432 Quiz 1 for Spring 2024

Thomas E. Love, Ph.D.

2024-02-22

Links

All links relevant to this Quiz will be found starting at 5 PM on 2024-02-22 at https://github.com/THOMASELOVE/432-quizzes-2024/tree/main/quiz1.

This will include links to:

- the Main Document (this pdf) containing the instructions and questions
- the Google Form Answer Sheet, and
- the data sets we are providing

Deadline

The deadline to complete your work and submit the Google Form Answer Sheet is Tuesday 2024-02-27 at Noon. All of your answers must be submitted through the Google Form Answer Sheet found on the links page by the deadline, without exception. The form will close at that time, and no extensions will be made available, so please do not wait until the last moment to submit. We will not accept any responses except through the Google Form.

Instructions

This PDF document is **32** pages long. There are **26** questions on this Quiz, not counting the Bonus question on Campuswire. It is to your advantage to answer all of the Questions. A blank response cannot possibly score better than an incorrect one, a guess might be correct (or at least partially correct), so you should definitely answer all of the questions.

The Google Form Answer Sheet

The Google Form Answer Sheet 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 Quiz 1 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 1 Answer Sheet. You will then receive a link at your CWRU email which will allow you to return to the Quiz 1 Answer Sheet as often as you like without losing your progress.

The Data Sets

I have provided **five** data sets (called **dat01.csv**, **dat11.csv**, **dat18.csv**, **dat23.Rds** and **dat24.csv**) mentioned in the Quiz. They may be helpful to you.

What does the Quiz cover?

Quiz 1 includes material from the first 11 classes in 432, including all of Jeff Leek's *How to be a Modern Scientist*.

Bonus Question on Campuswire

Remember to complete the bonus question for Quiz 1 now available on Campuswire. Look for the Quiz 1 Bonus Question: "How to be a modern scientist" post (it's #40), and reply to it on Campuswire in response to that Question by the deadline for this Quiz to obtain credit. The bonus question will be worth either 3 or 4 points (I'll decide based on how the rest of the Quiz goes) for a complete and well-written response, with some partial credit also potentially available.

Scoring and Timing

All questions are worth between **3** and **5** points, adding to a total of **100** points, again not counting the Bonus question on Campuswire. 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 5-6 hours to complete. I expect most students will take 3-8 hours, and some will take as little as 2 or as many as 10. 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 2024-02-29.

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 one way of getting help:

- You can ask your question via email to **431-help at case dot edu**.

During the Quiz period (2024-02-22 through 2024-02-27) we will not answer questions about Quiz 1 except through the email listed above. We promise to respond to all questions received before 9 AM on 2024-02-27 in a timely fashion.

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.

A few cautions about asking us questions

- 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.

Writing Code into the Answer Sheet

Occasionally, we ask you to provide R code in your response. Do not include the `library` command at any time for any of your code. Instead, assume in all questions that all relevant packages have been loaded in R. A list of R packages that Dr. Love used in building the Quiz and its answer sketch is available in the next section.

Packages and Settings used by Dr. Love

This doesn't mean that I used all of these packages (I did not), or that you need to use all of these packages, nor does it mean that you are prevented from using other packages we've discussed in class to complete the Quiz, but all of the packages that I did use in writing the Quiz and its answer sketch are listed below.

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

library(bestglm)
library(broom)
library(caret)
library(Epi)
library(glue)
library(gt)
library(janitor)
library(MASS)
library(mosaic)
library(naniar)
library(patchwork)
library(pROC)
library(rms)
library(rsample)
library(simputation)
library(survey)
library(tidyverse)

theme_set(theme_bw())
options(dplyr.summarise.inform = FALSE)
```

The `dat01` data (Q01 - Q10)

The data in the `dat01.csv` file contain information for 230 subjects on a binary `outcome` (Positive or Negative), a `size` (quantitative, between 20 and 130, measured in centimeters), an indicator of `status` (either Treated or Untrtd[1]), 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 `dat01` and use that tibble to develop your responses to questions Q01 through Q10.

1 Q01 (3 points)

Using your `dat01` tibble, fit a logistic regression model to predict the log odds of a Positive `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 in Q06. The Output for Q01 below will guide you as to what we're looking for.

You will have to create appropriate additional code in order to fit this `mod1` 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 `mod1` model, using the `glance()` tool from the `broom` package.

Your task on the answer sheet for Q01 is to specify that AIC value (rounded to zero decimal places.) The output below and on the next page should be of some help to you in ensuring you've fit the model correctly.

Output for Q01

An appropriate analyses starts with the following ingestion and cleaning of the data. Note that we did not filter for complete cases in Questions Q01-Q05.

```
dat01 <- read_csv("data/dat01.csv", show_col_types = FALSE) |>
  clean_names() |>
  mutate(across(where(is_character), as_factor)) |>
  mutate(out_positive = ifelse(outcome == "Positive", 1, 0),
           subject = as.character(subject))
```

[1] Note that I abbreviated "Untreated" as "Untrtd."

Note that the fitting of the actual `mod1` and the creation of `ses_grpf` are not shown here.

Here is a **partial** listing of the summary of the fitted `mod1` I created, so you can check to see that you've done what you needed to do.

```
Call:
glm(formula = out_positive ~ size + status + ses_grpf, family = binomial,
    data = dat01)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.288164   0.705362  -4.662 3.14e-06 ***
size           0.017488   0.006622   2.641  0.00827 **
statusTreated  0.687568   0.311386   2.208  0.02724 *
...

