432 Quiz A

Thomas E. Love

Deadline: 9 PM 2021-03-22. Version: 2021-03-10 11:48:03

Links

All links for the Quiz will be made available at https://github.com/THOMASELOVE/432-2021/tree/master/quizzes/quizA on the morning of 2021-03-12.

This will include links to:

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

Instructions

This PDF document is 25 pages long. There are 26 questions on this Quiz. It is to your advantage to answer all 26 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.

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 2021-03-12. All of your answers must be submitted through the Google Form by **9 PM** on Monday 2021-03-22, without exception. The form will close at that time, and no extensions 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.

The Data Sets

I have provided three data sets (called `set01.csv`, `set13.csv` and `set20.csv`) that are mentioned in the Quiz. They may be helpful to you.

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

During the Quiz period (2021-03-12 through 2021-03-22) 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 2021-03-22 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 Piazza in the `quiza` folder if we find an error in the Quiz that needs fixing.

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.

Scoring and Timing

All questions are worth 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. We'll warn you that several of the "longer" questions come early in the Quiz, in particular, we expect Questions 1-5, 8-9 and 11 to take "longer" to complete than the median question.

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 2021-03-25.

What does the Quiz cover?

Quiz A includes material from the first 9 classes in 432, as well as:

- Chapters 1-14 of the 432 course notes,
- all of Jeff Leek's *How to be a modern scientist* and
- Chapters 1, 2, 6 and 7 of Nate Silver's *The Signal and the Noise*.

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. 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 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.

```r
library(knitr)
library(janitor)
library(magrittr)
library(naniar)
library(simputation)
library(rms)
library(broom)
library(tidyverse)

# Note that all data files were downloaded onto
# my machine into a subfolder called data below
# my main R Project directory for Quiz A.

theme_set(theme_bw())
opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)
```

The set01 data (Q01 - Q11)

The data in the set01.csv file contain information for 235 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 set01 and use that tibble to respond to questions Q01 through Q11.

1 Q01 (4 points)

Using your set01 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 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 m1 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 m1 model, by running summary(m1). Some of the output from my version of that summary appears below.

Your task on the answer sheet for Q01 is to specify that AIC value (rounded to zero decimal places.)

Output for Q01

An appropriate analyses yields the following results. Note that you do not need to filter for complete cases in Questions Q01-Q05.

```
set01 <- read_csv("data/set01.csv") %>%
    mutate(goodout = ifelse(outcome == "Good", 1, 0),
           subject = as.character(subject))
```

Note that the fitting of the actual m1 and the creation of ses_grpf are not shown here.

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

```
 (Intercept)         size treatmentYes    ses_grpf2    ses_grpf3    ses_grpf4 
  0.03890977   1.01627013   1.91866871   2.15138271   2.35388280   2.93597422 
   ses_grpf5 
  3.58978368 
```

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

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

                   2.5 %     97.5 %
(Intercept)  0.008941697  0.1416091
size         1.003823118  1.0293922
treatmentYes 1.052768308  3.5236890
ses_grpf2    0.580207853  8.4766178
ses_grpf3    0.806354173  7.9313634
```

```
ses_grpf4    1.036378012  9.7167827
ses_grpf5    1.172567071 12.5724575
```

Here is a partial listing of the summary of the fitted `m1` I created.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3751  -0.8860  -0.6384   1.2100   2.1932

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 270.60  on 217  degrees of freedom
Residual deviance: 250.95  on 211  degrees of freedom
  (17 observations deleted due to missingness)
```

2 Q02 (5 points)

Interpret the `treatmentYes` value of 1.92 specified in the output for Q01 above, using a complete English sentence or two.

3 Q03 (5 points)

What does the provided 95% confidence interval (1.05, 3.52) for `treatmentYes` in the output for Q01 tell you about the `treatment` variable? Provide your response in the form of 1-2 complete English sentences.

4 Q04 (5 points)

Consider the Output for Q04, provided below. Why is the odds ratio shown in the Output for Q04 referring to `size` different from that shown in the earlier presentation (in the Output for Q01) of `exp(coef(m1))` for the `size` variable in the same model? Again provide your response in the form of 1-2 complete English sentences.

Output for Q04

The output below comes from another approach to fitting the identical logistic regression model that we saw in Q01, still using only the complete cases. I'll call this model `m1L`, to emphasize that it contains the same outcome and predictors as were used in `m1`.

```
summary(m1L)
```

```
             Effects              Response : goodout 

 Factor             Low   High  Diff. Effect   S.E.    Lower 0.95 Upper 0.95
 size               60.55 96.25 35.7   0.57617 0.22816  0.128990  1.023400  
  Odds Ratio        60.55 96.25 35.7   1.77920      NA  1.137700  2.782500  
 treatment - Yes:No  1.00  2.00   NA   0.65163 0.30736  0.049216  1.254000  
  Odds Ratio         1.00  2.00   NA   1.91870      NA  1.050400  3.504500  
 ses_grpf - 1:4      4.00  1.00   NA  -1.07700 0.56113 -2.176800  0.022762  
  Odds Ratio         4.00  1.00   NA   0.34060      NA  0.113400  1.023000  
 ses_grpf - 2:4      4.00  2.00   NA  -0.31093 0.53488 -1.359300  0.737410  
  Odds Ratio         4.00  2.00   NA   0.73277      NA  0.256850  2.090500  
 ses_grpf - 3:4      4.00  3.00   NA  -0.22097 0.40135 -1.007600  0.565670  
  Odds Ratio         4.00  3.00   NA   0.80174      NA  0.365090  1.760600  
 ses_grpf - 5:4      4.00  5.00   NA   0.20105 0.43418 -0.649930  1.052000  
  Odds Ratio         4.00  5.00   NA   1.22270      NA  0.522080  2.863500  
```

5 Q05 (4 points)

Again ignoring missingness in the `set01` 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 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.

