# 432 Quiz 2 Answer Sketch

Thomas E. Love

Deadline: 9 AM 2022-04-19. Version: 2022-04-18 18:43:44

## Links

All links for the Quiz will be made available at https://github.com/THOMASELOVE/432-2022/tree/main/quiz/quiz2 at 5 PM on 2022-04-14.

This will include links to:

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

## Instructions

This PDF document is **31** pages long. There are **30** 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.

### The Google Form Answer Sheet

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.

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 **5** data sets (called **quizd01.csv**, **quizd08.Rds**, **quizd17.Rds**, **quizd24.csv** and **quizd30.Rds**) 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 (2022-04-14 through 2022-04-19) we will not answer questions about the Quiz except through the two approaches listed above. We promise to respond to all questions received before 9 PM on 2022-04-18 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 `quiz2` 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 between **2** and **5** points, 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 4-6 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 2022-04-21. Remember that we have a Project B working day (with no 432 class) on Tuesday 2022-04-19.

## What does the Quiz cover?

Quiz B includes material from the first 22 classes in 432, including:

- all of Jeff Leek's *How to be a Modern Scientist* and
- all 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. 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 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.

```
library(conflicted)
library(here)
library(knitr)
library(janitor)
library(magrittr)
library(naniar)
library(patchwork)
library(countreg)
library(equatiomatic)
library(fivethirtyeight)
library(GGally)
library(ggrepel)
library(lars)
library(nhanesA)
library(nnet)
library(pROC)
library(pscl)
library(simputation)
library(survey)
library(survival)
library(survminer)
library(rms)        # includes Hmisc
library(tidymodels) # includes broom, rsample, etc.
library(tidyverse)  # includes dplyr, ggplot, etc.

# Note that all data files were downloaded onto
# my machine into a subfolder called data below
# my main R Project directory for Quiz 2.

theme_set(theme_bw())
opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)
conflict_prefer("summarize", "dplyr")
conflict_prefer("filter", "dplyr")
conflict_prefer("select", "dplyr")
conflict_prefer("zeroinfl", "pscl")
```

## Setup for Questions 1-3

The `quizd01.csv` file available on the course web site contains information on 1000 animal subjects who took part in an observational study. You will use this data set for Questions 1-3. The data includes information on:

- `alive` = the subject's vital status at the end of the study (`alive` = 1 if alive at the end of the study, 0 otherwise)
- `treated` = 1 if the subject received the treatment of interest and 0 if the subject received usual care
- `age`, in years, at the start of the study
- `female` = 1 if the subject is female, biologically
- `comor` = a count of comorbid conditions (maximum = 7)

# 1 Question 1

Create a tibble called `quizd01` in R containing the data in `quizd01.csv`. How many rows in your `quizd01` data contain at least one missing value?

## 1.1 Answer 1 is 97

```
quizd01 <- read_csv(here("data/quizd01.csv"))
```

```
Rows: 1000 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (6): subject, alive, age, treated, comor, female

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
miss_case_table(quizd01)
```

```
# A tibble: 3 x 3
  n_miss_in_case n_cases pct_cases
           <int>   <int>     <dbl>
1              0     903      90.3
2              1      94       9.4
3              2       3       0.3
```

There are 94 cases missing one observation, plus 3 more missing two. That's 97.

# 2 Question 2

Specify the R code you would use to fit a logistic regression model to predict `alive` on the basis of main effects of `treated`, `age`, `female` and `comor`, using multiple imputation to deal with missing values, and setting a seed of `43237` for the imputation work. In your response, don't show any of the code you used to create the tibble in Question 1. Assume all necessary packages have been pre-loaded using the `library()` function, and that the `quizd01` data have been successfully imported into R already.

In your imputation process, you should include all variables in the `quizd01` data other than the subject identifying code, run 20 imputations, and use `nk = c(0, 3), tlinear = TRUE, B = 10` and `pr= FALSE`. Do not show the results here, just the code.

## 2.1 Answer 2 is a lengthy piece of code.

You'll need to have:

- done multiple imputation, and
- included `alive`, `treated`, `age`, `female` and `comor` in the imputation model, and
- fit the outcome model using `fit.mult.impute`

Here's what I used.

```
set.seed(43237)

d <- datadist(quizd01)
options(datadist = "d")

imp_fit2 <- aregImpute(~ alive + treated + age + female + comor,
                       nk = c(0,3), tlinear = TRUE, data = quizd01,
                       B = 10, n.impute = 20, pr = FALSE)

m2_imp <- fit.mult.impute(alive ~ treated + age + female + comor,
                          fitter = lrm, xtrans = imp_fit2,
                          data = quizd01, x = T, y = T)
```

The result of applying this is:

```
set.seed(43237)

d <- datadist(quizd01)
options(datadist = "d")

imp_fit2 <- aregImpute(~ alive + treated + age + female + comor,
                       nk = c(0,3), tlinear = TRUE, data = quizd01,
                       B = 10, n.impute = 20, pr = FALSE)

m2_imp <- fit.mult.impute(alive ~ treated + age + female + comor,
                          fitter = lrm, xtrans = imp_fit2,
                          data = quizd01, x = T, y = T)
```

```
Variance Inflation Factors Due to Imputation:

Intercept   treated       age    female     comor
     1.09      1.15      1.32      1.03      2.97

Rate of Missing Information:

Intercept   treated       age    female     comor
     0.09      0.13      0.25      0.03      0.66

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

Intercept   treated       age    female     comor
  2621.01   1174.66    316.47  17180.85     43.14

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

  stats linear.predictors
```

```
m2_imp
```

```
Logistic Regression Model

 fit.mult.impute(formula = alive ~ treated + age + female + comor,
     fitter = lrm, xtrans = imp_fit2, data = quizd01, x = T, y = T)

                       Model Likelihood    Discrimination    Rank Discrim.
                             Ratio Test           Indexes          Indexes
 Obs          1000    LR chi2     348.49    R2       0.430    C       0.856
  0            738    d.f.             4    g        2.079    Dxy     0.712
  1            262    Pr(> chi2) <0.0001    gr       8.048    gamma   0.713
 max |deriv| 8e-06                          gp       0.276    tau-a   0.276
                                            Brier    0.123

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept  6.1694 0.6452   9.56 <0.0001
 treated    1.3981 0.2003   6.98 <0.0001
 age       -0.1871 0.0161 -11.61 <0.0001
 female    -0.0357 0.1791  -0.20 0.8420
 comor      0.3729 0.1086   3.43 0.0006
```

Notes:

- There was no reason to filter to complete cases on `alive` before doing this imputation.
- You needed to use `lrm` as the fitter in `fit.mult.impute` rather than `glm`.
- You needed to use the `set.seed` seed that I specified, which was 43237.
- You needed to include precisely the variables I did in each model.
- You needed to use the data frame called `quiz2C` and not something else, like `quiz2c`.
- You needed to fit both `aregImpute` and `fit.mult.impute`.

# 3 Question 3

Using the model you specified in Question 2, estimate the effect of treatment (vs. control) on the odds of being alive at the end of the study. Your odds ratio estimate should compare `treated` to `control`, while adjusting for the effects of `age`, `female` and `comor`. Provide both a point estimate and a 95% confidence interval. Use the format 1.23 (0.78, 1.89) for your response, rounding your estimates to two decimal places. Then interpret your point estimate concisely and correctly in complete English sentences. Do not include any R code or output in your response.

## 3.1 Answer 3 is 4.05, with 95% CI (2.73, 5.99) for the odds ratio.

To receive full credit, you'd have to describe this as an odds ratio appropriately, mentioning the key predictor (and the direction of the effect) and the outcome, and correctly specify the variables that are adjusted for in the model.

We can read the odds ratio estimate comparing `treated` to `control`, while adjusting for the effects of `age`, `female` and `comor` using the summary of the imputation model displayed below.

```
summary(m2_imp)
```

```
             Effects              Response : alive
```

| Factor | Low | High | Diff. | Effect | S.E. | Lower 0.95 | Upper 0.95 |
|---|---|---|---|---|---|---|---|
| treated | 0 | 1 | 1 | 1.398100 | 0.20027 | 1.005500 | 1.79060 |
| Odds Ratio | 0 | 1 | 1 | 4.047400 | NA | 2.733400 | 5.99310 |
| age | 43 | 57 | 14 | -2.619900 | 0.22567 | -3.062200 | -2.17760 |
| Odds Ratio | 43 | 57 | 14 | 0.072813 | NA | 0.046786 | 0.11332 |
| female | 0 | 1 | 1 | -0.035722 | 0.17915 | -0.386840 | 0.31540 |
| Odds Ratio | 0 | 1 | 1 | 0.964910 | NA | 0.679200 | 1.37080 |
| comor | 1 | 3 | 2 | 0.745860 | 0.21728 | 0.320000 | 1.17170 |
| Odds Ratio | 1 | 3 | 2 | 2.108200 | NA | 1.377100 | 3.22750 |

So an example of a good response would be:

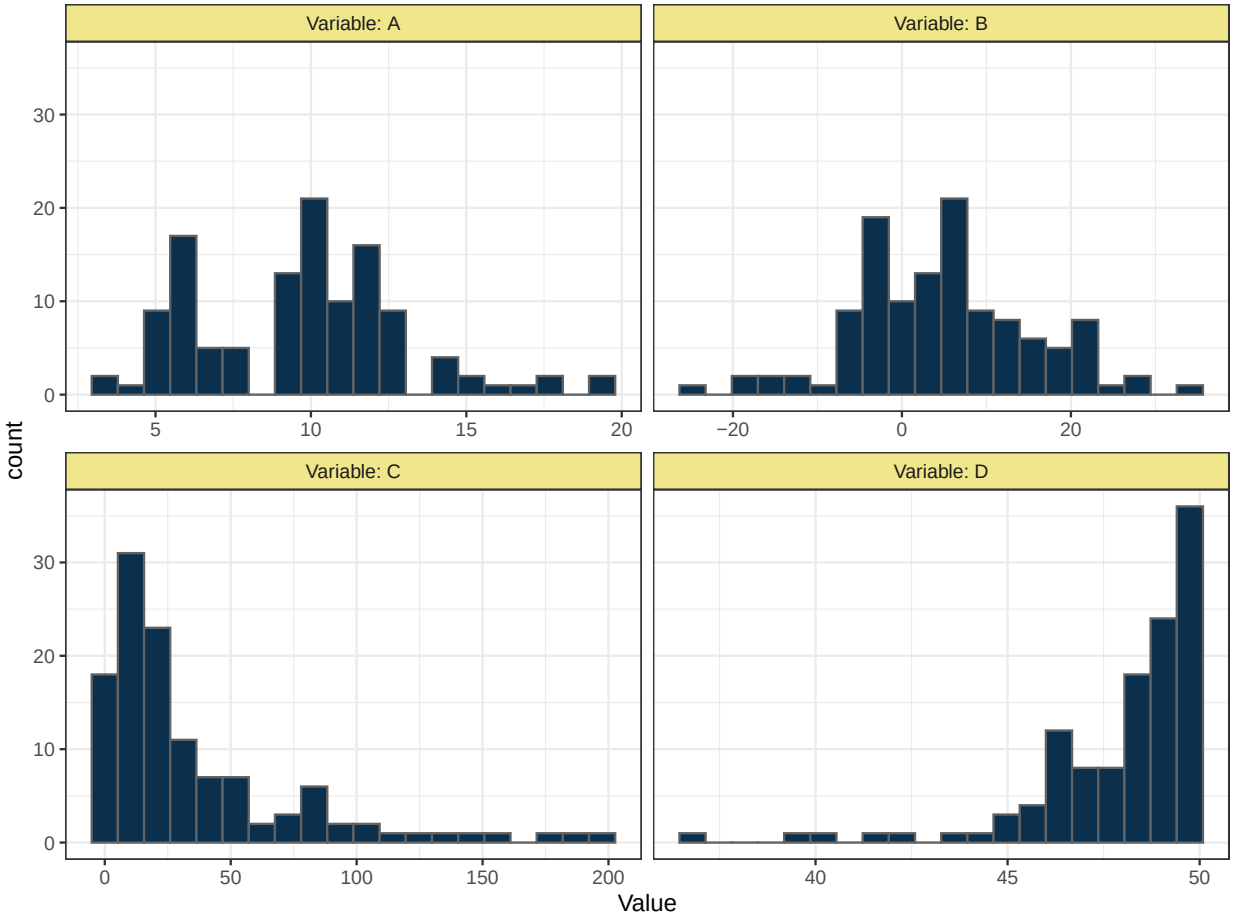
The odds of being alive at the end of the study for treated subjects are estimated to be 4.05 times as high as the odds for control subjects, adjusting for the effects of age, sex, and number of comorbidities.

# 4 Question 4

Which of the four variables plotted in the Display for Question 4 can be most effectively modeled by assuming a Normal distribution of the natural logarithm of the variable?

a. A
b. B
c. C
d. D
e. It is impossible to tell from the information provided.

### Display for Question 4




### 4.1 Answer 4 is C

- The logarithm is an excellent transformation to deal with right skew, in positive values. The histogram of variable C fits those specifications well, but those of the other variables do not.