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 266.02  on 213  degrees of freedom
Residual deviance: 246.16  on 207  degrees of freedom
  (16 observations deleted due to missingness)
```

2 Q02 (5 points)

Tidy the coefficients from your `mod1` and then interpret the relative odds associated with the `statusTreated` coefficient, after exponentiation. Be sure to specify the point estimate and a 95% confidence interval for this coefficient, all to two decimal places, *and* then interpret their meaning carefully, using two or more complete English sentences. Do not use the term "statistically significant" or any alternative phrasing of that concept, like "statistically detectable" in your response to this question.

3 Q03 (5 points)

Consider the Output for Q03, provided below. Why is the odds ratio shown in the Output for Q03 referring to `size` different from that shown in your tidied coefficients (that you developed in response to Question Q02) for the `size` variable in the same model? Again, provide your response in the form of 1-2 complete English sentences.

Output for Q03

The output below comes from another approach to fitting the same logistic regression model that we saw in Q01, still using only the complete cases. I'll call this model `mod1L`, to emphasize that it contains the same outcome and predictors as were used in `mod1`.

```
summary(mod1L)
```

```
           Effects              Response : out_positive

Factor                  Low High Diff. Effect   S.E.    Lower 0.95 Upper 0.95
size                    60  95   35     0.61220 0.23176  0.157950  1.06640
 Odds Ratio             60  95   35     1.84450      NA  1.171100  2.90500
status - Treated:Untrtd  1   2   NA     0.68769 0.31139  0.077376  1.29800
 Odds Ratio              1   2   NA     1.98910      NA  1.080400  3.66200
ses_grpf - 1:4           4   1   NA    -0.94722 0.56693 -2.058400  0.16395
 Odds Ratio              4   1   NA     0.38782      NA  0.127660  1.17820
ses_grpf - 2:4           4   2   NA    -0.18638 0.54310 -1.250800  0.87808
 Odds Ratio              4   2   NA     0.82996      NA  0.286260  2.40630
ses_grpf - 3:4           4   3   NA    -0.17970 0.40594 -0.975320  0.61593
 Odds Ratio              4   3   NA     0.83552      NA  0.377070  1.85140
ses_grpf - 5:4           4   5   NA     0.24541 0.43898 -0.614980  1.10580
 Odds Ratio              4   5   NA     1.27820      NA  0.540650  3.02170
```

4 Q04 (4 points)

Again working only with complete cases, which of the predictors in your model `mod01` would be included according to a best-subsets selection process using BIC as the information criterion? Use `method = "exhaustive", TopModels = 3, nvmax = "default"` as part of your function to obtain the result. CHECK ALL THAT APPLY.

a. `size`
b. `status`
c. `ses_groupf`
d. None of these variables.

5 Q05 (4 points)

Again ignoring missingness in the `dat01` 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). 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 status.
d. An interaction term between status and size.
e. An interaction term between status and SES grouping.
f. An interaction term between SES grouping and size.

6 Q06 (3 points)

a. How many subjects in the `dat01` tibble are missing data in at least one variable?

b. How many missing observations are there on the outcome for your logistic regression models in the `dat01` tibble?

Setting Up Q07 - Q10

Note that in Questions Q07 - Q10, you will again be using the `dat01` data, and you will fit a new model (which we'll call `mod2`) adding in the non-linear component that you specified in Q05 to what was fit in `mod1` and `mod1L`, while also accounting for missing data using **multiple imputation**.

7 Q07 (4 points)

The code listed below uses the `aregImpute()` function to fit a multiple imputation model, using `set.seed(4322024)`.

```
set.seed(4322024)
dat01_imp <- aregImpute(~ out_positive + status + ses_grpf + size,
                        nk = 0, data = dat01, B = 10,
                        n.impute = 15, x = TRUE, pr = FALSE)
```

Run the code above, to complete the imputation process, and then consider the results.

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 status variable
c. the size variable
d. the out_positive variable
e. It is impossible to tell.

8 Q08 (5 points)

Fit the outcome model called `mod2` using `fit.mult.impute()`. Your `mod2` model should incorporate the multiple imputations from Q07 that you stored in `dat01_imp` and the outcome model you develop should include each of the original set of predictors of `out_positive` augmented by the non-linear component you selected in Q05. Your fit of model `mod2` 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 `mod2`. In the Answer Sheet, your code should begin with