6 Q06 (3 points)

How many subjects in the `set01` tibble are missing data in at least one variable?

7 Q07 (3 points)

How many missing observations are there on the outcome for your logistic regression models in the `set01` tibble?

Setting Up Q08 - Q11

Note that in Questions Q08-Q11, you will again be using the `set01` data, and you will fit a new model (which we'll call `m2`) adding in the non-linear component that you specified in Q05 to what was fit in `m1` and `m1L`, while also accounting for missing data using **multiple imputation**.

8 Q08 (4 points)

The code listed below uses the `aregImpute()` function to fit a multiple imputation model, using `set.seed(2021)`.

```r
set.seed(2021)
set01_imp <- aregImpute(~ goodout + treatment + ses_grpf + size,
                        nk = 0, data = set01, 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 treatment variable
c. the size variable
d. the goodout variable
e. It is impossible to tell.

9 Q09 (5 points)

Fit the outcome model called `m2` using `fit.mult.impute()`. Your `m2` model should incorporate the multiple imputations from Q08 that you stored in `set01_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 Q05. Your fit of model `m2` 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 `m2`. In the Answer Sheet, your code should begin with

```
m2 <- fit.mult.impute(
```

10 Q10 (3 points)

What is the in-sample estimated area under the ROC curve for your `m2`, rounded to three decimal places?

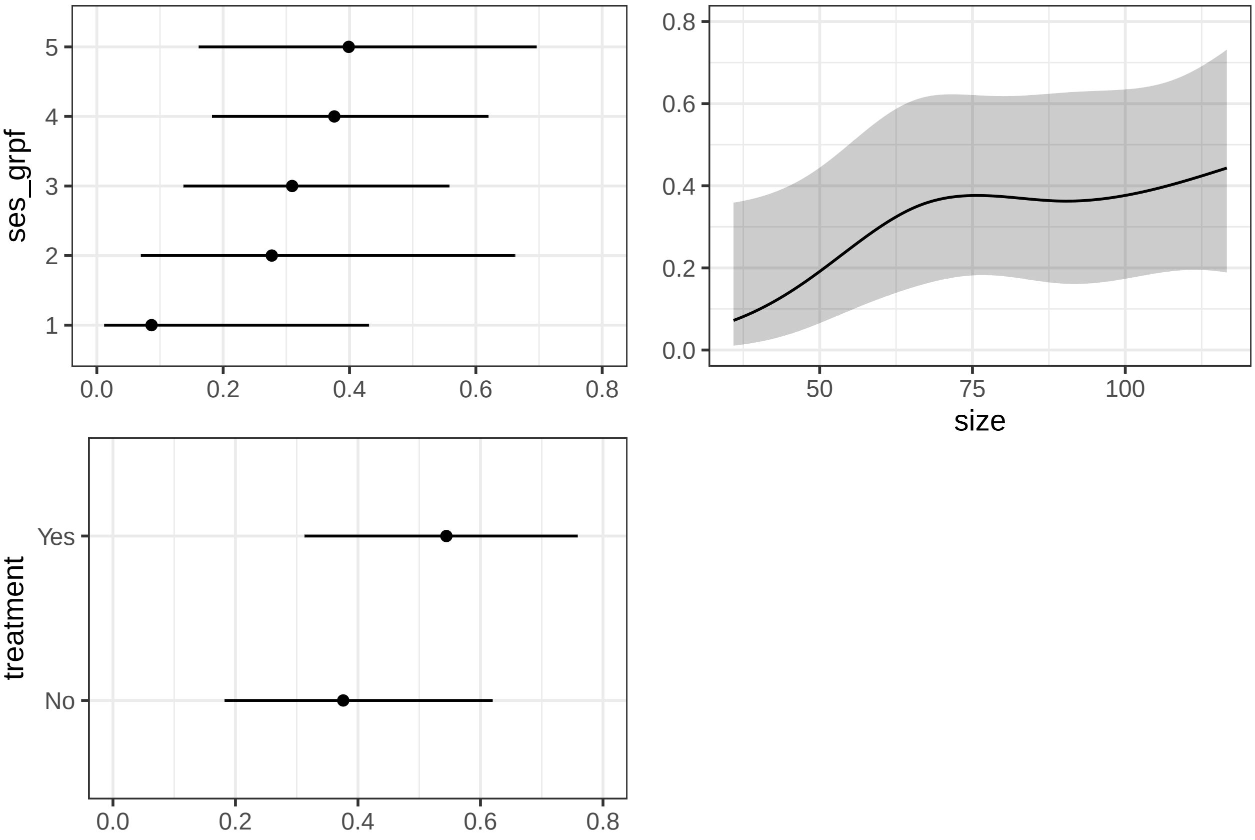
11 Q11 (4 points)

Consider the four sets of plots for Q11 printed below, 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 `m2` model?

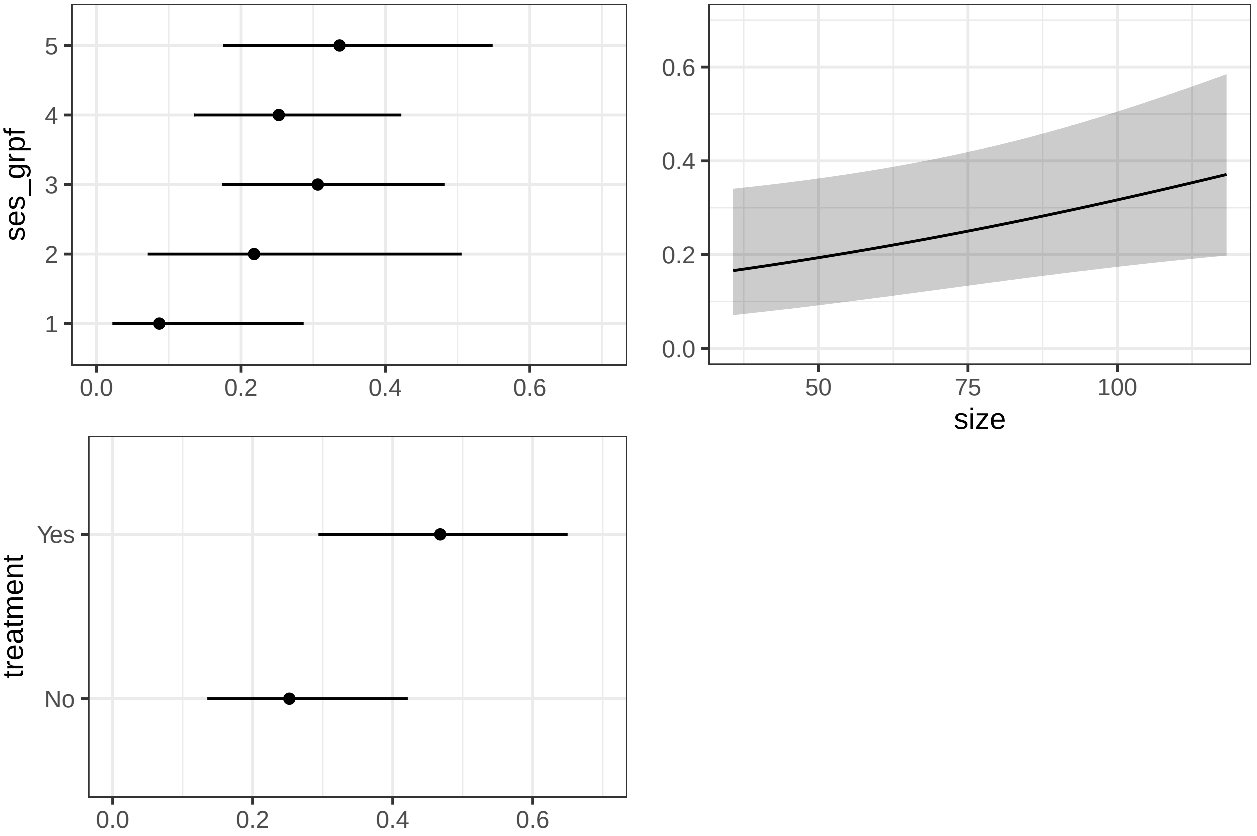
a. Plot Sets A and C
b. Plot Sets A and D
c. Plot Sets B and C
d. Plot Sets B and D

Just to confirm, exactly two of these plots do come from `m2` and the other two do not.

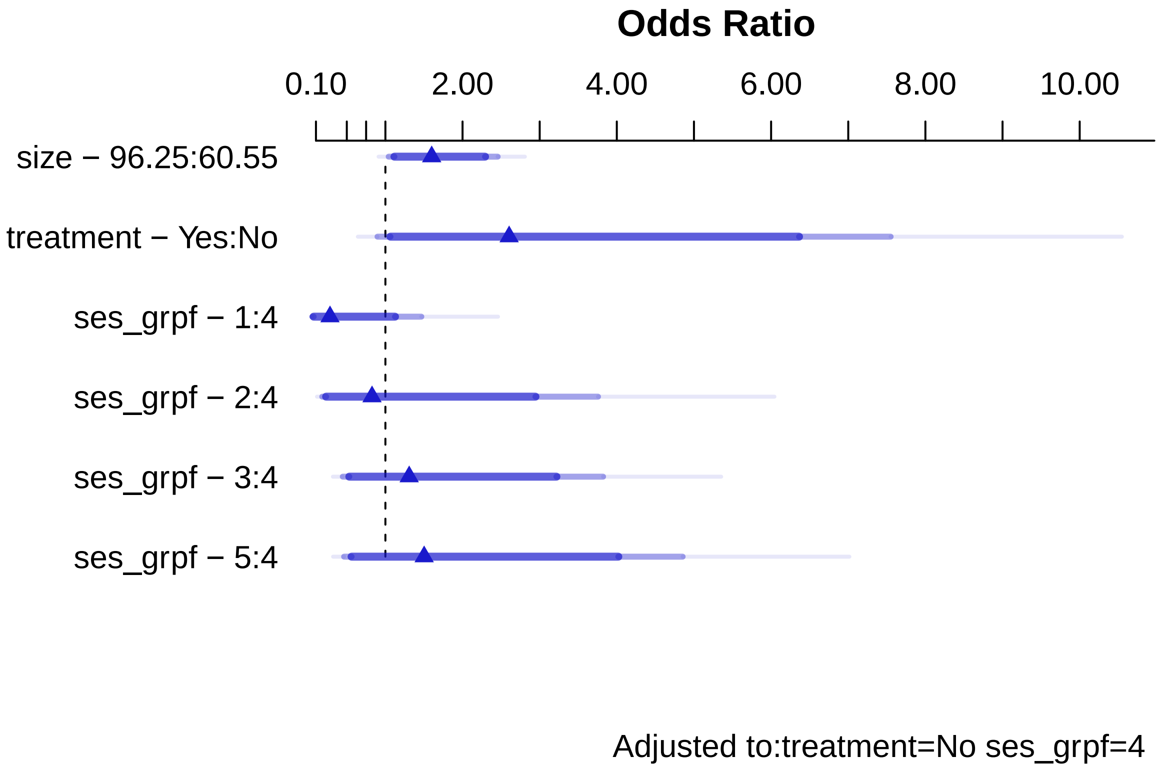
Plot Set A for Q11



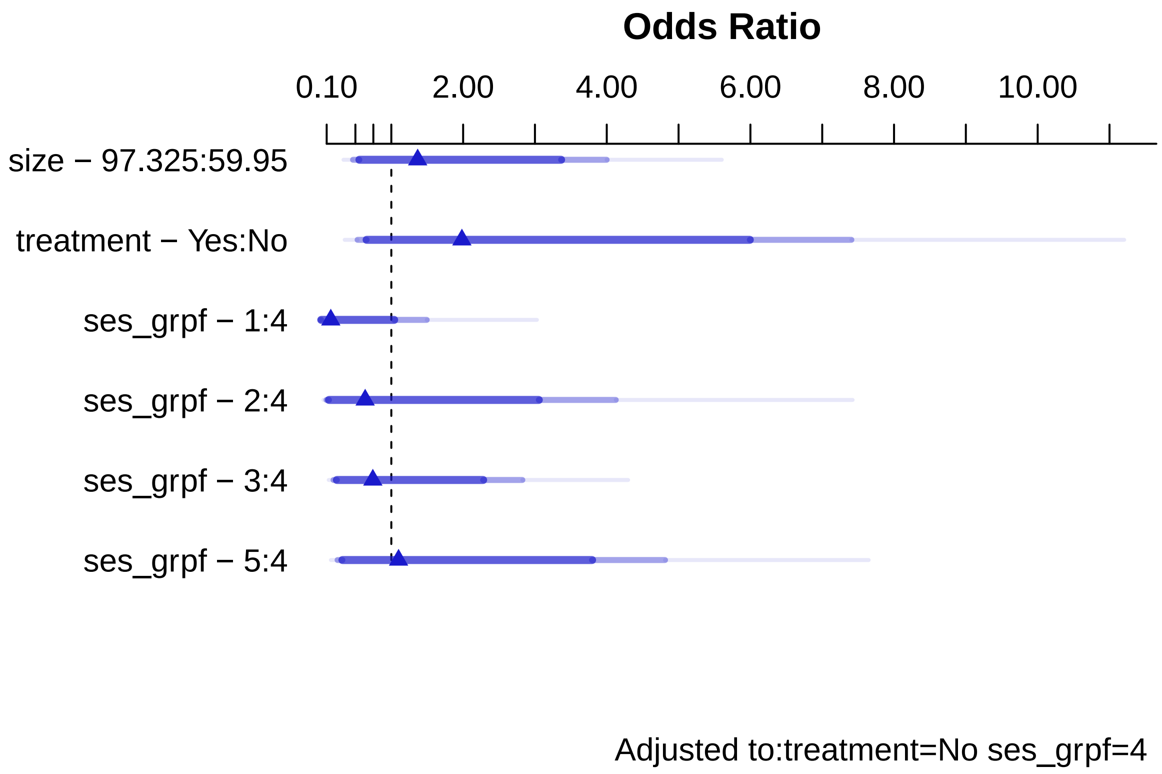
Plot Set B for Q11



Plot Set C for Q11



Plot Set D for Q11



This is the end of the output for Q11.

12 Q12 (4 points)

In *The Signal and the Noise*, Nate Silver encourages his readers to behave like foxes, as compared to hedgehogs, when forecasting. Which of the following statements describe "foxes"? (SELECT ALL THAT APPLY.)

a. Someone willing to acknowledge mistakes in their predictions.
b. Someone who relies more on observation than on grand theories.
