432 Quiz 2 for Spring 2025

Thomas E. Love, Ph.D.

2025-04-17 03:27 am

Table of contents

0.1 General Instructions

This PDF document is **53** pages long. There are **25** questions on this Quiz.

0.1.1 Links

All of the links you need for this Quiz will be found at https://thomaselove.github.io/432-2025/quiz2.html.

This will include details on how to access:

- the Main Document (this pdf) containing the instructions and questions
- the Google Form Answer Sheet, and
- the **6** data sets we have provided to you
- a "template" for a Quarto file for doing the Quiz that includes all of the code provided in the Quiz to load R packages and obtain the necessary data

If there are any changes to the Quiz after it is first posted, those changes will be posted to the Quiz 2 links page **and** you will be notified of any such changes in an email sent to you by Dr. Love.

0.1.2 Deadline

The deadline to complete your work and submit the Google Form Answer Sheet is **Monday, 2025-04-28 at 9 AM**. 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 one hour after the deadline, so please do not wait until the last moment to submit. We will not accept any responses except through the Google Form.

0.1.3 The Google Form Answer Sheet

The Google Form Answer Sheet, found at **LINK TO BE PROVIDED**, 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 log into Google via CWRU to access it.

If you wish to work on some of Quiz 2 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 2 Answer Sheet. You will then receive a link at your CWRU email which

will allow you to return to the Quiz 2 Answer Sheet as often as you like without losing your progress.

0.1.4 Footnotes are hints.

There are **23** footnotes in this document, including this one[1].

0.1.5 What does the Quiz cover?

Quiz 2 includes material from the first 23 classes in 432, including all of Jeff Leek's *How to be a Modern Scientist*.

0.1.6 Scoring and Timing

Each question is worth **4** points except for Questions 12 and 13, which are worth **6** points each adding to a total of **104** points, but we'll pretend it's out of 100. If a question has two parts (a and b), then each part is worth half of the total points for that question. 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 around 6 hours to complete, assuming a little less than 10 minutes per question. I expect most students will take 3-10 hours, and some will take as little as 2 or as many as 12. Again, it is **not** a good idea to spend a long time on any one question. There are 25 questions. If you spend 20 minutes on each one (which is more time than it should take you to answer most of them, at least) then you will take about 8 hours and 20 minutes to complete the Quiz.

Dr. Love will grade the Quiz, and an answer sketch, grading rubric, and grades will be available by Tuesday 2025-04-29.

0.1.7 Should I Answer All of the Items?

A blank response cannot score better than an incorrect one, a guess might be correct (or at least partially correct), so you should answer all of the items.

[1] Read the footnotes. That's where we put (some of) the hints.

0.2 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 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:

- You can ask your question via email to **Thomas dot Love at case dot edu**, or
- You can ask your question of Dr. Love directly before, after or during Class 28 on 2025-04-24.

During the Quiz period (2025-04-18 through 2025-04-28) we will not answer questions about Quiz 2 except through the email and class session above.

- Please check your email **and** the Quiz 2 links page *before* submitting the final version of your Quiz, to see if Dr. Love has posted any changes there.
- Dr. Love promises to respond to all questions received before 9 PM on 2025-04-27 in a timely fashion.

0.2.1 When Should I ask for Help?

We recommend the following process.

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

0.2.2 A few suggestions about asking questions

- Specific questions are more likely to get helpful answers.
- We will not review your code or your English for you.
- We will not tell you if your answer is correct, or if it is complete.

0.3 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 are loaded in R.