```
mod2 <- fit.mult.impute(
```

9 Q09 (4 points)

a. (2 points) What is the in-sample estimated area under the ROC curve for the `mod2` you fit in Q08, rounded to three decimal places?

b. (2 points) Use bootstrap validation on the summaries for your `mod2` from Q08, with B = 40 replications and set your seed to be `20240225`. What is the optimism-corrected validated estimate of the area under the ROC curve for your model `mod2`, rounded to three decimal places?

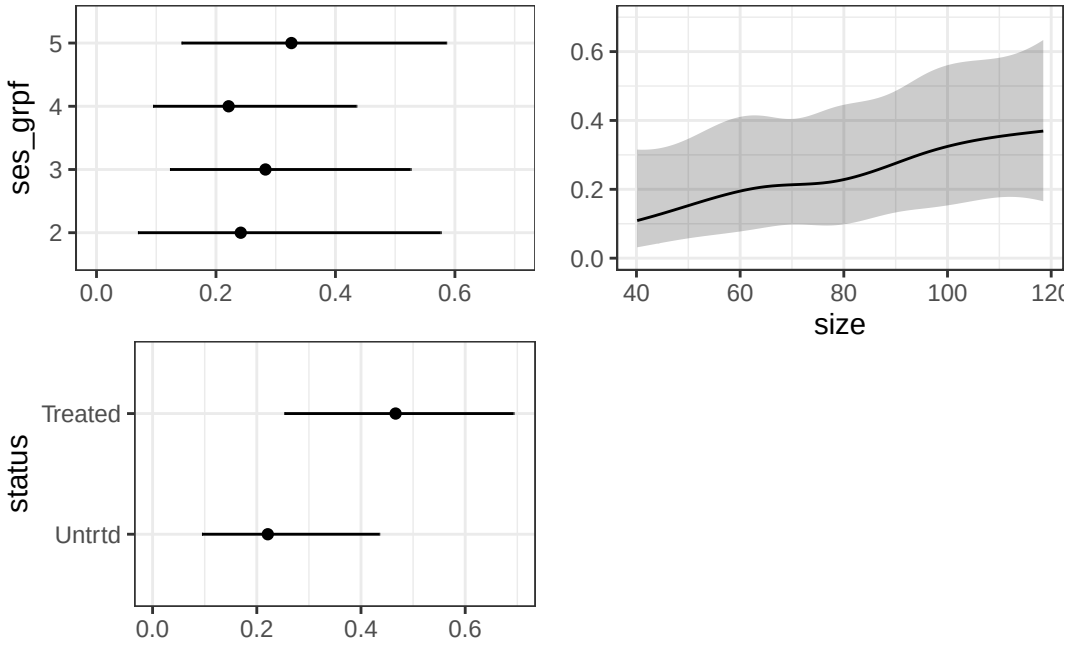
10 Q10 (4 points)

Consider the four sets of plots for Q10 printed on the next four pages, developed using `ggplot(Predict(modelname, fun = plogis))` for plot sets A and B, and using `plot(summary(modelname))` for plot sets C and D. Which two of these four sets of plots come from your `mod2` model?

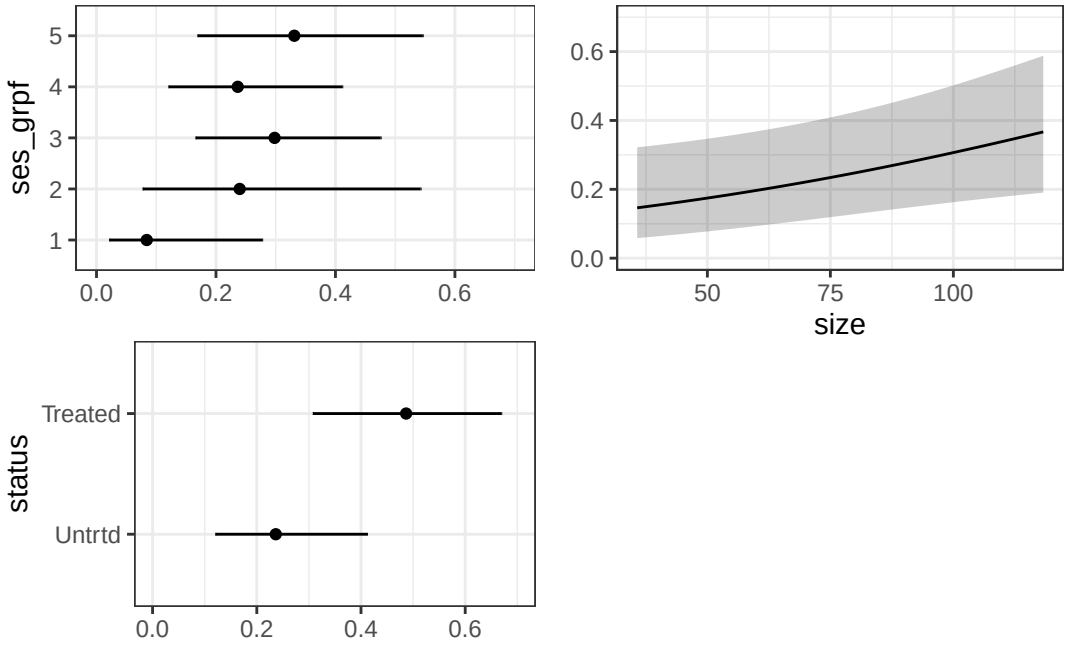
a. Plot Sets A and C
b. Plot Sets A and D
c. Plot Sets B and C
d. Plot Sets B and D

Just to confirm, exactly two of these plots do come from `mod2` and the other two do not.

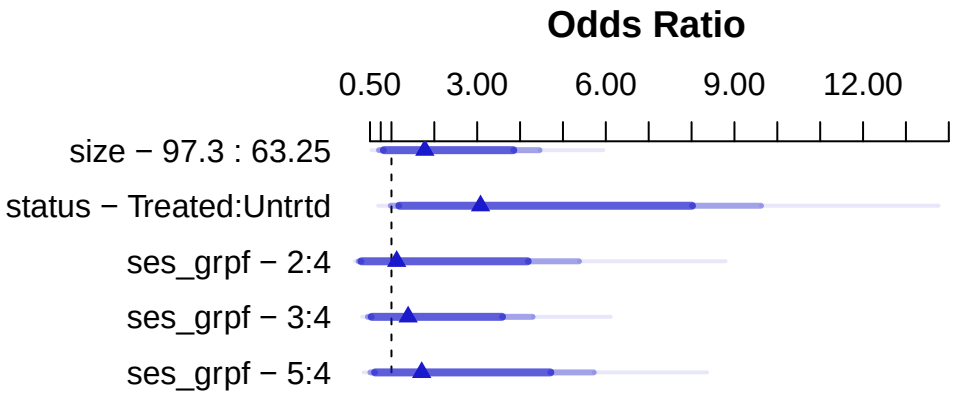