c. Someone who expects the world to abide by relatively simple relationships once the signal has been separated from noise.
d. Someone who establishes a planned approach at the start, and uses new data just to refine the original plan.
e. Someone who is willing to bet on their forecasts.
f. Someone who is highly focused on learning about one particular subgroup of information as thoroughly as possible.

The `set13` data (Q13-Q17)

The `set13.csv` data file will be used for Questions 13-17. I'd ingest the data into R as a tibble called `set13`, containing three variables.

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

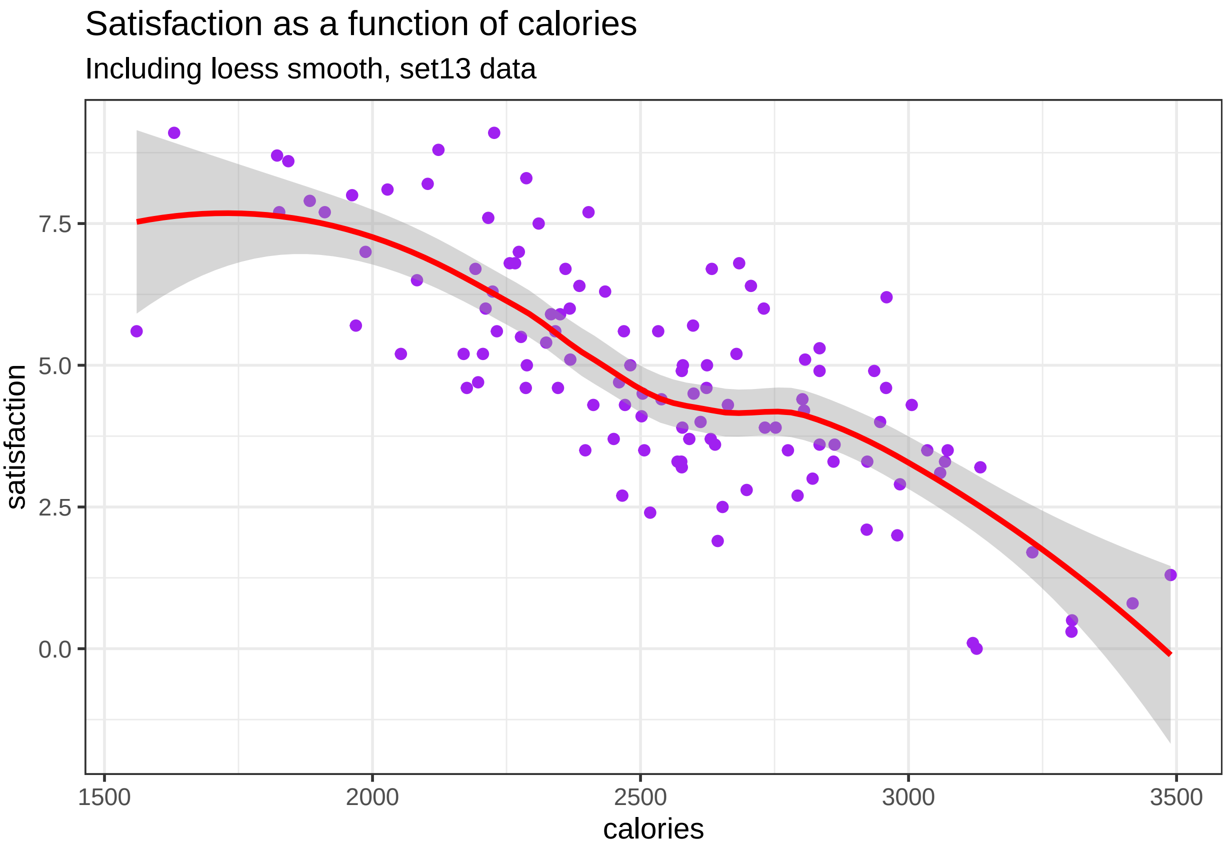
13 Q13 (3 points)

How many of the subjects in `set13` have both `calories` above 2500 and `satisfaction` below 5?

14 Q14 (5 points)

Using the `set13.csv` data set, a student attempted unsuccessfully to generate the Q14 Target Plot shown below, in R, developing the code shown in the Q14 Code Attempt also shown below. Explain, in a couple of sentences, how you would FIX the code in the Q14 Code Attempt to generate the Q14 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.

Q14 Target Plot



Q14 Code Attempt