## 5 Question 5

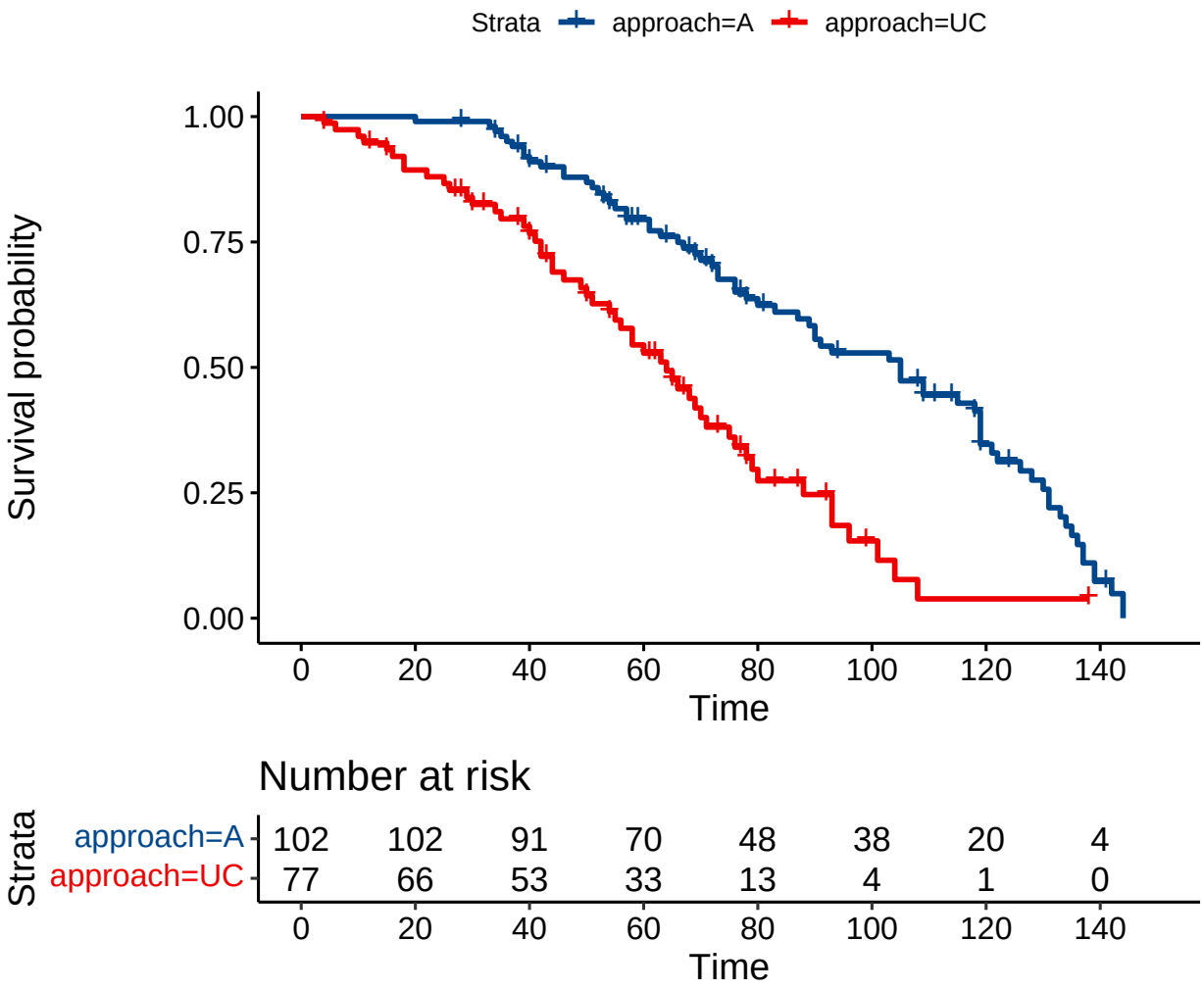
The Display for Question 5 shows Kaplan-Meier estimated survival curves for 179 patients, each of whom either received approach A, or received usual care (UC). Time is measured in days. Censoring is indicated with cross marks, and be sure to CAREFULLY determine which group is represented in each of the lines in the plot. The log rank test (not shown here) has a chi-square estimate of 24.8 with one degree of freedom. Which of the following statements are true?

**CHECK ALL RESPONSES THAT ARE TRUE.**

a. The UC patients had a median survival time about 40 days longer than the Approach A patients.
b. The UC patients had a median survival time about 40 days shorter than the Approach A patients.
c. The log rank test will have a $p$ value which suggests no detectable difference at the 5% significance level between the survival curves of the two approaches.

d. The log rank test will have a $p$ value which suggests a detectable difference at the 5% significance level between the survival curves of the two approaches.
e. None of the statements above are true.

### Display for Question 5




### 5.1 Answer 5 is B, D

Reading over from the point of 50% survival, the median in the (top) blue group is a little over 100 days, while the median in the (bottom) red group is about 60 days, so there is a difference of about 40 days, with the red group having the shorter median survival time. From the legend and the number still at risk at 100 days, it is clear that the red data represents usual care (only four drops in the survival function occur at or after 100 days, as indicated in the number at risk) and that the blue data represent approach A. So statement `b` is correct, and `a` is not.

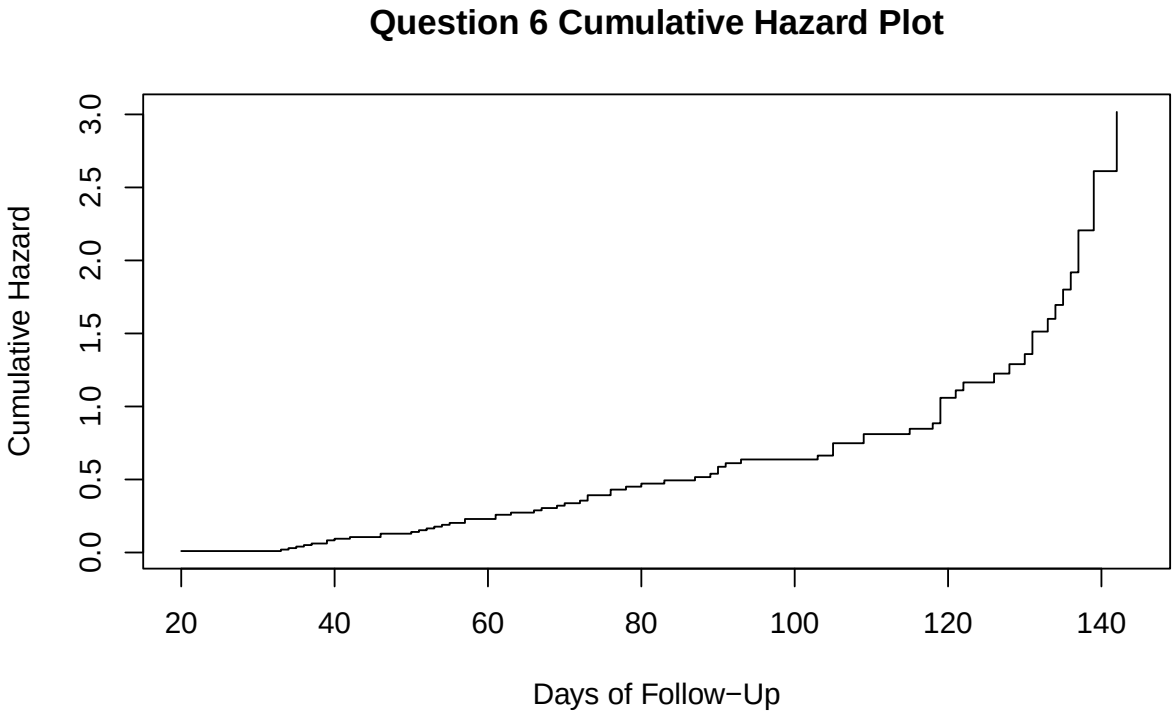
A chi-square value of 24.8 on one degree of freedom yields a $p$ value of `1 - pchisq(24.8,1)` which is a good deal smaller than 0.05, so statement `d` is correct and `c` is not.

# 6 Question 6

A cumulative hazard function (using the inverse Kaplan-Meier methodology) was fit to one of the two approaches (A or UC) discussed in Question 5, to produce the Display for Question 6. In building this plot, I used the same data that produced the Kaplan-Meier curve shown in Question 5. If one of the two approaches (A or UC) produced this cumulative hazard plot, which approach was it?

a. Approach A
b. Usual Care
c. It is impossible to tell from the information provided.

### Display for Question 6




## 6.1 Answer 6 is A

It is in fact the Approach A data. This is easiest to see by looking at the data from day 100 to 140. In this Figure, the cumulative hazard function has lots of small steps up in that time period, which is consistent with the data in Approach A (the blue curve in the Figure for Question 5), which is the subgroup in wich many patients were still at risk. The Usual Care data, since all but four patients were no longer at risk by Day 100, cannot be correct.

# 7 Question 7

In *The Signal and The Noise*, Nate Silver writes repeatedly about a Bayesian way of thinking about uncertainty, for instance in Chapters 8 and 13. Which one of the following statistical methods are consistent with a Bayesian approach to thinking about variation and uncertainty?

**CHECK ALL OF THE CORRECT RESPONSES.**

a. Updating our forecasts as new information appears.
b. Gambling using a strategy derived from a probability model.
c. Combining information from multiple sources to build a model.
d. Significance testing of a null hypothesis, using, say, Fisher's exact test.
e. Establishing a researchable hypothesis prior to data collection.

## 7.1 Answer 7 is A, B, C and E (everything but `d`)

See, for instance, this quote from Silver in the "Bob the Bayesian" section of Chapter 8.

> The problem with Fisher's notion of hypothesis testing is not with having hypotheses but with the way Fisher recommends that we test them.

Each of the other strategies mentioned (besides `d`) is clearly part of the Bayesian approach, and is explicitly described as such in the book.

# Setup for Questions 8 - 10

The `quizd08.Rds` data used in Questions 8-10 has been provided to you. Dr. Love created these data from NHANES 2011-12 Demographics and Questionnaire data, specifically the `DEMO_G` (Demographics), `HSQ_G` (Current Health Status) and `PAQ_G` files, which are described at https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2011.

| Item | Description | Possible Responses |
|---|---|---|
| `SEQN` | Subject id code | 62161 through 71912 |
| `WTINT2YR` | Full sample 2 year interview weight | min = 8045, max = 168807 |
| `RIDAGEYR` | Age in years at screening | min = 21, max = 49 |
| `RIAGENDR` | Sex | 1 = Male, 2 = Female |
| `FEMALE` | Sex | 1 = Female, 0 = Male |
| `RIDRETH3` | Race/Ethnicity | categories listed below |
| `HSD010` | General Health Condition | see below |
| `HSQ571` | Donated blood in past year | see below |
| `PAQ665` | Moderate recreational activities | see below |

- `RIDRETH3` categories are
  - 1 = Mexican American,
  - 2 = Other Hispanic,
  - 3 = Non-Hispanic White,

  - 4 = Non-Hispanic Black,
  - 6 = Non-Hispanic Asian,
  - 7 = Other Race including Multi-Racial
- `HSD010` Would you say your health in general is
  - 1 = Excellent,
  - 2 = Very Good,
  - 3 = Good,
  - 4 = Fair, or
  - 5 = Poor?
  - (Note that 7 = Refused, 9 = Don't know in this variable)
- `HSQ571` During the past 12 months have you donated blood?
  - 1 = Yes,
  - 2 = No,
  - 7 = Refused to answer
  - 9 = Don't Know.
- `PAQ665` Do you do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or golf for at least 10 minutes continuously?
  - 1 = Yes,
  - 2 = No,
  - 7 = Refused to answer
  - 9 = Don't Know.