Plot Set A for Q10



Plot Set B for Q10

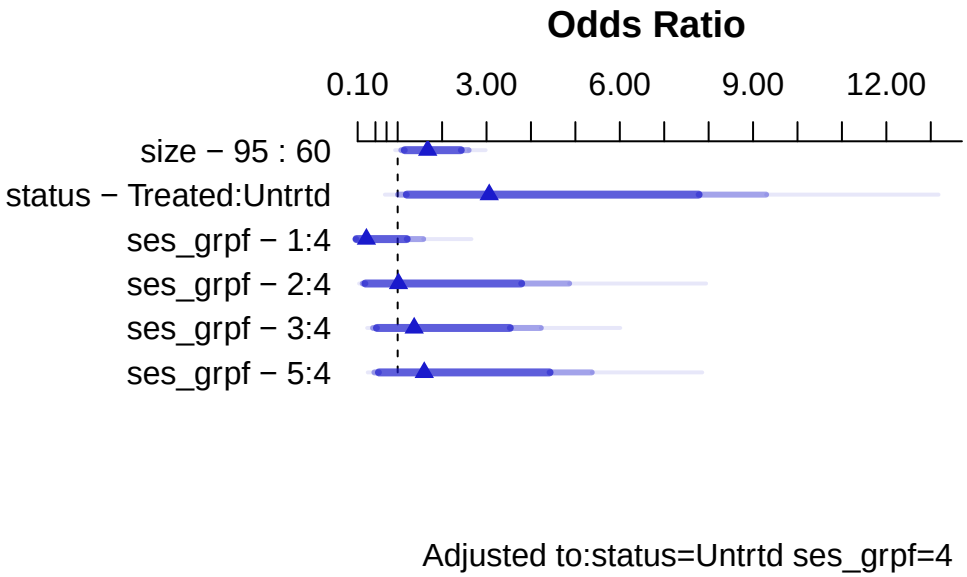


Plot Set C for Q10



Adjusted to:status=Untrtd ses_grpf=4

Plot Set D for Q10



This is the end of the output for Q10.

The `dat11` data (Q11-Q13)

The `dat11.csv` data file provided to you will be used for Questions 11-13. Ingest the data into R as a tibble called `dat11`, containing three variables.

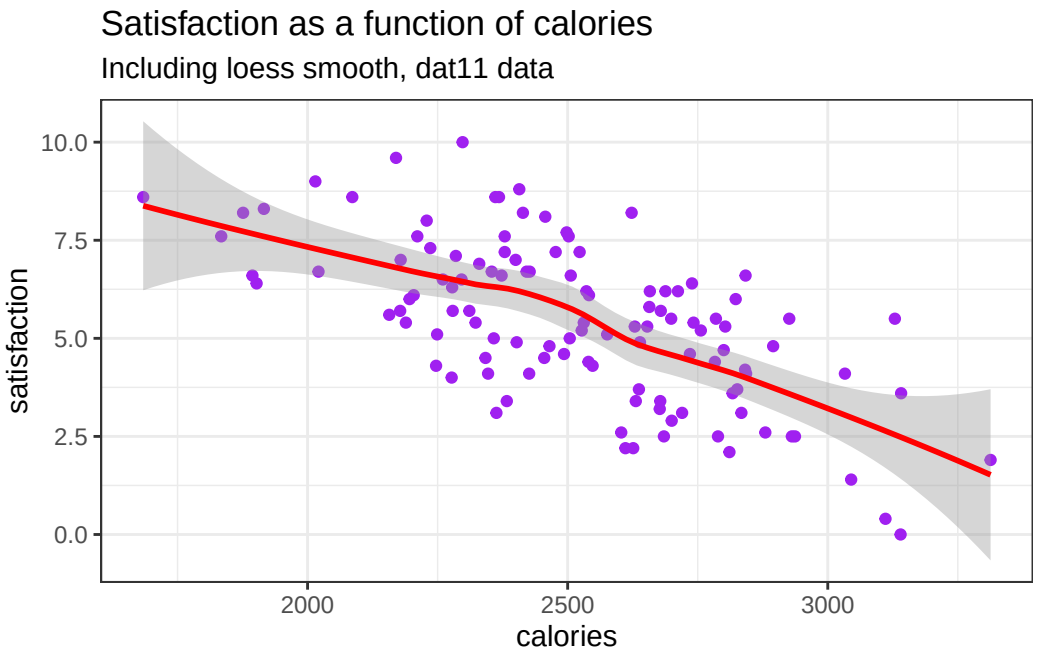
- `subject` is an identifying code
- `calories` is quantitative
- `satisfaction` is quantitative, as well.

11 Q11 (5 points)

Using the `dat11.csv` data set, a student attempted unsuccessfully to generate the Q11 Target Plot shown below, in R, developing the code shown in the Q11 Code Attempt shown at the top of the next page.

Explain, in a couple of sentences, how you would FIX the code in the Q11 Code Attempt to generate the Q11 Target Plot. Be specific about the changes you would make. Note the colors in the Target Plot are "purple" for the points and "red" for the smooth fit.

Q11 Target Plot



Q11 Code Attempt