0.4 Packages and Settings used by Dr. Love

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

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

library(janitor); library(naniar)
library(here); library(conflicted)
library(rms)
library(bestglm); library(broom); library(car); library(caret)
library(cobalt); library(countreg); library(cutpointr); library(GGally)
library(glue); library(gt); library(haven); library(MASS)
library(mice); library(mosaic); library(nnet); library(olsrr)
library(patchwork); library(pROC); library(pscl); library(pwr)
library(readxl); library(ROCR); library(rsample); library(survey)
library(survival); library(survminer); library(tableone); library(topmodels)
library(xfun); library(yardstick)

library(easystats)
library(tidyverse)

conflicts_prefer(dplyr::filter, dplyr::select, dplyr::summarize, dplyr::count,
                 base::mean, base::sum, base::max, janitor::clean_names,
                 car::vif, Matrix::update, rms::Predict, pscl::zeroinfl,
                 ROCR::performance)

options(dplyr.summarise.inform = FALSE)

theme_set(theme_lucid())
```

0.5 The Data Sets

I have provided **six** data sets (called **data1.Rds**, **data2.csv**, **data3.xlsx**, **data4.dta**, **data5.sav** and **data6.csv**) that are described below. They may be helpful to you. You'll find these data sets in our Shared Drive in the folder called **432 Quiz 2 Materials (due 2025-04-28 at 9 AM)** and its **data** sub-folder.

I suggest you download these six data sets to a folder called `data` in your R project for Quiz 2, as I have assumed that you've done this in creating the code below.

0.5.1 The `data1.Rds` file (used in Questions 1-3)

In a study of 146 adult (ages 34-82) subjects with Parkinson's disease who have received deep brain stimulation surgery, we have collected a series of pre-operative neuropsychological assessments (measures of cognitive ability and of anxiety, depression, and quality of life) as well as follow-up data on whether the subject was later (post-surgery) diagnosed with dementia[2].

The data in `data1.rds` include the following variables:

Variable	Description
`subject`	subject ID
`days_btw`	days between DBS surgery and last in-study follow-up
`dem_dx`	new dementia diagnosis after DBS surgery (1 = yes, 0 = no)
`bnt`	standardized score on Boston Naming test (100 = average healthy score, higher = healthier)
`long_mem`	score on long-term memory capacity (100 = average healthy score, higher = healthier)
`sdmt`	standardized score on Symbol Digit Modalities test (100 = average healthy score, higher = healthier)
`wech_iq`	Wechsler abbreviated scale of intelligence (100 = average, higher = healthier)
`pdq_tot`	quality of life score for Parkinson's (higher = more symptoms and problems; worse health)
`pdq_grp`	three groups based on `pdq_tot` (names indicate `pdq_tot` ranges, and are [5,30], (30,50], and (50, 101])
`bdi`	Beck Depression Inventory score (higher = more self-reported depression)
`bai`	Beck Anxiety Inventory score (higher = more self-reported anxiety)

[2] Trenton George gathered related data for his 432 Project B in 2024. Thanks, Trenton!

Here is the code that I used to load these data into R. You are encouraged to use it, as well.

```
data1 <- read_rds(here("data/data1.Rds"))
```

Here is a simple summary of the 11 variables in my `data1` tibble, across all 146 observations.