Here are a few summaries of the data in `quizd08.Rds`.

```
quizd08 <- readRDS("data/quizd08.Rds")

glimpse(quizd08)
```

```
Rows: 2,640
Columns: 9
$ SEQN     <labelled> 62161, 62164, 62169, 62172, 62176, 62180, 62184, 62189, ~
$ WTINT2YR <labelled> 102641.41, 127351.37, 14391.78, 26960.77, 53830.60, 2045~
$ RIDAGEYR <labelled> 22, 44, 21, 43, 34, 35, 26, 30, 35, 42, 36, 28, 35, 38, ~
$ RIAGENDR <fct> 1, 2, 1, 2, 2, 1, 1, 2, 1, 1, 1, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2~
$ RIDRETH3 <fct> 3, 3, 6, 4, 3, 3, 4, 6, 4, 6, 1, 3, 3, 2, 4, 4, 4, 4, 3, 2, 3~
$ HSD010   <fct> 3, NA, 3, 3, 2, 3, 3, 2, NA, 2, 2, 3, NA, 3, 3, 3, 3, 3, 3, 3~
$ HSQ571   <fct> 2, NA, 2, 2, 2, 2, 2, 2, NA, 2, 2, 1, NA, 2, 2, 2, 2, 2, 2, 2~
$ PAQ665   <fct> 2, 1, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1, 2, 1, 2, 2, 1, 1, 2, 1, 1~
$ FEMALE   <labelled> 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0,~
```

```
describe(quizd08)
```

```
quizd08

 9  Variables      2640  Observations
--------------------------------------------------------------------------------
SEQN : Respondent sequence number
       n  missing distinct     Info     Mean      Gmd      .05      .10
    2640        0     2640        1    67061     3295    62631    63115
     .25      .50      .75      .90      .95
   64612    67015    69584    70967    71467

lowest : 62161 62164 62169 62172 62176, highest: 71901 71904 71909 71911 71912
--------------------------------------------------------------------------------
WTINT2YR : Full sample 2 year interview weight
```

```
       n  missing distinct     Info     Mean      Gmd      .05      .10
    2640        0     1464        1    42760    34431    13367    15267
     .25      .50      .75      .90      .95
   19076    29123    48995   105142   123110

lowest :   8044.550   9131.449   9145.762   9177.296   9585.653
highest: 154664.072 154825.467 155208.038 163194.688 168806.615
---------------------------------------------------------------------------
RIDAGEYR : Age in years at screening
       n  missing distinct     Info     Mean      Gmd      .05      .10
    2640        0       29    0.999    34.65    9.656       22       23
     .25      .50      .75      .90      .95
      27       35       42       46       48

lowest : 21 22 23 24 25, highest: 45 46 47 48 49
---------------------------------------------------------------------------
RIAGENDR
       n  missing distinct
    2640        0        2

Value          1     2
Frequency   1304  1336
Proportion 0.494 0.506
---------------------------------------------------------------------------
RIDRETH3
       n  missing distinct
    2640        0        6

lowest : 1 2 3 4 6, highest: 2 3 4 6 7

Value          1     2     3     4     6     7
Frequency    312   240   919   634   440    95
Proportion 0.118 0.091 0.348 0.240 0.167 0.036
---------------------------------------------------------------------------
HSD010
       n  missing distinct
    2249      391        5

lowest : 1 2 3 4 5, highest: 1 2 3 4 5

Value          1     2     3     4     5
Frequency    269   690   927   322    41
Proportion 0.120 0.307 0.412 0.143 0.018
---------------------------------------------------------------------------
HSQ571
       n  missing distinct
    2249      391        3

Value          1     2     9
Frequency    107  2139     3
Proportion 0.048 0.951 0.001
---------------------------------------------------------------------------
PAQ665
       n  missing distinct
```

```
    2640        0        2

Value          1     2
Frequency   1244  1396
Proportion 0.471 0.529
--------------------------------------------------------------------------------
FEMALE : Gender
       n  missing distinct     Info      Sum     Mean      Gmd
    2640        0        2     0.75     1336   0.5061   0.5001

--------------------------------------------------------------------------------
```

# 8 Question 8

Three of the following five commands were used in creating the `quizd08.Rds` data frame that we have made available for you, in some cases when the variables involved were numeric (as `nhanesA` provides them) and not yet specified as factors. Which two of the commands below were not used?

**CHECK EACH of the two commands that were NOT used.**

a. `filter(RIDAGEYR > 20 & RIDAGEYR < 50)`
b. `filter(PAQ665 < 3)`
c. `quizd08 <- zap_label(quizd08)`
d. `mutate(RIDRETH3 = factor(RIDRETH3))`
e. `mutate(FEMALE = 2 - RIAGENDR)`

## 8.1 Answer 8 is that C and E were not used

- `c` was not used, because there are still labels for some of the variables (each of the ones that have not been converted to factors, in fact) in `quizd08`.
- `e` was not used, because such an approach would have yielded 1 for Male and 0 for Female, and that's the opposite of what we have (or what we would want) in the FEMALE variable.
- The other three functions were, in fact, used, in the development of `quizd08`, as you can see from the code below.

```
temp1 <- nhanes('DEMO_G')
temp2 <- nhanes('HSQ_G')
temp3 <- nhanes('PAQ_G')

temp12 <- inner_join(temp1, temp2, by = "SEQN")
temp123 <- inner_join(temp12, temp3, by = "SEQN")

quizd08 <- temp123 %>%
  select(SEQN, WTINT2YR, RIDAGEYR, RIAGENDR, RIDRETH3,
         HSD010, HSQ571, PAQ665) %>%
  filter(RIDAGEYR > 20 & RIDAGEYR < 50) %>%
  filter(PAQ665 < 3) %>%
  mutate(FEMALE = RIAGENDR - 1) %>%
  mutate(RIAGENDR = factor(RIAGENDR),
         RIDRETH3 = factor(RIDRETH3),
         HSD010 = factor(HSD010),
         HSQ571 = factor(HSQ571),
         PAQ665 = factor(PAQ665)) %>%
```

```
    tibble()

saveRDS(quizd08, file = "data/quizd08.Rds")
```

# 9 Question 9

What percentage of the rows included in the `quizd08` data describe subjects who have described their General Health as either "Excellent" or "Very Good"?

Be sure to use a **complete-case** analysis to deal with missing data on the General Health variable.

Please express your response as a percentage between 0 and 100, including a single decimal place.

## 9.1 Answer 9 is 42.6

```
quizd08 <- readRDS("data/quizd08.Rds")
```

```
q09 <- quizd08 %>%
  filter(complete.cases(HSD010))

q09 %>%
  tabyl(HSD010) %>%
  adorn_totals()
```

```
 HSD010    n    percent
      1  269 0.11960871
      2  690 0.30680302
      3  927 0.41218319
      4  322 0.14317474
      5   41 0.01823032
  Total 2249 1.00000000
```

In `HSD010`, category 1 represents "Excellent" and category 2 represents "Very Good". So that's a total of 269 + 690 = 959 subjects who meet that standard, out of a total of 2249 subjects with data on `HSD010`. That's a proportion of 959 / 2249 = 0.42641, or 42.6%, since we need to present this as a percentage with one decimal place.

If you'd neglected to remove the observations with missing values, as instructed, then you would have obtained the following result...

```
quizd08 %>%
  tabyl(HSD010) %>%
  adorn_totals()
```

```
 HSD010    n   percent valid_percent
      1  269 0.1018939    0.11960871
      2  690 0.2613636    0.30680302
      3  927 0.3511364    0.41218319
      4  322 0.1219697    0.14317474
      5   41 0.0155303    0.01823032
   <NA>  391 0.1481061            NA
  Total 2640 1.0000000    1.00000000
```

Following through on that logic, your incorrect response would have been calculated with (269 + 690) / 2640 = 0.36326 or 36.3%.

# 10 Question 10

Next, please answer the question asked in Question 9 again, but this time accounting for the sampling weights used in `WTINT2YR`, again using a **complete-case** analysis to deal with missing General Health values.

What is the resulting estimate of the percentage of the US non-institutionalized adult population within the ages of 21-49 who would describe their General Health as either "Excellent" or "Very Good"?

As you did in Question 9, express your response to Question 10 as a percentage, with a single decimal place.

## 10.1 Answer 10 is 47.2

```
q10_design <-
  svydesign(id = ~SEQN, weights = ~ WTINT2YR, data = q09)

q10_design <- update(q10_design, one = 1)


## get weighted counts for HSD010 levels
svyby( ~ one, ~ HSD010, q10_design, svytotal)
```

```
  HSD010      one        se
1      1 13069131  975968.6
2      2 33700950 1486118.4
3      3 39219547 1440165.3
4      4 11700806  775454.8
5      5  1490001  295977.3
```

```
## total weighted count across all HSD010 responses
svytotal( ~ one, q10_design )
```

```
       total      SE
one 99180436 1695768
```

- Our sample-weighted counts for Excellent (13069131) and Very Good (33700950) sum up to 46770081.
- The total sample-weighted count for all non-missing responses to `HSD010` is 99180436.
- So our sample-weighted proportion is 46770081 / 99180436 = 0.47157, so our sample-weighted percentage (again using one decimal place) is 47.2%.

If you'd neglected to remove the observations with missing values, as instructed, then you would have obtained the following result...

```
q10_design_x <-
  svydesign(id = ~SEQN, weights = ~ WTINT2YR, data = quizd08)

q10_design_x <- update(q10_design_x, one = 1)


## get weighted counts for HSD010 levels
svyby( ~ one, ~ HSD010, q10_design_x, svytotal)
```

```
  HSD010      one        se
1      1 13069131  981684.4
2      2 33700950 1511034.4
3      3 39219547 1474880.6
4      4 11700806  781223.4
5      5  1490001  296214.5
```

```
## total weighted count across all HSD010 responses
svytotal( ~ one, q10_design_x )
```

```
        total      SE
one 112886988 1784406
```

- So that incorrect response would have been (13069131 + 33700950) / 112886988 = 0.414, or 41.4%.

# 11 Question 11

Suppose you are trying to build a regression model to predict a patient's self-reported overall health (where the available responses are Excellent, Very Good, Good, Fair or Poor) where you want to treat the health assessments as categorical. Which of the following models would represent the most appropriate choice?

a. An ordinary least squares model.
b. A Cox proportional hazards model.
c. A proportional odds logistic regression model.
d. A zero-inflated negative binomial model.
e. A binary logistic regression model.
f. None of these models would be appropriate.

## 11.1 Answer 11 is C

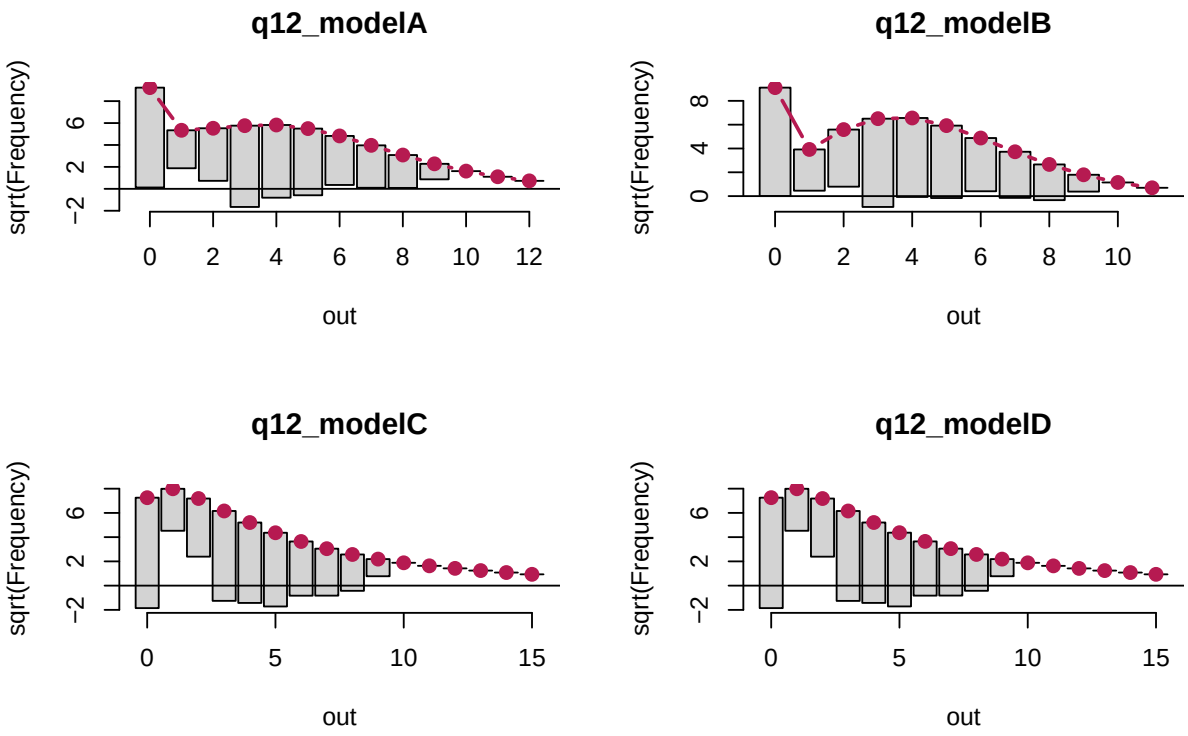
- A proportional odds logistic regression model is used to describe ordered multi-categorical outcomes. The others are not.

# 12 Question 12

The Display for Question 12 shows four rootograms, using four different count regression models to fit the same outcome, which is named `out`. Which model (A, B, C, D) shows the best fit to the data?

a. Model A
b. Model B
c. Model C
d. Model D
e. It is impossible to tell from the information provided.

### 12.1 Display for Question 12




### 12.2 Answer 12 is B

- Model B clearly shows the best fit to the data, of the four models provided. The modeled counts are very close to the actual counts, across the range.

## 13 Question 13

Patients with genital herpes (HSV-2) were enrolled in a study of a new recombinant herpes vaccine based on the antigen glycoprotein, gD2. Patients were required to have a history of at least 6 HSV-2 episodes in the 12 months prior to study enrollment and be in remission at the time of vaccination. Patients were randomly assigned to receive either a single gD2 vaccine injection or a placebo, and their conditions were followed for 1 year. The response is the time (in weeks) to first recurrence of HSV-2 following immunization for each patient. Our aim is to determine whether the distributions of the times to recurrence differ between the two vaccination groups. A few subjects dropped out of the study before the year was up and before their HSV-2 recurred.