```
ggplot(dat11, aes(x = calories, y = satisfaction)) +
    geom_point() +
    geom_smooth(formula = y ~ x, method = "lm") +
    labs(title = "Satisfaction as a function of calories",
         subtitle = "Including loess smooth, dat11 data")
```

12 Q12 (4 points)

Using the `dat11` tibble, specify the code required to fit (using `lm`) a model called `mod12` that predicts the `satisfaction` score across these subjects using an orthogonal polynomial of degree 3 in the `calories` variable. Then summarize the `mod12` model you built. What is the observed R^2 value for your model `mod12`, expressed as a proportion, and rounded to three decimal places?

13 Q13 (3 points)

A new model in R (which I'll call `mod13`) was fit to the `dat11` data, now using an orthogonal polynomial of degree 2. The `glance` function applied to `mod13` shows an AIC of 430.2 and a BIC of 441.2. Compare these results to those you obtain for the `mod12` you fit in Q12. Which of the following conclusions is most appropriate based on these results?

a. The cubic term in Model `mod12` is not helpful according to either AIC or BIC.
b. The cubic term in Model `mod12` is helpful according to exactly one of AIC or BIC.
c. The cubic term in Model `mod12` is helpful according to both AIC and BIC.
d. None of these conclusions are appropriate.

14 Q14 (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 THAT APPLY]

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

15 Q15 (4 points)

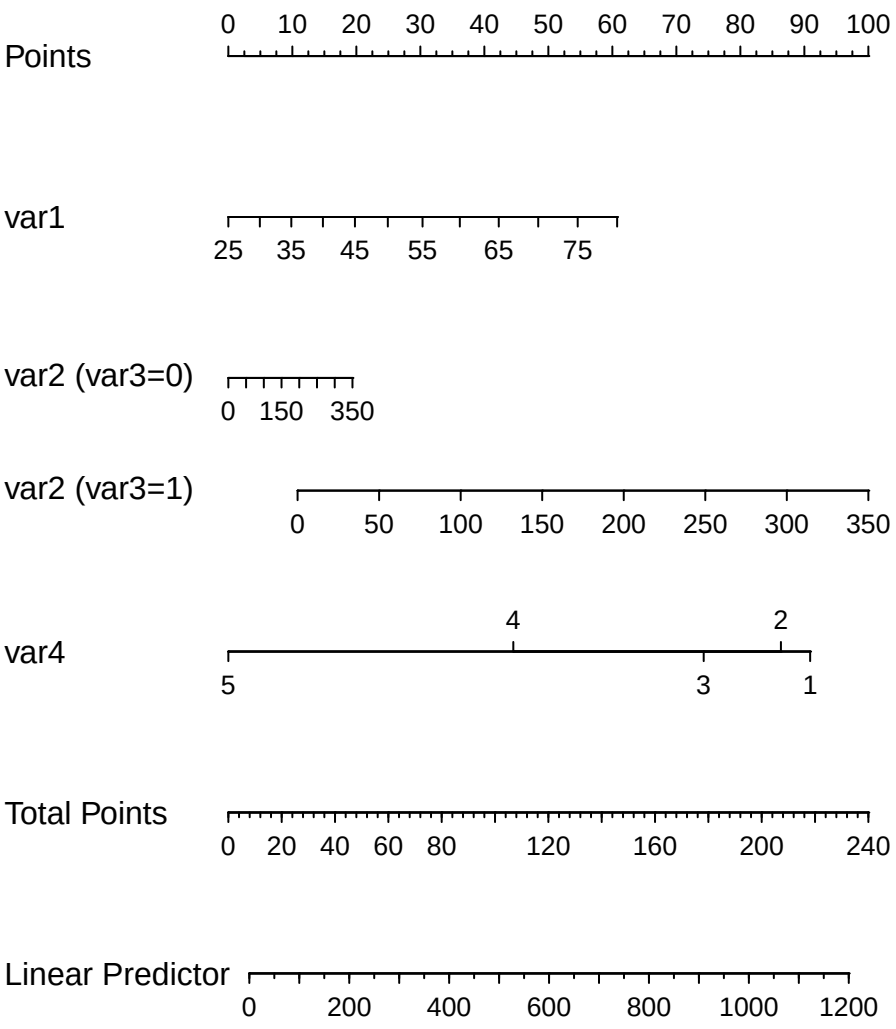
In Q15, we look at a new data set, not provided to you. Use the nomogram shown in the Output for Q15 on the next page to make a prediction about the outcome variable (measured in hours) for each of two subjects (named Noah and Sophia), based on the model described by that nomogram.

While each subject has `var4` = 5, Noah has `var1` = 45, `var2` = 150 and `var3` = 0, and Sophia has `var1` = 30, `var2` = 200 and `var3` = 1.

Which of the following descriptions 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.

Output for Q15



This is the end of the output for Q15.

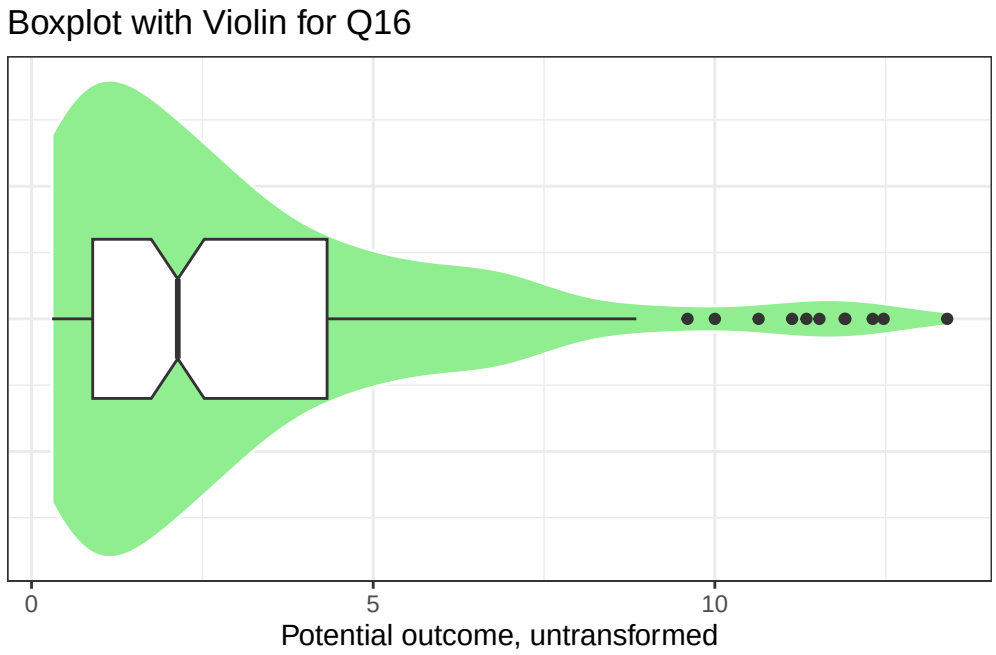
16 Q16 (3 points)

Consider the information provided below (in the Output for Q16) on the distribution of a potential outcome variable in a linear regression model to be built using the `dat16` tibble, which contains data on 195 subjects. Note that I have deliberately not provided you with these data.

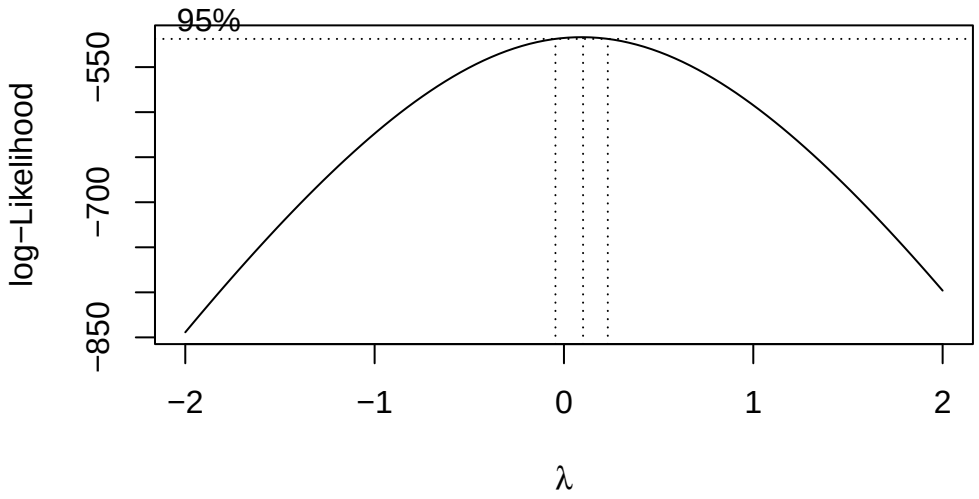
Based on the three pieces of output provided, which of the following transformations of the `outcome` data would be most appropriate?

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.

Output 1 of 3 for Q16



Output 2 of 3 for Q16: Box-Cox plot



Output 3 of 3 for Q16: Hmisc::describe()