```
ggplot(set13, 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, set13 data")
```

15 Q15 (3 points)

Using the `set13` tibble, specify the code required to fit (using `lm`) a model called `m15` that predicts the `satisfaction` score across these subjects using an orthogonal polynomial of degree 3 in the `calories` variable. Then summarize the `m15` model you built in Q15.

Now, what is the observed R^2 value for your model `m15`, expressed as a proportion, and rounded to three decimal places?

16 Q16 (3 points)

A new model in R (which I'll call `m16`) was fit to the `set13` data, now using an orthogonal polynomial of degree 2. The `glance` function applied to `m16` shows an AIC of 388.3 and a BIC of 399.3. Compare these results to `m15`. Which of the following conclusions is most appropriate based on these results?

a. The cubic term in Model `m15` is not helpful according to either AIC or BIC.
b. The cubic term in Model `m15` is helpful according to exactly one of AIC or BIC.
c. The cubic term in Model `m15` is helpful according to both AIC and BIC.
d. None of these conclusions are appropriate.

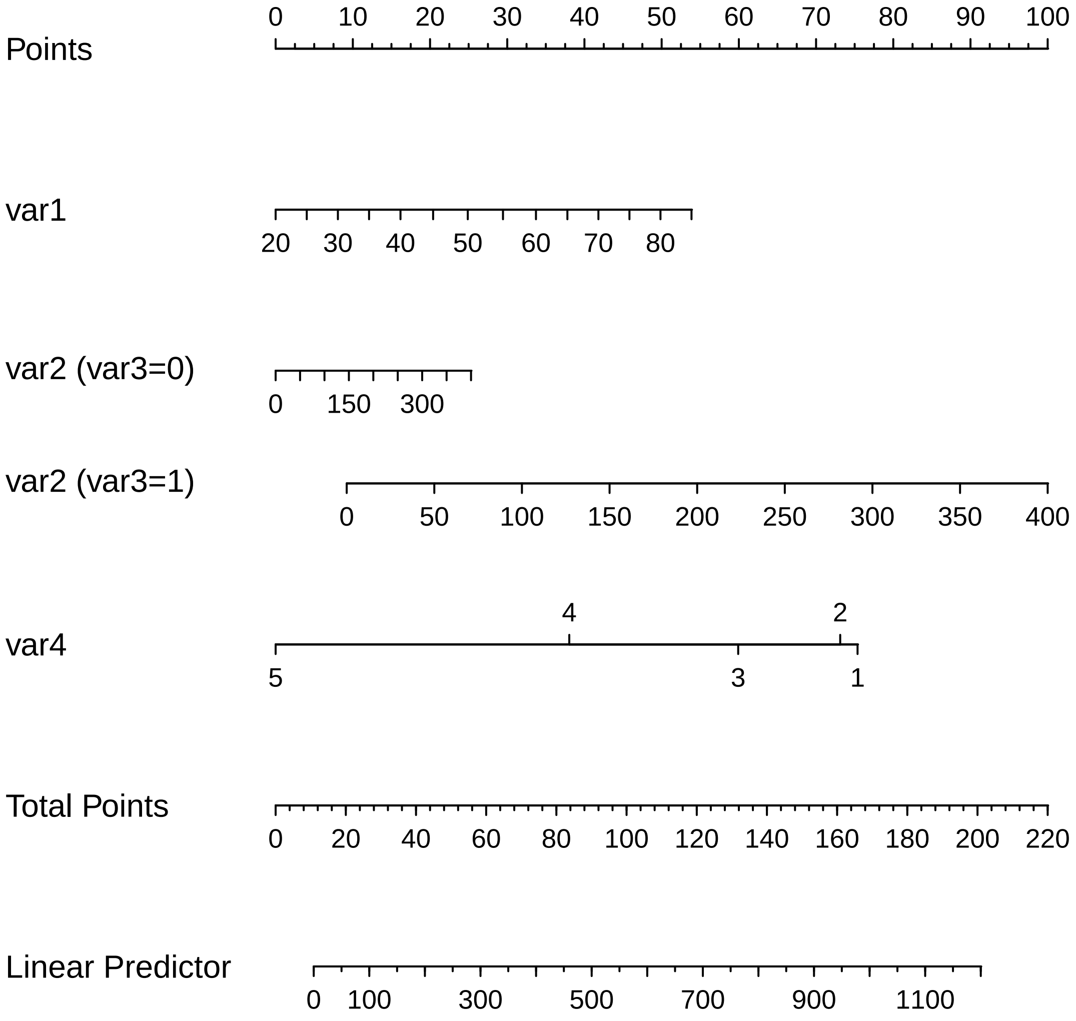
17 Q17 (4 points)

In Q17, we look at a new model. Use the nomogram in the Output for Q17 below to make a prediction about the outcome variable (which is measured in hours) for two subjects. They each have `var4` = 3, but Noah has `var1` = 45, `var2` = 150 and `var3` = 0. 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 days 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 Q17