Which of the possible modeling strategies listed below might be appropriate for the study. More than one strategy might be appropriate, and you should indicate all appropriate choices.

**CHECK ALL APPROPRIATE CHOICES**

a. Linear regression to predict the outcome.
b. (Binary) Logistic regression to predict the outcome.
c. Poisson regression to predict the outcome.

d. Multinomial logistic regression to predict the outcome.
e. Ordinal logistic regression to predict the outcome.
f. Cox proportional hazards model to predict the outcome.
g. Log rank test of the intervention's impact on the outcome.
h. None of these modeling strategies are appropriate.

### 13.1 Answer 13 is F and G

We are interested in a time to recurrence outcome with right censoring, and we're comparing two groups. Either a Cox model or simply a log rank test could work.

## Setup for Questions 14 - 16

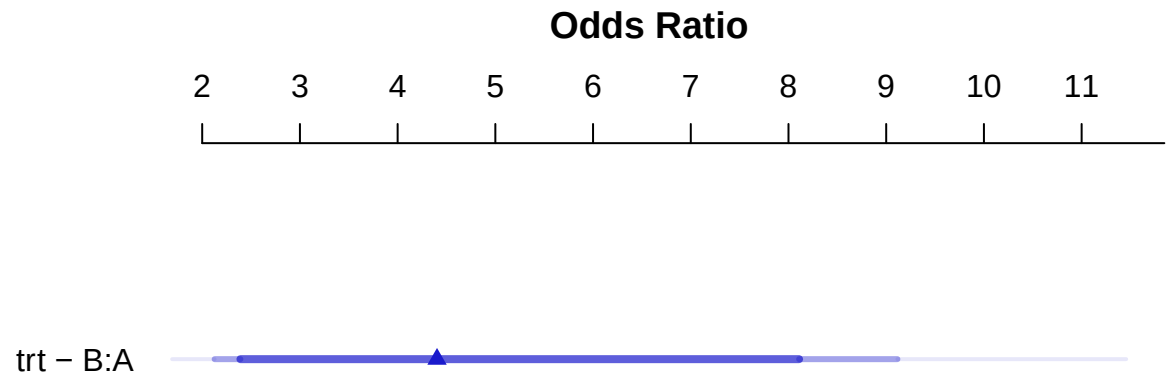
141 subjects were enrolled in a study comparing two treatments (`trt` = A or B). The outcome of interest is a binary result (1 = good, 0 = bad), and we also have information on baseline severity of illness (on a 0-10 scale, where 10 is most severe) and on insurance status (`ins` = Private, Public or None). Consider the output and plot below.

```
dat14 <- read_csv("data/dat14.csv")

d <- datadist(dat14)
options(datadist="d")
model14 <- lrm(result ~ trt, data=dat14, x=TRUE, y=TRUE)
model14
```

```
Logistic Regression Model

 lrm(formula = result ~ trt, data = dat14, x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                             Ratio Test           Indexes          Indexes
 Obs           141    LR chi2      17.27    R2       0.155    C       0.675
  0             58    d.f.             1    g        0.746    Dxy     0.351
  1             83    Pr(> chi2) <0.0001    gr       2.108    gamma   0.630
 max |deriv| 5e-12                          gp       0.171    tau-a   0.171
                                            Brier    0.213

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -0.3037 0.2368 -1.28  0.1997
 trt=B      1.4823 0.3712  3.99  <0.0001
```

### Display for Question 14

```
plot(summary(model14))
```




# 14 Question 14

Which of the following statements are true, according to the Model provided in the Setup for Questions 14 and 15, and the Display for Question 14?

**CHECK ALL OF THE TRUE STATEMENTS.**

a. The odds of a good result are estimated to be 1.48 times as large with treatment B as with treatment A.
b. The odds of a good result are more than 4 times as large with treatment B as with treatment A.
c. The probability of a good result is nearly 5 times as high with treatment B as with treatment A.
d. More people had a good result than a bad result in this study.
e. None of the above statements are true.

## 14.1 Answer 14 is B and D

- Statement `a` is not true. The 1.48 refers to log odds, not odds.
- Statement `b` is true. From the plot, the odds ratio associated with trt B appears to be between 4 and 5, which is certainly more than 4.
- Statement `c` is not true. The ratio we see in the plot describes the odds ratio not the probability ratio.
- Statement `d` is true. We had 83 `1`s and 58 `0`s in our results, so that's more good than bad.

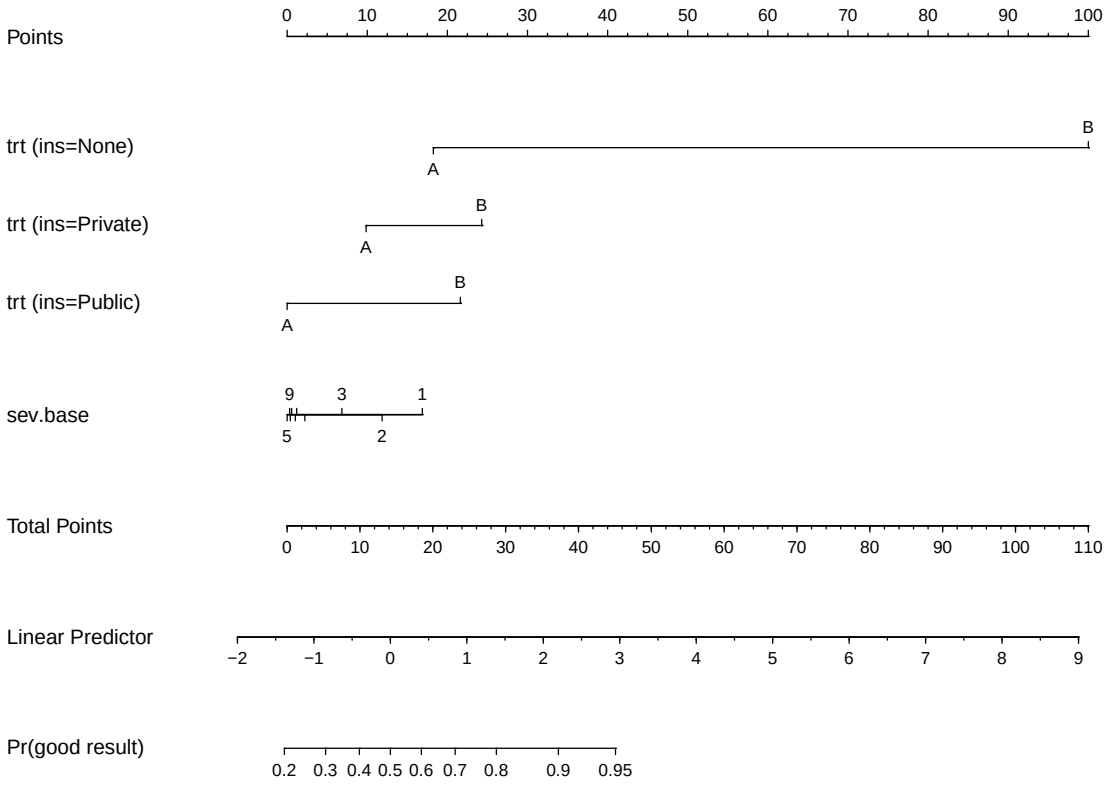
# 15 Question 15

An augmented model was fit to the data discussed in Question 14, incorporating some non-linear terms in the predictors we've seen, and a nomogram of that augmented model in the Display for Question 15. If you want to see a larger version of the nomogram, it is available on our Quiz 2 website, in the file `q15-nomogram.png`

Based on the nomogram provided in the Display for Question 15 and 16, which of the following responses includes the predicted probability of a good result for a subject with Private insurance, treatment A, and baseline severity 1?

a. Less than 0.25
b. Between 0.25 and 0.49
c. Between 0.5 and 0.74
d. 0.75 or larger
e. It is impossible to tell from the information provided.

## 15.1 Display for Question 15 and 16



### 15.2 Answer 15 is D

- Private insurance with treatment A yields approximately 10 points, and severity 1 yields a little less than 20 more, for a total near 30.
- Total Points = 30 is a linear predictor of about 1.5, which gives a predicted probability for a good result between 0.8 and 0.9.

## 16 Question 16

Based on the nomogram shown in the Display for Question 15 and 16, which of the following augmentations to the model shown in Question 14 were made?

**CHECK ALL OF THE RESPONSES THAT APPLY.**

a. A restricted cubic spline or polynomial function in `sev.base`
b. An interaction term between `sev.base` and `treatment`
c. An interaction term between `sev.base` and `insured`
d. An interaction term between `treatment` and `insured`
e. None of the above

### 16.1 Answer 16 is A and D

```
model15 <- lrm(result ~ trt + rcs(sev.base,5) +
                   trt * ins, data=dat14, x=TRUE, y=TRUE)
```

The actual model fit is shown above, so that you can verify that I fit a spline in severity and the interaction of treatment with insurance status.

## 17 Question 17

A subset of the data from the BRFSS SMART study developed in Chapter 2 of the Course Notes are used in Questions 17 and 18 with some modifications. The three variables of interest are:

- `mmsa`, which is either CIN (Cincinnati), CLE (Cleveland-Elyria), COL (Columbus), or DAY (Dayton)
- `vax_pneumo`, which is either "Vax" if the subject had received a vaccination against pneumonia, and "NoVax" if not
- `binge`, which is either "Yes" or "No" (the standard for "Yes" is sex-specific: males having five or more drinks on one occasion in the past 30 days, females having four or more drinks on one occasion in the past 30 days)

The `quizd17.Rds` data set of interest is available to you on our web site, if you want it. I used it to fit the following four models to predict `mmsa` based on the two predictors.

```
quizd17 <- readRDS("data/quizd17.Rds")
options(contrasts = c("contr.treatment", "contr.poly"))

m17_1 <- multinom(mmsa ~ 1,
                  data = quizd17, trace = FALSE)
m17_B <- multinom(mmsa ~ binge,
                  data = quizd17, trace = FALSE)
m17_V <- multinom(mmsa ~ vax_pneumo,
                  data = quizd17, trace = FALSE)
m17_BV <- multinom(mmsa ~ binge + vax_pneumo,
```

```
                  data = quizd17, trace = FALSE)
m17_SAT <- multinom(mmsa ~ binge * vax_pneumo,
                  data = quizd17, trace = FALSE)
```

Here are the AIC results for these models:

```
AIC(m17_1, m17_B, m17_V, m17_BV, m17_SAT)
```

```
        df      AIC
m17_1    3 11944.56
m17_B    6 11937.27
m17_V    6 11940.38
m17_BV   9 11934.83
m17_SAT 12 11936.76
```

Note that one of the models I fit is preferable to the others, on the basis of the Akaike Information Criterion. Call that the preferred model. Which of the models is the preferred model?

a. The model with an intercept only.
b. The model using only the main effect of `binge`.
c. The model using only the main effect of `vax_pneumo`.
d. The model using only the main effects of `binge` and `vax_pneumo`.
e. The model including the interaction of `binge` and `vax_pneumo`.

### 17.1 Answer 17 is D

The preferred model (lowest AIC) is the BV model, so that model will include both the main effects of `binge` and `vax_pneumo` but not their interaction.

## 18 Question 18

Your job is to consider the five displays for Question 18 that follow, and determine which of those displays describe the preferred model that you identified back in Question 17. More than one response may be correct.

**CHECK ALL RESPONSES THAT APPLY.**

a. Display A for Question 18
b. Display B for Question 18
c. Display C for Question 18
d. Display D for Question 18
e. Display E for Question 18
f. None of these.