```
select(dat16, outcome)

 1  Variables      195  Observations
--------------------------------------------------------------------------------
outcome
       n  missing distinct     Info     Mean      Gmd      .05      .10
     195        0      163        1    3.093    3.012    0.320    0.404
     .25      .50      .75      .90      .95
   0.895    2.140    4.325    7.014    9.720

lowest : 0.3   0.31  0.32  0.34  0.35 , highest: 11.9  11.91 12.31 12.47 13.4
--------------------------------------------------------------------------------
```

This is the end of the output for Q16.

17 Q17 (3 points)

The table below describes the result of using 10-fold cross-validation to compare seven candidate linear regression models (labeled modelA, modelB, modelC, modelD, modelE, modelF, and modelG) for a data set predicting a quantitative outcome. The table below summarizes cross-validation R-square (labeled `Rsquared`), the root mean squared prediction error (labeled `RMSE`), and the mean absolute prediction error (labeled `MAE`).

model	Rsquared	RMSE	MAE
modelA	0.5904	5.5745	4.4624
modelB	0.5959	5.5466	4.4418
modelC	0.6006	5.5297	4.4369
modelD	0.5952	5.5327	4.4393
modelE	0.5495	5.8535	4.7791
modelF	0.5948	5.5243	4.4490
modelG	0.5449	5.8724	4.7504

According to the table provided above, which model shows the strongest results in terms of:

Rows:

a. cross-validated R-square
b. root mean squared prediction error
c. mean absolute prediction error

Columns:

1. modelA
2. modelB
3. modelC
4. modelD
5. modelE
6. modelF
7. modelG

18 Q18 (5 points)

The `dat18.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.

Ingest the data into a tibble called `dat18`.

Suppose you now want to create a variable called `gov_ins` within the `dat18` 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 Code Attempt for Q19. Fix the call to the `mutate` function in that code so that your resulting code will actually do what is required.

On the answer sheet, your response should begin with `mutate(gov_ins =`

Code Attempt for Q18

```
dat18 <- dat18 |>
  mutate(across(where(is_character), as_factor)) |>
  mutate(gov_ins = factor(insurance,
                          Medicare or Medicaid = Yes,
                          Commercial or Uninsured = No))
```

19 Q19 (3 points)

You are building a linear regression model for an 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 Spearman ρ^2 plot provided for Q19 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 = dat19, x = TRUE, y = TRUE)`.

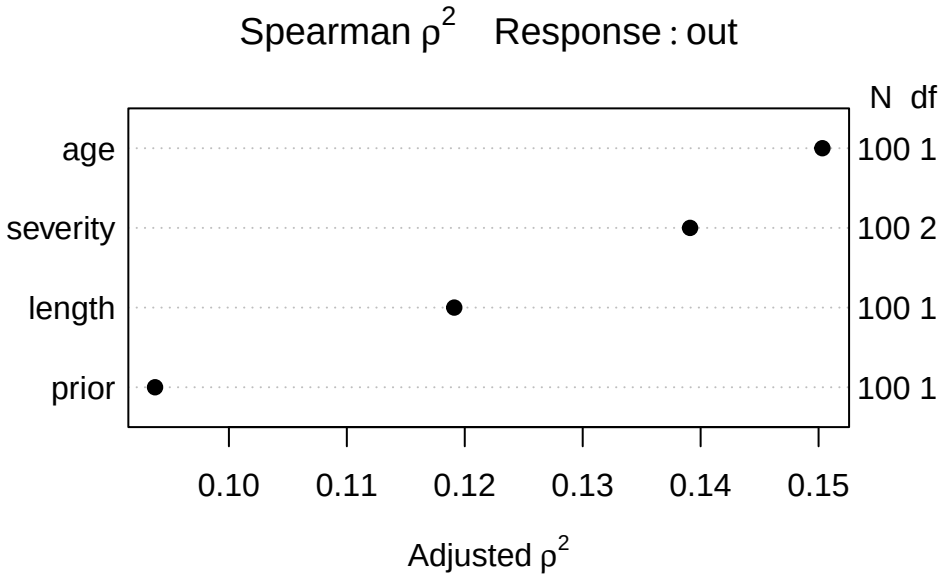
So the actual specification for Model A, for example, would be