This is the end of the output for Q17.

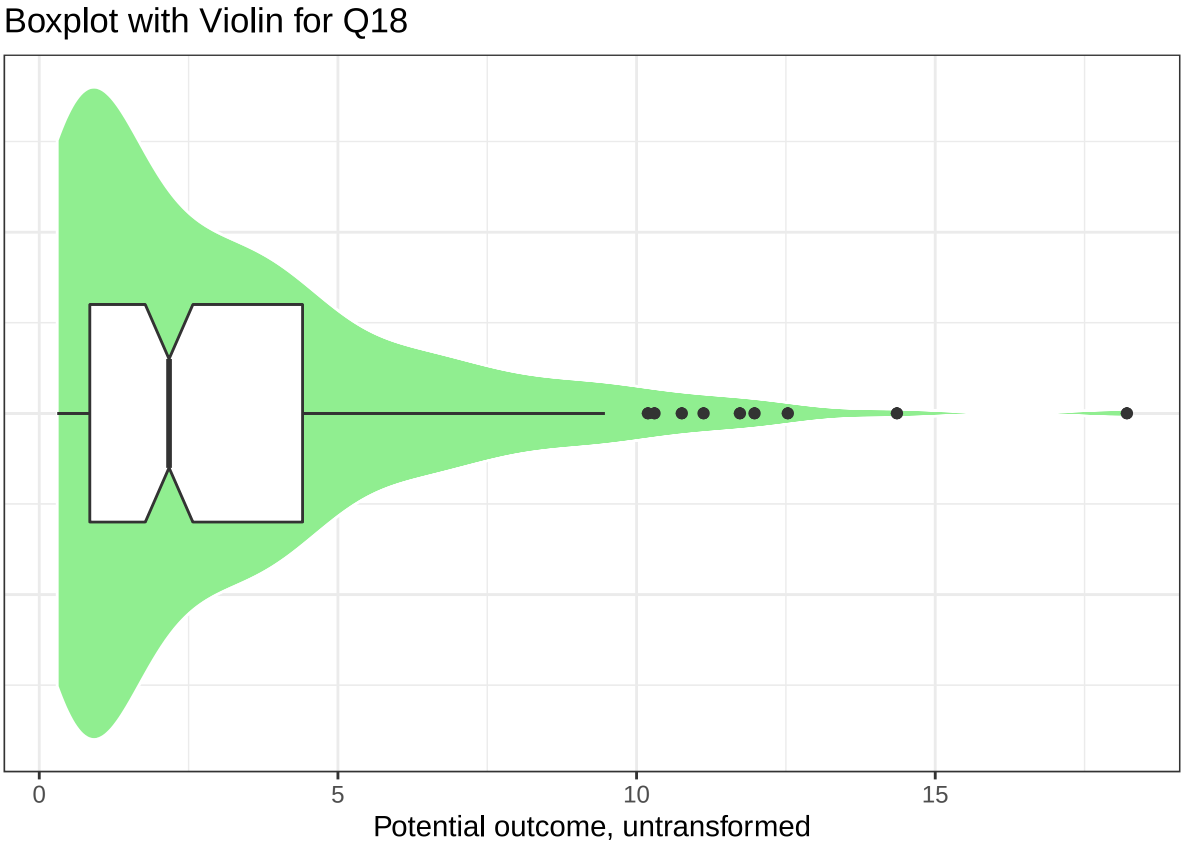
18 Q18 (3 points)

Consider the information provided below (in the Output for Q18) on the distribution of a potential outcome variable in a linear regression model to be built using the `dat18` tibble. Note that I have deliberately not provided you with these data.

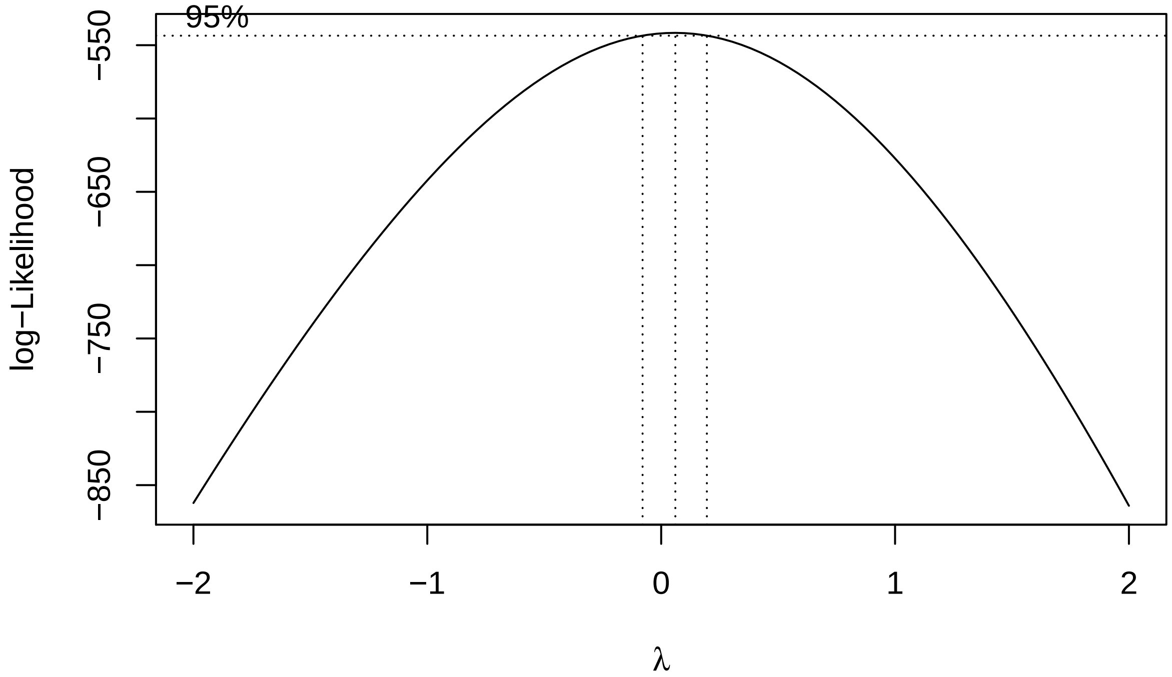
Based on the three pieces of output provided, which of the following transformations of the `outcome` data would be most appropriate?

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.

Output 1 of 3 for Q18



Output 2 of 3 for Q18: Box-Cox plot



Output 3 of 3 for Q18

Table 1: Data summary

Name	Piped data
Number of rows	200
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
outcome	0	1	3.15	3.11	0.3	0.85	2.17	4.41	18.21

This is the end of the output for Q18.

19 Q19 (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

20 Q20 (5 points)

The `set20.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 `set20`.

Suppose you now want to create a variable called `gov_ins` within the `set20` 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 Q20. Fix the call to the `mutate` function in that code so that your resulting code will actually do what is required. Your response should begin with `mutate(gov_ins =` on the answer sheet.

Code Attempt for Q20