### Display A for Question 18

```
tidy(displayA, exponentiate = TRUE, conf.int = TRUE) %>%
  kable(digits = 3)
```

| y.level | term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|---|
| CLE | (Intercept) | 0.601 | 0.063 | -8.067 | 0.000 | 0.531 | 0.680 |
| CLE | bingeYes | 0.938 | 0.116 | -0.557 | 0.578 | 0.748 | 1.176 |
| CLE | vax_pneumoVax | 1.255 | 0.084 | 2.695 | 0.007 | 1.064 | 1.481 |

| y.level | term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|---|
| COL | (Intercept) | 1.149 | 0.053 | 2.616 | 0.009 | 1.036 | 1.275 |
| COL | bingeYes | 0.776 | 0.102 | -2.489 | 0.013 | 0.636 | 0.948 |
| COL | vax_pneumoVax | 1.125 | 0.073 | 1.627 | 0.104 | 0.976 | 1.297 |
| DAY | (Intercept) | 0.333 | 0.078 | -14.087 | 0.000 | 0.285 | 0.388 |
| DAY | bingeYes | 0.645 | 0.161 | -2.732 | 0.006 | 0.470 | 0.883 |
| DAY | vax_pneumoVax | 1.219 | 0.105 | 1.884 | 0.060 | 0.992 | 1.497 |

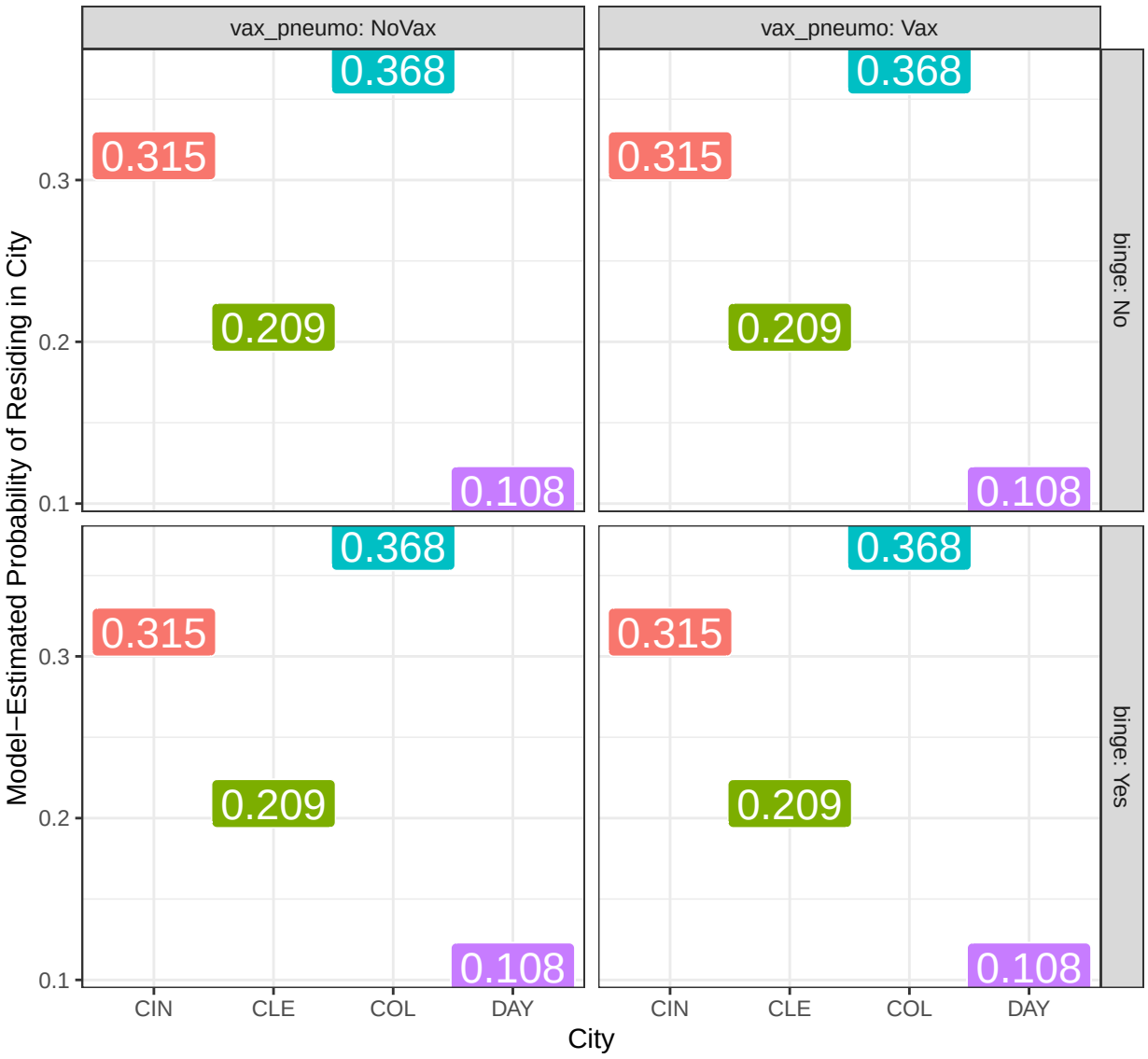
### Display B for Question 18

```
tidy(displayB, exponentiate = TRUE, conf.int = TRUE) %>%
  kable(digits = 3)
```

| y.level | term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|---|
| CLE | (Intercept) | 0.601 | 0.066 | -7.743 | 0.000 | 0.529 | 0.684 |
| CLE | bingeYes | 0.934 | 0.143 | -0.479 | 0.632 | 0.705 | 1.237 |
| CLE | vax_pneumoVax | 1.252 | 0.091 | 2.472 | 0.013 | 1.048 | 1.497 |
| CLE | bingeYes:vax_pneumoVax | 1.016 | 0.243 | 0.066 | 0.948 | 0.632 | 1.635 |
| COL | (Intercept) | 1.144 | 0.055 | 2.445 | 0.014 | 1.027 | 1.275 |
| COL | bingeYes | 0.797 | 0.124 | -1.832 | 0.067 | 0.626 | 1.016 |
| COL | vax_pneumoVax | 1.135 | 0.078 | 1.622 | 0.105 | 0.974 | 1.321 |
| COL | bingeYes:vax_pneumoVax | 0.925 | 0.218 | -0.359 | 0.720 | 0.603 | 1.418 |
| DAY | (Intercept) | 0.344 | 0.080 | -13.419 | 0.000 | 0.294 | 0.402 |
| DAY | bingeYes | 0.511 | 0.213 | -3.151 | 0.002 | 0.336 | 0.776 |
| DAY | vax_pneumoVax | 1.145 | 0.111 | 1.220 | 0.222 | 0.921 | 1.424 |
| DAY | bingeYes:vax_pneumoVax | 1.777 | 0.325 | 1.768 | 0.077 | 0.939 | 3.363 |

### Display C for Question 18

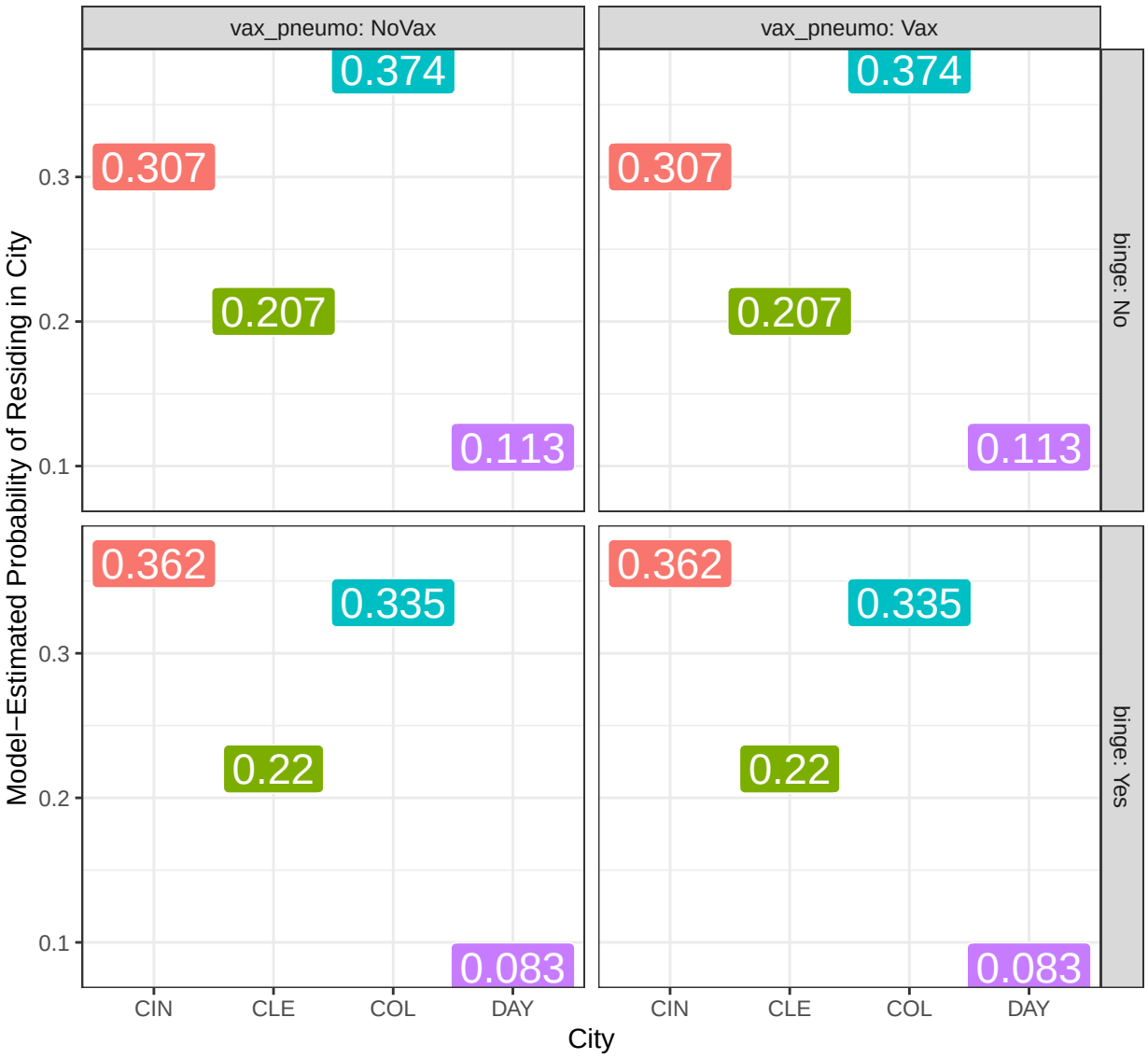
```
ggplot(displayC, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 6) +
  guides(fill = "none") +
  facet_grid(binge ~ vax_pneumo, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Residing in City",
       x = "City")
```




Note that the `prob` variable shows the fitted probability of the subject residing in each city, according to the model fit for this display.

### Display D for Question 18

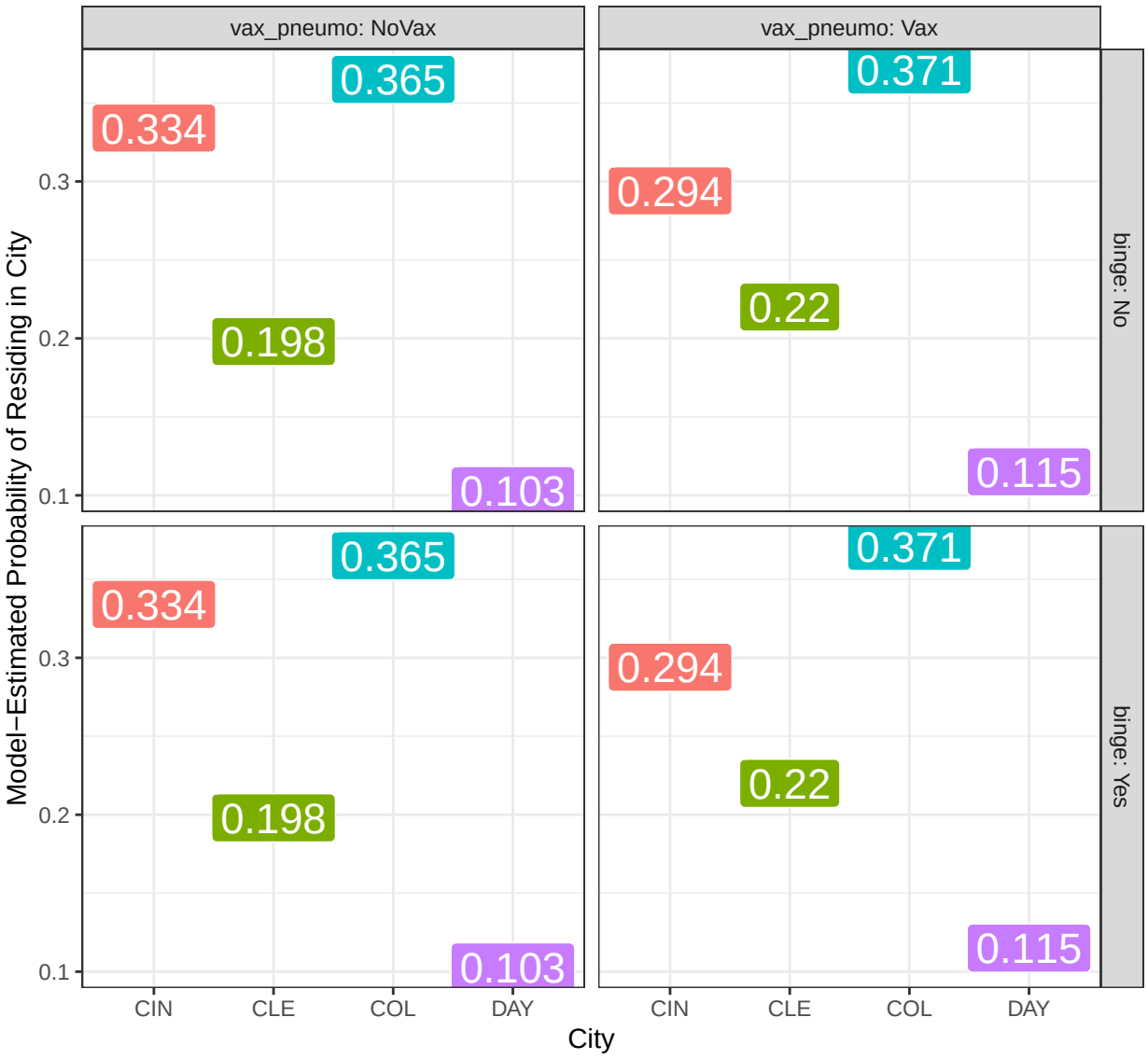
```
ggplot(displayD, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 6) +
  guides(fill = "none") +
  facet_grid(binge ~ vax_pneumo, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Residing in City",
       x = "City")
```




Note that the `prob` variable shows the fitted probability of the subject residing in each city, according to the model fit for this display.