```
dd <- datadist(dat19); options(datadist = "dd")
modelA <- ols(out ~ age*severity + prior*length,
              data = dat19, x = TRUE, y = TRUE)
```

Now, which of the models specified above does the best job of meeting the requirements for Q19?

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

Spearman Plot for Q19



This is the end of the output for Q19.

20 Q20 (3 points)

In Q20, 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`.

Consider the validation summaries provided for the four potential models shown in the Output for Q20. Which of the models shown in the Output for Q20 below displays the strongest R^2 and 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 the above.

Output for Q20

```
d <- datadist(dat20)
options(datadist = "d")

m20w <- ols(outcome ~ c + d,
            data = dat20, x = TRUE, y = TRUE)
```

```
m20x <- ols(outcome ~ c + d + e + f,
            data = dat20, x = TRUE, y = TRUE)
m20y <- ols(outcome ~ a + b + c + d + e + f + g,
            data = dat20, x = TRUE, y = TRUE)
m20z <- ols(outcome ~ c + d + g,
             data = dat20, x = TRUE, y = TRUE)
```

```
set.seed(4321); validate(m20w, B = 40)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5129   0.5085  0.5099  -0.0013          0.5142 40
MSE          25.3309  25.1917 25.4894  -0.2976         25.6285 40
g             5.8893   5.8022  5.8675  -0.0653          5.9546 40
Intercept     0.0000   0.0000 -0.4041   0.4041         -0.4041 40
Slope         1.0000   1.0000  1.0074  -0.0074          1.0074 40
```

```
set.seed(4322); validate(m20x, B = 40)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5204   0.5287  0.5158   0.0128          0.5076 40
MSE          24.9392  24.4181 25.1792  -0.7611         25.7003 40
g             5.9519   5.9711  5.9257   0.0453          5.9066 40
Intercept     0.0000   0.0000  0.2362  -0.2362          0.2362 40
Slope         1.0000   1.0000  0.9963   0.0037          0.9963 40
```

```
set.seed(4323); validate(m20y, B = 40)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5226   0.5328  0.5143   0.0185          0.5042 40
MSE          24.8249  24.5320 25.2600  -0.7280         25.5529 40
g             5.9657   6.0498  5.9304   0.1194          5.8463 40
Intercept     0.0000   0.0000  1.1936  -1.1936          1.1936 40
Slope         1.0000   1.0000  0.9798   0.0202          0.9798 40
```

```
set.seed(4324); validate(m20z, B = 40)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5161   0.5203  0.5120   0.0084          0.5078 40
MSE          25.1638  24.8948 25.3808  -0.4860         25.6498 40
g             5.9174   5.9223  5.8995   0.0228          5.8946 40
Intercept     0.0000   0.0000  0.2782  -0.2782          0.2782 40
Slope         1.0000   1.0000  0.9956   0.0044          0.9956 40
```

This is the end of the output for Q20.

21 Q21 (3 points)

In attempting to measure the complex relationships between four potential treatments and primary insurance on a summary measure of health obtained after treatment among 360 Northeast Ohio residents, two linear models were developed, called `model21A` and `model21B`. Each of the 360 subjects received exactly one of the four Treatments (although Treatments W and X were selected more often than Y or Z), and the sample was obtained to include equal numbers of Medicare, Medicaid and Commercially insured subjects.

Consider the Output for Q21 provided below. What was included in `model21B` but not included in `model21A`?

21.1 Output for Q21

```
anova(model21A)
```

```
Analysis of Variance Table

Response: health
           Df Sum Sq Mean Sq F value    Pr(>F)
treatment   3  34774 11591.2  9.1824 7.262e-06 ***
insurance   2  13768  6884.1  5.4534  0.004649 **
Residuals 354 446866  1262.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model21A, model21B)
```

```
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    354 446866
2    348 433191  6     13675 1.8309 0.09223 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the end of the output for Q21.

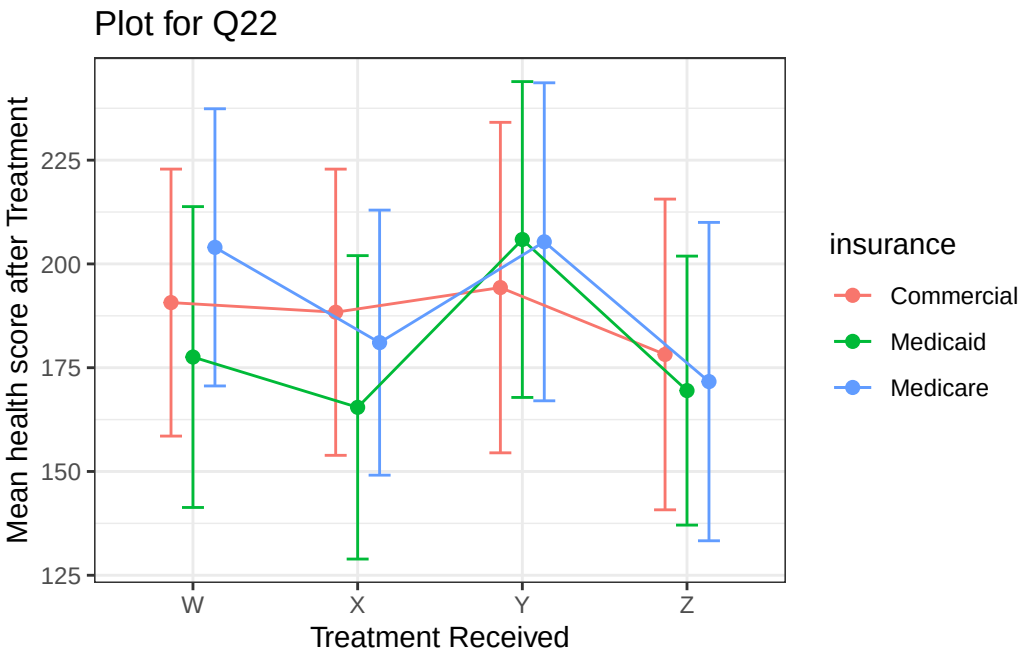
22 Q22 (3 points)

Consider again the situation described in Q21. In the Output for Q22 shown below, we built an additional plot to help us study those two models. Specifically, the plot shows group means with intervals indicating one standard deviation in either direction.