```
set20 <- set20 %>%
    type.convert() %>%
    mutate(gov_ins = factor(insurance,
                            Medicare or Medicaid = Yes,
                            Commercial or Uninsured = No))
```

21 Q21 (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 Q21 below, 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 = dat21, x = TRUE, y = TRUE)`.

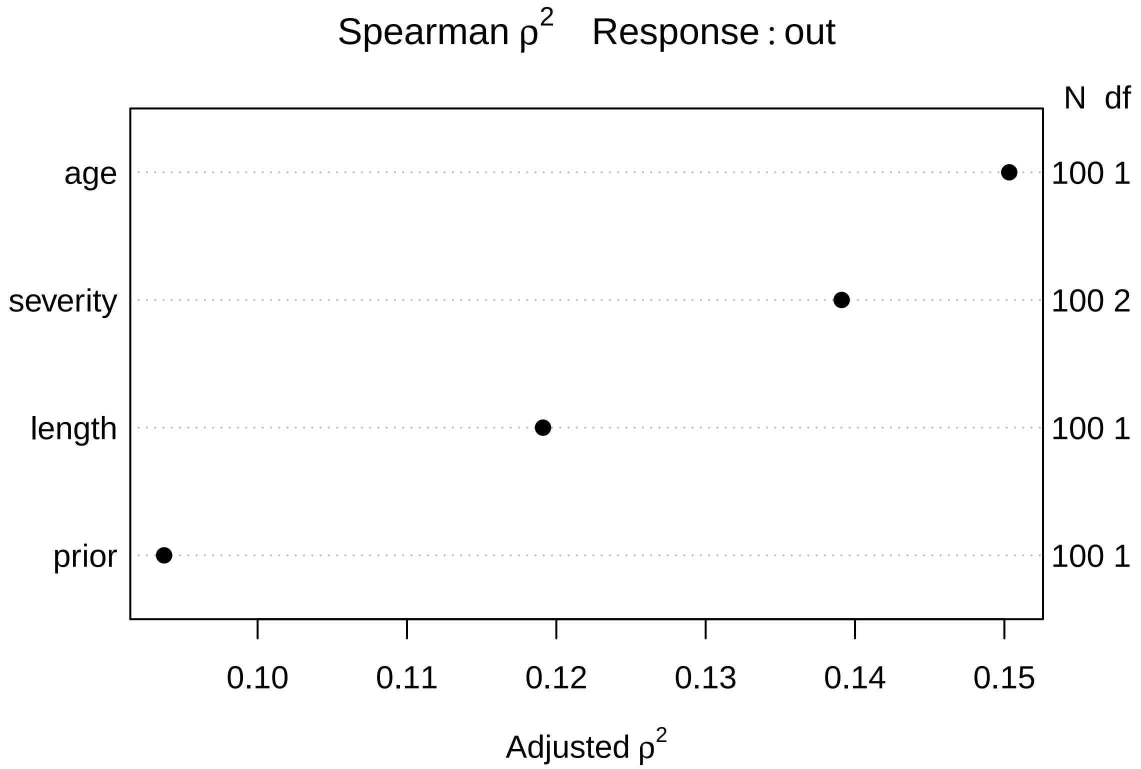
So the actual specification for Model A, for example, would be