### Display E for Question 18

```
ggplot(displayE, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 6) +
  guides(fill = "none") +
  facet_grid(binge ~ vax_pneumo, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Residing in City",
       x = "City")
```




Note that the `prob` variable shows the fitted probability of the subject residing in each city, according to the model fit for this display.

## 18.1 Answer 18 is A (and only a.)

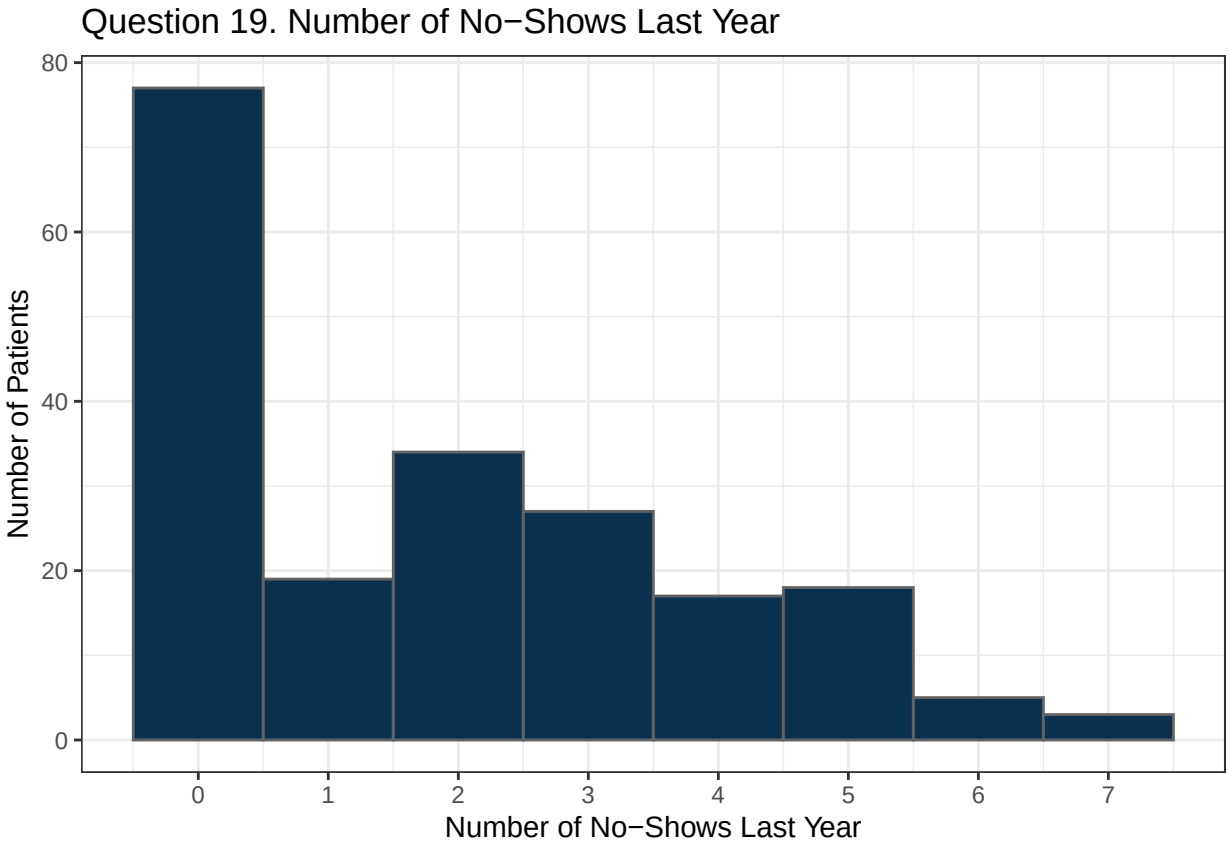
- Display A shows coefficients for both `binge` and `vax_pneumo` but not their interaction, so that is the preferred model.
- Display B shows coefficients for `binge` and `vax_pneumo` but also for their interaction, so that is not the right model.
- Display C shows no changes in the fitted probabilities associated with either `binge` or `vax_pneumo` so that cannot be the preferred model, which uses both main effects.
- Display D shows the same fitted probabilities regardless of `vax_pneumo` status, depending only on `binge`, so that cannot be the preferred model.
- Display E shows the same fitted probabilities regardless of `binge` status, depending only on `vax_pneumo` so that cannot be the preferred model.

# 19 Question 19

Suppose you are trying to build a regression model to predict `noshow`, the number of times a patient will "no show" an appointment for medical care in the next 12 months, on the basis of several characteristics related to their health, demographics, and satisfaction levels with prior visits. The `noshow` data on 200 patients from last year are visualized in the Display for Question 19. Which of the following models is most likely to be appropriate?

a. A Cox proportional-hazards model.
b. A proportional odds logistic regression.
c. A binary logistic regression model.
d. A zero-inflated Poisson model.
e. A multinomial logistic regression model.
f. None of these models will be appropriate.

### Display for Question 19




## 19.1 Answer 19 is D

- From the Display, this is a count outcome, with (it appears) some extra zeros. That looks like a zero-inflated Poisson regression would be the best choice.

# 20 Question 20

Which of the following strategies is described as a sound (or more fox-like) strategy for making predictions according to Nate Silver in *The Signal and the Noise*?

**CHECK ALL RESPONSES THAT APPLY.**

a. Betting on the forecast for next year made by the individual who had the best result last year in a prediction contest.
b. Focusing on the principal cause of an outcome, more than correlating factors.
c. Extrapolating (on an exponential scale) from a current trend.
d. Calculating the basic reproduction number $R_0$ at the start of a disease's sweep through a community.
e. Applying an SIR model (susceptible - infectious - recovery) to the prediction of infectious disease in a similar way for all members of a population.

## 20.1 Answer 20 is B (and only `b`.)

See Chapters 5 and 7 of *The Signal and the Noise*.

# 21 Question 21

Suppose you are trying to build a regression model to predict whether or not a patient hospitalized with heart failure will need to return to the hospital in the 30 days after they are released. You gather a series of predictors that should be useful.

Which of the following models would be most appropriate?

**SELECT THE SINGLE BEST RESPONSE.**

a. A multinomial logit model.
b. A binary logistic regression model.
c. An ordinary least squares model.
d. A Cox proportional hazards model.
e. None of these models would be appropriate.

## 21.1 Answer 21 is B

- A binary outcome (rehospitalized or not rehospitalized) is what we have, so a plain old binary logistic regression is the best choice of model from these options.

# 22 Question 22

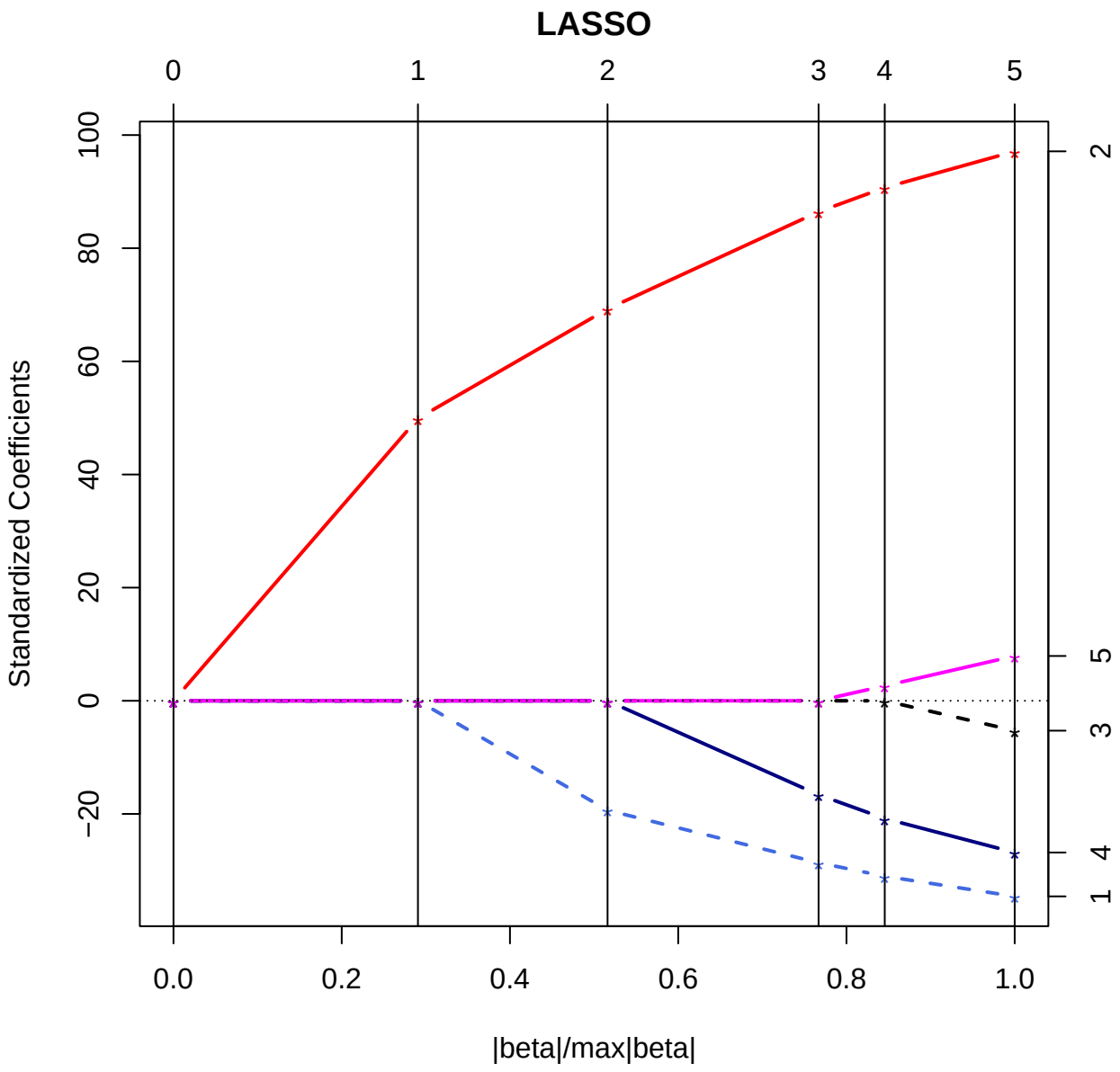
Consider the lasso plot shown in the Display for Question 22 describing a situation with five candidate predictors. If the |beta| / max|beta| value at which the cross-validated mean square error was minimized in this situation was 0.6, then specify the predictors that the plot suggests should be included in the model.

