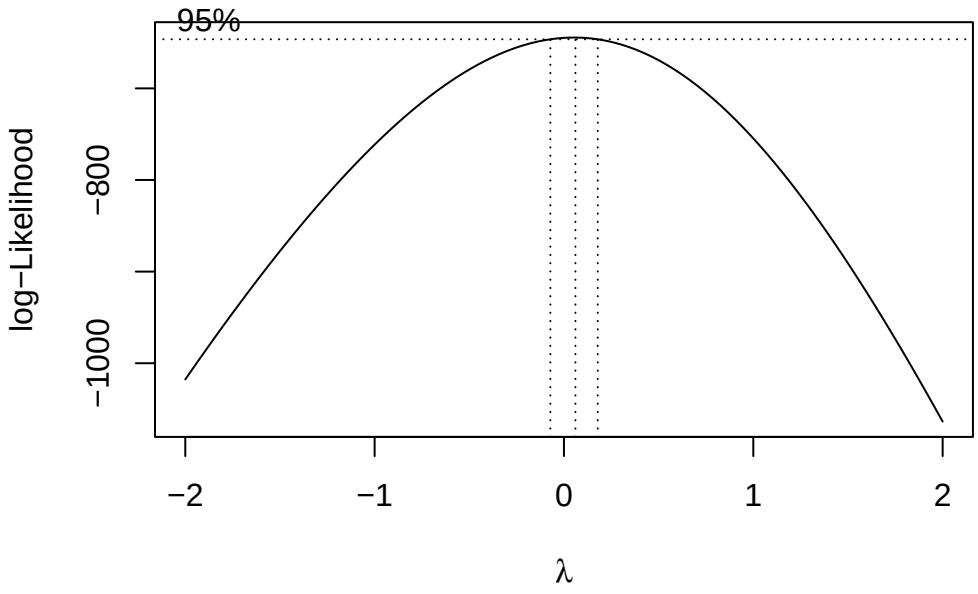
The Q28 Output continues on the next page.

28.2 Q28 Output (conclusion)

```
favstats(~ outcome, data = dat28) |> kable(digits = 2)
```

min	Q1	median	Q3	max	mean	sd	n	missing
0.3	0.96	2.15	3.99	18.77	2.99	2.95	240	0

Below, please find the Box-Cox plot fit to our kitchen sink model for the outcome using all available predictors.



This is the end of the Quiz.

Be sure to complete the Affirmation at the end of the Answer Sheet, and that you have submitted your Answer Sheet, and received your copy in your CWRU email by the deadline.