```
dd <- datadist(dat21); options(datadist = "dd")
modelA <- ols(out ~ age*prior + severity*length,
              data = dat21, x = TRUE, y = TRUE)
```

Now, which of the models specified above does the best job of meeting the requirements for Q21?

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

Spearman Plot for Q21



This is the end of the output for Q21.

22 Q22 (3 points)

In Q22, 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 Q22. Which of the models shown in the Output for Q22 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 Q22

```
d <- datadist(dat22)
options(datadist = "d")

m22w <- ols(outcome ~ c + d,
           data = dat22, x = TRUE, y = TRUE)
m22x <- ols(outcome ~ c + d + e + f,
           data = dat22, x = TRUE, y = TRUE)
m22y <- ols(outcome ~ a + b + c + d + e + f + g,
           data = dat22, x = TRUE, y = TRUE)
m22z <- ols(outcome ~ c + d + g,
            data = dat22, x = TRUE, y = TRUE)
```

```
set.seed(4321); validate(m22w)
```

```
          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(m22x)
```

```
          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(m22y)
```

```
          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(m22z)
```

```
          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 Q22.

23 Q23 (4 points)

In the early chapters of *The Signal and the Noise*, Nate Silver describes several of the challenges involved in making a variety of different types of forecasts. Which of the following statements are backed up by evidence in Chapters 1, 2 and 6 of the book? (SELECT ALL THAT APPLY.)

a. Tracking forecasts is of great interest to the scientific community,
b. In this century, there is increasing demand from the public for accuracy in forecasts.
c. Throwing out data that does not match expectations materially improves the quality of forecasts based on those data.
d. Grouping sets of forecasts and taking an average typically outperforms the forecasts of individuals.
e. The amount of confidence with which a person makes a forecast is a strong positive indicator of that forecast's accuracy.

24 Q24 (4 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 `model24A` and `model24B`. 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 Q24 provided below. What was included in `model24B` but not included in `model24A`?

24.1 Output for Q24

```
anova(model24A)
```

```
Analysis of Variance Table

Response: health
           Df Sum Sq Mean Sq F value    Pr(>F)    
treatment   3  59583 19861.1 16.8301 3.053e-10 ***
insurance   2  22141 11070.5  9.3811 0.0001072 ***
Residuals 354 417753  1180.1                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model24A, model24B)
```

```
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 417753                              
2    348 399070  6     18683 2.7154 0.01368 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the end of the output for Q24.

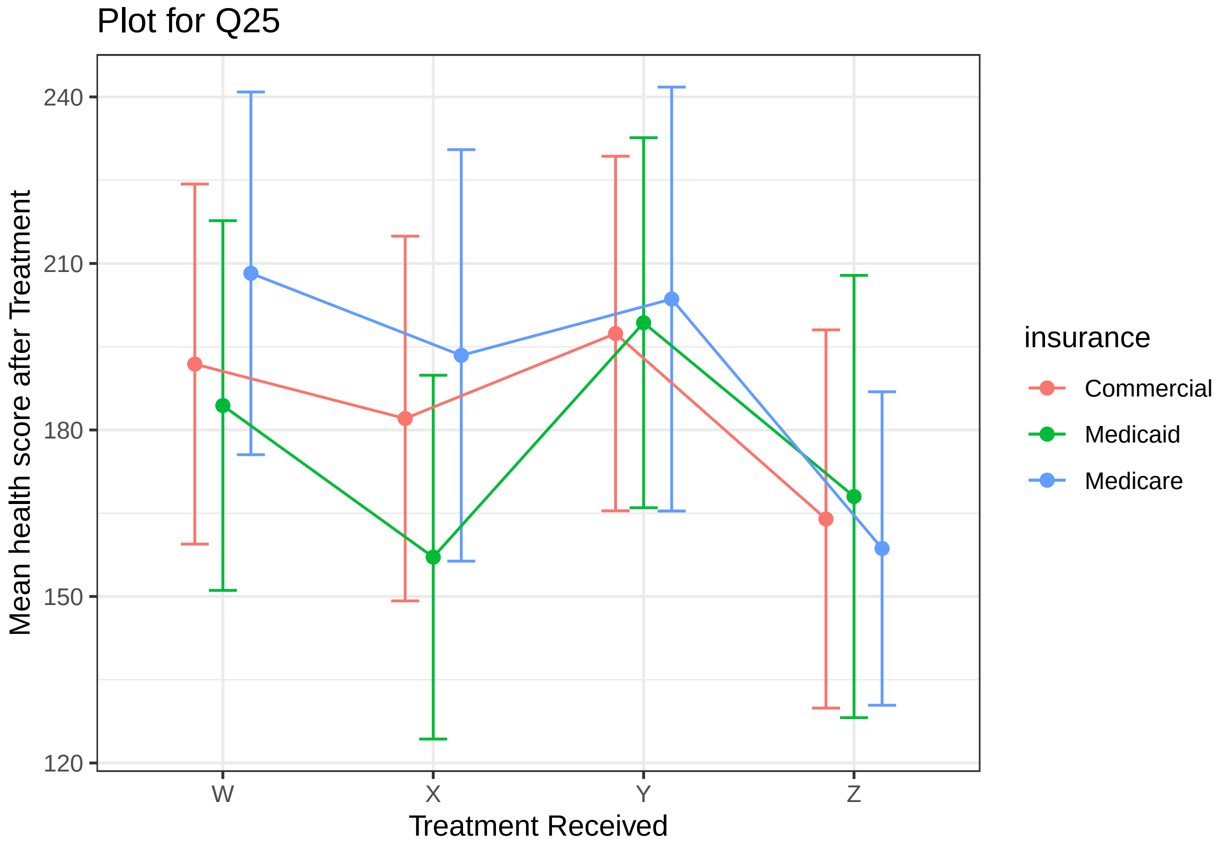
25 Q25 (3 points)

Consider again the situation described in Q24. In the Output for Q25 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 Q25 suggest about the best choice of model, comparing `model24A` to `model24B`?

a. `model24A` seems like the better choice.
b. `model24B` seems like the better choice.
c. This plot does not help us make the decision.

Output for Q25



This is the end of the output for Q25.

26 Q26 (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.

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.