**CHECK ALL PREDICTORS THAT SHOULD BE INCLUDED.**

a. Variable 1
b. Variable 2
c. Variable 3
d. Variable 4

e. Variable 5
f. None of the above

### Display for Question 22




### 22.1 Answer 22 is A, B and E

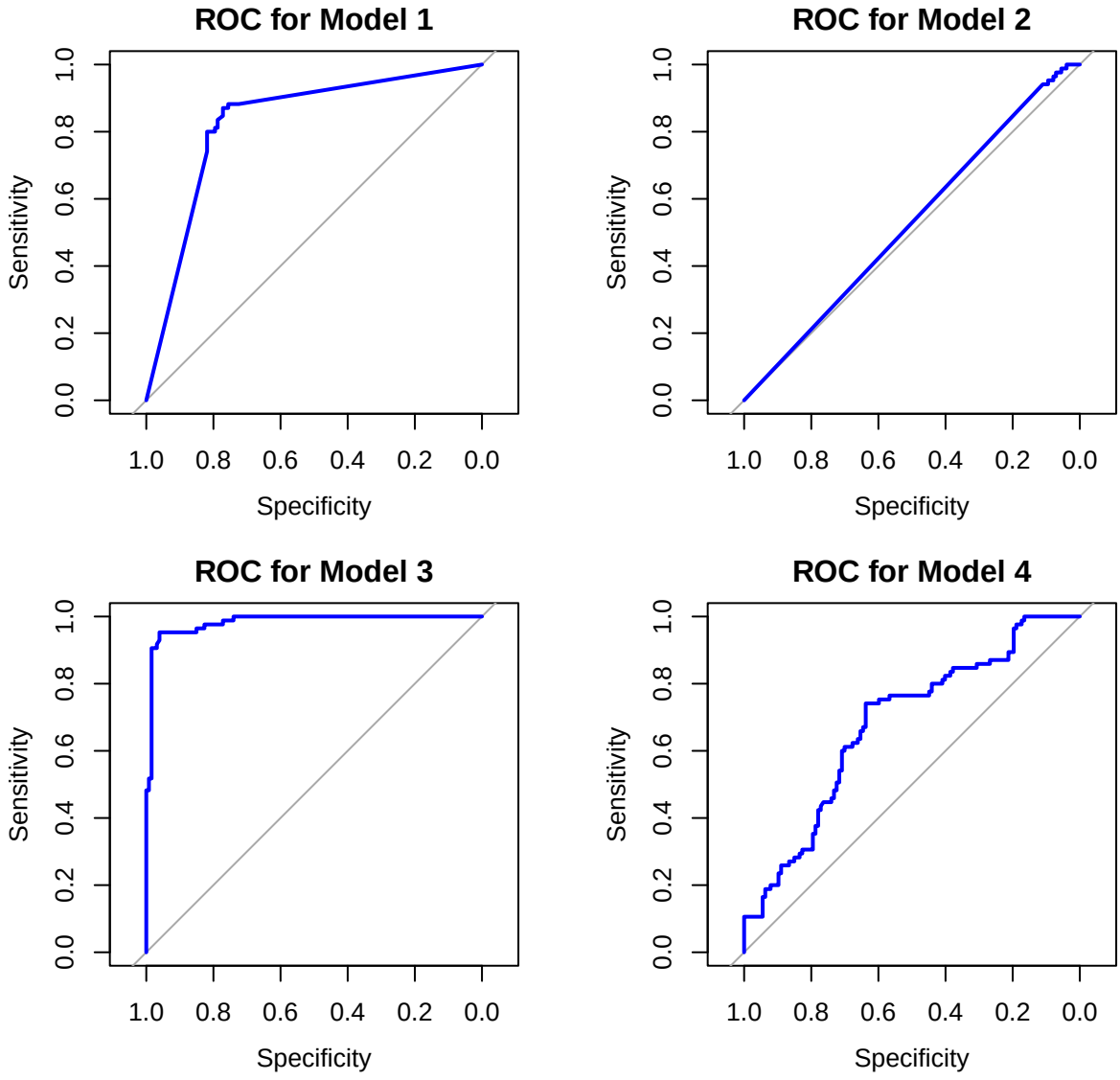
The Lasso shows that only variables 3 and 4 have been deleted from the model at a fraction of 0.6, so variables 1, 2 and 5 remain.

# 23 Question 23

The grid of four receiver operating characteristic curve plots shown in the Display for Question 23 describes four different logistic regression models fit to predict the same outcome. Indicate the C statistic associated with each model.

- Rows are Model 1, 2, 3, 4
- Columns are C statistics equal to 0.53, 0.68, 0.83, 0.98

## Display for Question 23



### 23.1 Answer 23 is that 1 goes with 0.83, 2 with 0.53, 3 with 0.98 and 4 with 0.68

To prove that these answers are correct, the actual ROC statistics are shown below. The C statistic measures the area under the plotted blue curves.

```
roc(dat23$outcome ~ dat23$modelB) # listed as Model 1
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```
Call:
roc.formula(formula = dat23$outcome ~ dat23$modelB)

Data: dat23$modelB in 127 controls (dat23$outcome 0) < 85 cases (dat23$outcome 1).
Area under the curve: 0.8289
```

```
roc(dat23$outcome ~ dat23$modelC) # listed as Model 2
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```
Call:
roc.formula(formula = dat23$outcome ~ dat23$modelC)

Data: dat23$modelC in 127 controls (dat23$outcome 0) < 85 cases (dat23$outcome 1).
Area under the curve: 0.5264
```

```
roc(dat23$outcome ~ dat23$modelD) # listed as Model 3
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```
Call:
roc.formula(formula = dat23$outcome ~ dat23$modelD)

Data: dat23$modelD in 127 controls (dat23$outcome 0) < 85 cases (dat23$outcome 1).
Area under the curve: 0.9824
```

```
roc(dat23$outcome ~ dat23$modelA) # listed as Model 4
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```
Call:
roc.formula(formula = dat23$outcome ~ dat23$modelA)

Data: dat23$modelA in 127 controls (dat23$outcome 0) < 85 cases (dat23$outcome 1).
Area under the curve: 0.6792
```

## Setup for Questions 24 - 26

The `quizd24.csv` data set (which will be used in Questions 24-26) is available to you on the course web site.

The outcome of interest in that data set, labeled `y`, is the number of standards (out of 6) met by subjects involved in an alcoholism treatment program. Subjects are released from the program when they meet all six standards. The data in `y` describe the number of standards met after one week of treatment for 200 recent

subjects. Measures `x1`, `x2` and `x3` are predictors of `y`, whose main effects (only) are of interest to us. `x1` and `x3` are quantitative measures, and `x2` indicates whether or not the subject has completed a specific group of tasks.

# 24 Question 24

Fit a Poisson regression model to the data in the `quizd24.csv` file, and compare your result to what you obtain using a negative binomial regression. Treat variable `x2` as a number (1/0) rather than converting it to a factor.

Which of the statements listed in the Display for Question 24 are true?

**CHECK EACH OF THE CORRECT RESPONSES.**

a. Statement I
b. Statement II
c. Statement III
d. None of these three statements.

## Display for Question 24

- Statement I. A main effects model fit with Poisson regression provides a statistically significantly worse fit (at the 95% confidence level) than a model fit with Negative Binomial regression.
- Statement II. The rootogram for the Poisson model indicates a substantially better fit than the rootogram for the Negative Binomial model.
- Statement III. The rootogram for the Poisson model indicates a substantially worse fit than the rootogram for the Negative Binomial model.

## 24.1 Answer 24 is None of the above

- None of the statements are true. The Poisson and Negative Binomial regression models are nearly identical, and show no detectable difference (the difference in the log likelihood functions is essentially zero) in the likelihood ratio test, and there are no meaningful differences between the rootograms.

```
quiz24 <- read_csv(here("data/quizd24.csv"),
                   show_col_types = FALSE)
mod24_p <- glm(y ~ x1 + x2 + x3, family = poisson(),
               data = quiz24)
mod24_nb <- MASS::glm.nb(y ~ x1 + x2 + x3,
                         link = log, data = quiz24)

logLik(mod24_p)
```

```
'log Lik.' -296.0254 (df=4)
```

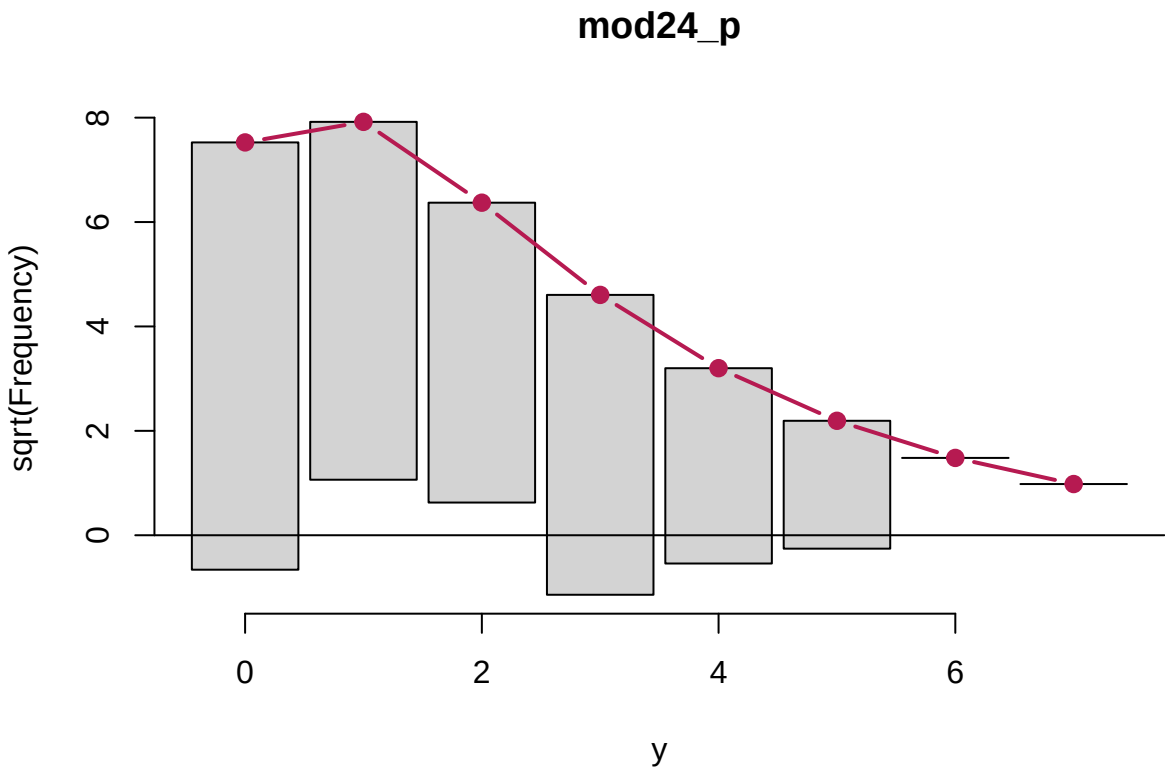
```
logLik(mod24_nb)
```

```
'log Lik.' -296.0262 (df=5)
```

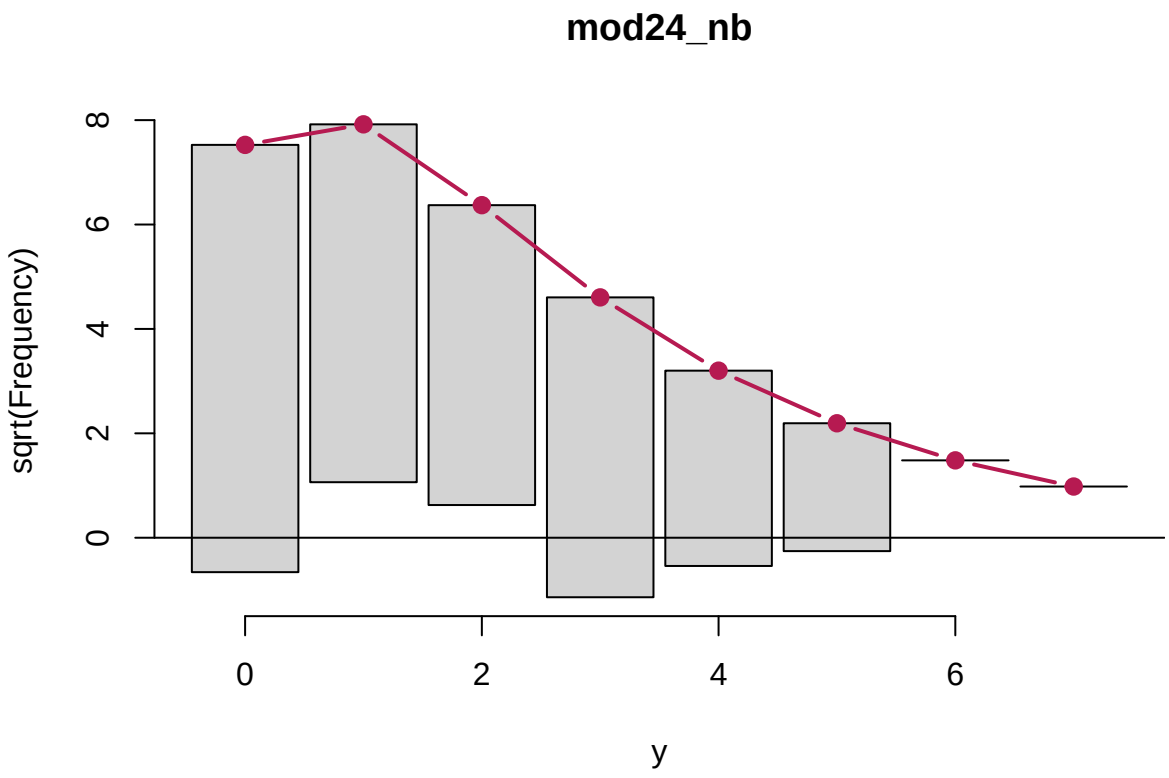
```
pchisq(2 * (logLik(mod24_nb) - logLik(mod24_p)),
       df = 1, lower.tail = FALSE)
```

```
'log Lik.' 1 (df=5)
```

```
rootogram(mod24_p)
```




```
rootogram(mod24_nb)
```




## 25 Question 25

Consider three new subjects, who are named Abigail, Brad and Chen. Here are their values of the predictors.

| Name | x1 | x2 | x3 |
|---|---|---|---|
| Abigail | 4 | 1 | 0 |
| Brad | 3 | 0 | 4 |
| Chen | 2 | 1 | 6 |

Use the Poisson regression model you fit in Question 24 to make a prediction for y for the three new subjects listed here. Rank the three new subjects in order of their predicted y, from highest (first) to lowest.

a. Abigail has the highest predicted y, then Brad then Chen
b. Abigail is highest, then Chen then Brad
c. Brad is highest, then Abigail then Chen
d. Brad is highest, then Chen then Abigail
e. Chen is highest, then Abigail then Brad
f. Chen is highest, then Brad then Abigail

### 25.1 Answer 25 is E

Let's make predictions from the Poisson model.

```
nd25 = tibble(name = c("Abigail", "Brad", "Chen"),
                      x1 = c(4, 1, 0),
                      x2 = c(3, 0, 4),
                      x3 = c(2, 1, 6))

predict(mod24_p, newdata = nd25, type = "response")
```

```
         1          2          3
11.0554323  0.4824522 27.5377361
```

- Chen is highest, then Abigail, then Brad.