What does the Output for Q22 suggest about the best choice of model, comparing `model21A` to `model21B`?

a. `model21A` seems like the better choice.
b. `model21B` seems like the better choice.
c. This plot does not help us make the decision.

Output for Q22



This is the end of the output for Q22.

23 Q23 (4 points)

I have used the `dat23.Rds` file provided to you to predict `stent` using `ves1proc` and `abcix` in two models, one called `mod23a` and one called `mod23b`. as shown below.

```
dat23 <- read_rds("data/dat23.Rds")

mod23a <- lm(stent ~ abcix + ves1proc, data = dat23)

tidy(mod23a, conf.int = TRUE, conf.level = 0.90) |>
    select(term, estimate, std.error, conf.low, conf.high, p.value)
```

```
# A tibble: 3 x 6
  term        estimate std.error conf.low conf.high  p.value
  <chr>          <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
1 (Intercept)   0.630     0.0394   0.565    0.695   3.30e-51
2 abcix         0.132     0.0336   0.0764   0.187   9.32e- 5
3 ves1proc     -0.0394    0.0233  -0.0778  -0.00107 9.09e- 2
```

```
mod23b <- glm(stent ~ abcix + ves1proc, family = binomial(),
              data = dat23)

tidy(mod23b, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.90) |>
    select(term, estimate, std.error, conf.low, conf.high, p.value)
```

```
# A tibble: 3 x 6
  term        estimate std.error conf.low conf.high  p.value
  <chr>          <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
1 (Intercept)    1.74      0.175    1.30      2.32  0.00164
2 abcix          1.79      0.150    1.40      2.29  0.000102
3 ves1proc       0.836     0.105    0.704     0.995 0.0881
```

Question 23 continues on the next page.

Q23 (continued)

For each statement below, identify whether it is true about `mod23a`, `mod23b`, both models, or neither model.

Columns:

1. `mod23a`
2. `mod23b`
3. both models
4. neither model

Rows:

a. The model predicts the probability that a subject will receive a stent.
b. If subjects A and B have the same `abcix` status, but A has one more `ves1proc` than B, A is predicted to have a larger probability of receiving a stent.
c. If subjects A and B have the same `abcix` status, but A has one more `ves1proc` than B, A is predicted to have a smaller probability of receiving a stent.
d. This is a logistic regression model.

Questions Q24 and Q25 use the `dat24` data set

286 male patients were examined. Each exhibited one of several reasons to suspect problems with their prostate glands. These data are available in the `dat24.csv` data set. For each patient, the following data are provided:

- `ptnum` = patient identification code
- `age` = age (in years)
- `dre` = digital rectal examination result (0 = negative, 1 = positive)
- `tru` = transurethral ultrasound result (0 = negative, 1 = positive)
- `psa` = prostate-specific antigen level (in ng/ml)
- `vol` = volume of prostate (in ml)
- `psad` = prostate-specific antigen density level (this is just psa / vol)
- `biopsy` = biopsy result (0 = negative, 1 = positive)

```
dat24 <- read_csv("data/dat24.csv", show_col_types = FALSE) |>
  clean_names() |>
  mutate(ptnum = as.character(ptnum))
```

```
summary(dat24)
```

```
   ptnum                age              dre              tru        
 Length:286         Min.   :47.00   Min.   :0.0000   Min.   :0.0000  
 Class :character   1st Qu.:61.00   1st Qu.:0.0000   1st Qu.:0.0000  
 Mode  :character   Median :66.00   Median :1.0000   Median :0.0000  
                    Mean   :66.72   Mean   :0.6084   Mean   :0.4755  
                    3rd Qu.:72.75   3rd Qu.:1.0000   3rd Qu.:1.0000  
                    Max.   :91.00   Max.   :1.0000   Max.   :1.0000  
      psa               vol              psad            biopsy      
 Min.   :  0.300   Min.   :  3.30   Min.   :0.0100   Min.   :0.0000  
 1st Qu.:  3.125   1st Qu.: 24.59   1st Qu.:0.0800   1st Qu.:0.0000  
 Median :  5.850   Median : 32.80   Median :0.1600   Median :0.0000  
 Mean   :  8.928   Mean   : 36.67   Mean   :0.2729   Mean   :0.3147  
 3rd Qu.:  8.000   3rd Qu.: 43.80   3rd Qu.:0.2800   3rd Qu.:1.0000  
 Max.   :221.000   Max.   :114.03   Max.   :4.5500   Max.   :1.0000  
```

The outcome which we are interested in predicting is the `biopsy` result, which we will assume indicates the "truth" in this case as to whether the patient actually has prostate cancer.

24 Q24 (5 points)

To begin, use the `dat24` data to build a regression model to predict whether the patient actually has prostate cancer on the basis of their PSA level, prostate volume, transurethral ultrasound result, digital rectal examination result and age.

An increase in which of the following predictors show a positive association (relative odds greater than 1) with our outcome? (Note that more than one response may be selected, and that this question has nothing to do with the notion of statistical significance.)

a. the subject's age
b. the subject's prostate-specific antigen level
c. the subject's prostate volume
d. the result of the subject's transurethral ultrasound
e. the result of the subject's digital rectal exam
f. None of the above

25 Q25 (3 points)

Suppose you decide to use a cutpoint of a fitted probability of **0.3 or higher** for `biopsy` as your prediction rule to predict that the patient actually should be further screened for prostate cancer. Create a confusion matrix for the model you developed in Question Q24. Use that matrix to specify:

a. the sensitivity
b. the specificity
c. the positive predictive value

under the prediction rule we've specified above. Specify your responses as **proportions** rounded to two decimal places.

26 Q26 (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? (SELECT ALL THAT APPLY.)

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

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.