# 26 Question 26

Now, instead of treating `y` in `quizd24.csv` as a count variable, treat it as an ordinal category, and fit a new model that is appropriate for such an outcome using again the main effects of x1, x2 and x3 as predictors. Use that model to predict the actual category that our three new subjects (Abigail, Brad and Chen) will fall into, and compare that to the results you found in Question 25.

Which of the three new subjects get a different predicted count with this ordinal categorical regression model, than they do when you round the predicted count made with the Poisson model to an integer?

**CHECK EACH OF THE CORRECT RESPONSES.**

a. Abigail
b. Brad
c. Chen
d. None of the three subjects.
e. It is impossible to tell from the information provided.

## 26.1 Answer 26 is A and C.

Let's get the predictions from the proportional odds logistic regression model:

```
mod26_polr <-
    polr(factor(y) ~ x1 + x2 + x3,
         data = quiz24, Hess = TRUE)

predict(mod26_polr, nd25)
```

```
[1] 5 0 5
Levels: 0 1 2 3 4 5
```

- For Abigail, the polr model predicts 5, and the Poisson predicts 11.06, which rounds to 11.
- For Brad, the polr model predicts 0, and the Poisson predicts 0.48, which rounds to 0.
- For Chen, the polr model predicts 5, and the Poisson predicts 27.5, which rounds to 28.
- So one of the predictions is the same; the one for Brad.

# 27 Question 27

We have available the `startday` and `exitday` for each subject in a tobacco cessation study, comparing three `treatment`s (called A, B and usual care.) The study began collecting subjects on day 0. The `exitreason` variable shows the reason why each subject exited the study, either because they achieved the outcome

(`achieved`), they stopped coming to appointments and were thus lost to follow up (`lost`), or because the study ended (`studyend`).

A summary of the relevant data from the study follows.

```
data27
```

```
# A tibble: 140 x 4
   startday exitday exitreason treatment
      <dbl>   <dbl> <fct>      <fct>
 1        0    31.5 achieved   UC
 2        0    38.2 achieved   B
 3        0    24.8 lost       UC
 4        0    47.3 lost       UC
 5        0    38.1 lost       UC
 6        0    19.6 achieved   B
 7        0    28.1 achieved   UC
 8        0    14.3 achieved   B
 9        0    39.3 lost       B
10        0    23.8 achieved   UC
# ... with 130 more rows
```

```
describe(data27)
```

```
data27

 4  Variables      140  Observations
--------------------------------------------------------------------------------
startday
       n  missing distinct     Info     Mean      Gmd      .05      .10
     140        0       23    0.975    20.14    14.56     0.00     0.00
     .25      .50      .75      .90      .95
    0.00    25.50    30.25    34.00    35.10

lowest :  0 16 18 20 21, highest: 35 37 38 39 40
--------------------------------------------------------------------------------
exitday
       n  missing distinct     Info     Mean      Gmd      .05      .10
     140        0      140        1    57.08    21.57    23.81    31.38
     .25      .50      .75      .90      .95
   42.71    58.75    72.30    80.31    85.29

lowest : 14.29152 18.29017 19.62402 21.25336 22.37179
highest: 87.49991 87.53402 87.93659 91.80623 94.05758
--------------------------------------------------------------------------------
exitreason
       n  missing distinct
     140        0        3

Value      achieved     lost studyend
Frequency        43       31       66
Proportion    0.307    0.221    0.471
--------------------------------------------------------------------------------
treatment
       n  missing distinct
     140        0        3
```

```
Value          A    UC     B
Frequency     32    72    36
Proportion 0.229 0.514 0.257
-------------------------------------------------------------------------------
```

Suppose you want to add a survival object called `S` to the `data27` data, and want to treat the subjects who did not achieve the outcome as being right-censored, then fit a log rank test to compare the three `treatment` groups in terms of that survival object. Which of the chunks of R code shown in the Display for Question 27 will accomplish this?

**CHECK ALL THE ANSWERS THAT ARE CORRECT.**

a. Chunk I
b. Chunk II
c. Chunk III
d. None of these Chunks

### Display for Question 27

**Chunk I**

```
data27$S = Surv(time = data27$exitday - data27$startday,
                event = data27$exitreason %in% c("lost", "studyend"))
survdiff(S ~ treatment, data = data27)
```

**Chunk II**

```
survdiff(Surv(time = data27$exitday,
              event = data27$exitreason) ~ treatment)
```

**Chunk III**

```
data27$S = Surv(time = data27$exitday - data27$startday,
                event = data27$exitreason == "achieved")
survdiff(S ~ treatment, data = data27)
```

### 27.1 Answer 27 is C

- Chunk I creates the wrong survival object, flipping the designation of censored and observed times.
- Chunk II doesn't create a survival object, so that won't work.
- Chunk III gets everything right.

## 28 Question 28

The experimental anti-hypertensive agent, GB2995, was studied in 4 study centers to compare 12 weeks of treatment using GB2995 as the sole therapy vs. using GB2995 in combination with a stable dose of a marketed calcium-channel blocker when given to patients with moderate to severe hypertension. Our response is the the decrease in diastolic blood pressure at 12 weeks, adjusted for the patient's age and a measure of the severity of hypertension at study entry. Our aim is to determine whether the data show any difference in response between the two treatment groups. Which of the following modeling approaches would be the single best choice in this scenario?

a. Linear regression to predict the outcome.
b. (Binary) Logistic regression to predict the outcome.

c. Poisson regression to predict the outcome.
d. Multinomial logistic regression to predict the outcome.
e. Ordinal logistic regression to predict the outcome.
f. Cox proportional hazards model to predict the outcome.
g. Log rank test of the intervention's impact on the outcome.
h. None of these modeling strategies are appropriate.

### 28.1 Answer 28 is A

We have a continuous outcome, without any censoring. Linear regression is the best choice. This is, in essence an ANOVA or ANCOVA model.

## 29 Question 29

In *How To Be a Modern Scientist*, Jeff Leek describes some hurdles likely to affect the transition towards reproducibility in scientific work, and some potential solutions related to data sharing. According to Leek, which of these statements are true?

**CHECK ALL OF THE TRUE STATEMENTS.**

a. It is hard to create serious research quality data sets that can be used by others.
b. Existing structures for advancement in academia sometimes are in conflict with the promotion of reproducible research.
c. There is no intermediate form of credit for data generators that counts more heavily than a regular publication.
d. Code books are often formatted using Word or another text editor.
e. The person collecting the data should provide pseudocode to help the statistician in tidying and data management activities.
f. None of the statements above are true.

### 29.1 Answer 29 is that statements A through E are all true.

See the section in *How to be a modern scientist* on "Data Sharing."

## 30 Question 30

A study of self-reported mental health status (recorded in the `status` variable as 1 = good, 0 = not good) was completed in 198 working adults ages 30-35 for whom the following variables were collected:

- `alone` = do they live alone (1 = yes, 0 = no)
- `bmi` = body-mass index category (normal, overweight or obese)
- `employed` = years of employment in current job
- `sex` = M or F
- `sbp` = systolic blood pressure (in mm Hg)

I have provided the `quizd30.Rds` file, which may be useful to you.

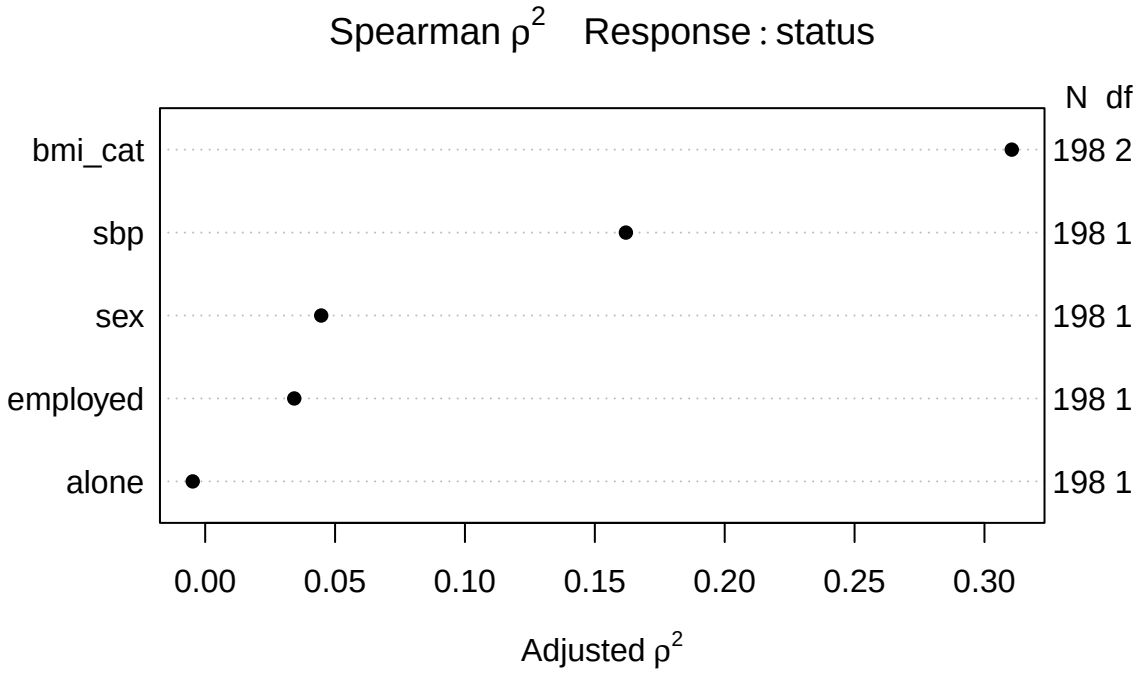
Consider the Display for Item 30. Suppose you are willing to fit a model that uses exactly two additional non-linear predictor terms, using a maximum of four additional degrees of freedom beyond those contained in the main effects model. **Which two** of the following terms would be the most appropriate ones to choose?

a. A restricted cubic spline in `alone` with four knots

b. A restricted cubic spline in `bmi` with four knots
c. A restricted cubic spline in `employed` with four knots
d. A restricted cubic spline in `sex` with four knots
e. A restricted cubic spline in `sbp` with four knots
f. An interaction term between `bmi` and the main effect of `alone`
g. An interaction term between `bmi` and the main effect of `employed`
h. An interaction term between `bmi` and the main effect of `sex`
i. An interaction term between `bmi` and the main effect of `sbp`

### Display for Item 30

```
quizd30 <- readRDS("data/quizd30.Rds")
plot(spearman2(status ~ alone + bmi_cat + employed + sex + sbp,
               data = quizd30), pch = 16)
```




### 30.1 Answer 30 is `E` and `I`

Since `bmi_cat` is highest on the Spearman plot, the most appropriate thing to add is an interaction term between `bmi_cat` and the next variable down on the Spearman plot, which is `sbp`. After that, we move down to `sbp` and a spline would be appropriate.

It's important to realize that `bmi_cat` here is a categorical variable, and thus it's not possible to build a spline for it.

Adding the terms described in `E` and `I` use a total of 10 degrees of freedom, as compared to the 6 used in the model only including main effects, as can be seen in the summary below. So we have used an additional 4 degrees of freedom, as we've been asked to do.

```
d <- datadist(quizd30)
options(datadist = "d")

mod_30_main <- lrm(status ~ alone + bmi_cat + employed +
                       sex + sbp, data = quizd30,
                   x = T, y = T)

mod_30_new <- lrm(status ~ alone + rcs(sbp, 4) + bmi_cat +
                      bmi_cat %ia% sbp + employed + sex,
                  data = quizd30,
                   x = T, y = T)

anova(mod_30_main)
```

```
                Wald Statistics          Response: status

 Factor     Chi-Square d.f. P
 alone       0.37      1    0.5438
 bmi_cat    32.02      2    <.0001
 employed   21.36      1    <.0001
 sex        21.12      1    <.0001
 sbp        29.63      1    <.0001
 TOTAL      35.48      6    <.0001
```

```
anova(mod_30_new)
```

```
                Wald Statistics          Response: status

 Factor                                     Chi-Square d.f. P
 alone                                       0.98       1   0.3226
 sbp  (Factor+Higher Order Factors)         29.81       5   <.0001
  All Interactions                           2.58       2   0.2758
  Nonlinear                                  1.58       2   0.4530
 bmi_cat  (Factor+Higher Order Factors)     31.09       4   <.0001
  All Interactions                           2.58       2   0.2758
 bmi_cat * sbp  (Factor+Higher Order Factors)  2.58     2   0.2758
 employed                                   20.64       1   <.0001
 sex                                        20.26       1   <.0001
 TOTAL NONLINEAR + INTERACTION               4.56       4   0.3354
 TOTAL                                      34.53      10   0.0002
```

## 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.