```
summary(data1)
```

```
  subject             days_btw        dem_dx            bnt        
Length:146         Min.   :  50   Min.   :0.0000   Min.   :  0.00  
Class :character   1st Qu.:1256   1st Qu.:0.0000   1st Qu.: 88.86  
Mode  :character   Median :1624   Median :0.0000   Median :102.52  
                   Mean   :1564   Mean   :0.2329   Mean   : 97.07  
                   3rd Qu.:2006   3rd Qu.:0.0000   3rd Qu.:114.52  
                   Max.   :2920   Max.   :1.0000   Max.   :129.03  
   long_mem           sdmt          wech_iq       pdq_tot          pdq_grp  
Min.   : 38.80   Min.   : 21.94   Min.   : 67   Min.   :  5.00   [5,30]  :50  
1st Qu.: 80.55   1st Qu.: 79.31   1st Qu.: 95   1st Qu.: 26.00   (30,50] :48  
Median : 96.65   Median : 89.20   Median :105   Median : 41.00   (50,101]:48  
Mean   : 94.70   Mean   : 87.11   Mean   :104   Mean   : 42.53                
3rd Qu.:108.07   3rd Qu.: 96.81   3rd Qu.:113   3rd Qu.: 55.75                
Max.   :142.69   Max.   :129.24   Max.   :133   Max.   :101.00                
     bdi              bai        
Min.   : 0.00   Min.   : 0.00  
1st Qu.: 6.00   1st Qu.: 6.00  
Median : 9.00   Median :11.00  
Mean   :10.36   Mean   :11.53  
3rd Qu.:14.00   3rd Qu.:16.75  
Max.   :32.00   Max.   :29.00  
```

0.5.2 The `data2.csv` file (used in Questions 5-6 and 8-9)

The subjects in the `data2.csv` file provided to you include responses describing 2265 children at the age of 12 who did not have a diagnosis of autism and who spoke English in their household, obtained from the National Survey of Child Health for 2022[3]. The variables included are specified in the table below.

Variable	Description
`subject`	child (subject) ID code created by Dr. Love
`age`	child's age in years (12 for all subjects)
`female`	child's biological sex at birth (1 = female, 0 = male)
`race_eth`	child's race/ethnicity (4 levels: Hispanic, NH_White, NH_Black, NH_Other)
`hstatus`	child's health status (5 levels: Excellent, VeryGood, Good, Fair, Poor)
`conditions`	Number of current health conditions reported from a list of 25 (no child had > 11 here)
`teeth`	child's dental health status (5 levels: Excellent, VeryGood, Good, Fair, Poor)
`height`	Child's height in centimeters (cm) based on parent's recollection
`bmi`	child's body mass index status (3 levels: Under, Normal, Over)
`bodyimage`	child's concern over body image (3 levels: VeryMuch, Somewhat, NotAtAll)
`sh_ideas`	How well children share ideas or talk about things that really matter with their parents? (3 levels: VeryWell, Somewhat, NotWell)
`grades`	Approximate grade point average on a scale from 0 (F) to 4 (A)
`activity`	# days in past week child was physically active for 60 minutes or more
`income`	household income (4 levels: 1_belowFPL, 2_Low, 3_Mid, 4_High)
`ad_educ`	maximum educational attainment for an adult in household (4 levels: 1_Elem, 2_HS_GED, 3_SomeC, 4_CollG)
`mom_h`	Mom's current mental and physical health both excellent or very good (Yes or No)
`dad_h`	Dad's current mental and physical health both excellent or very good (Yes or No)
`smoking`	Does anyone living in this child's household use cigarettes, cigars, or pipe tobacco? (Yes or No)
`premature`	1 if child was born more than 3 weeks before their due date, 0 if not
`bully`	Did this child bully others, pick on them, or exclude them (do not include siblings or dating partners) in the past 12 months? (1 = yes, 0 = no)

[3]Data source: https://www.childhealthdata.org/dataset/download?rq=15923. Trenton George gathered related data for his 432 Project A in 2024. Thanks again, Trenton!

Variable	Description
bullied	Was this child bullied, picked on, or excluded by other children (do not include siblings or dating partners) in the past 12 months? (1 = yes, 0 = no)
wf_system	Does the child receive care in a well-functioning health care system? (Yes or No)
forgone	During the past 12 months, was there any time when this child needed health care but it was not received?" Here health care is meant to include medical care as well as other kinds of care like dental care, vision care, and mental health services. (Yes or No)
n_support	Child lives in a supportive neighborhood (Yes or No)
momage	Mom's age at child's birth (in years)
wt_samp	Sampling weight for this subject
year	2022 for all subjects

Here is the code that I used to load these data into R. You are encouraged to use it, as well.

```
data2 <- read_csv(here("data/data2.csv"), show_col_types = FALSE) |>
  clean_names() |>
  mutate(across(where(is.character), as_factor)) |>
  mutate(income = fct_relevel(income, "1_belowFPL", "2_Low", "3_Mid", "4_High")) |>
  mutate(mom_h = fct_relevel(mom_h, "No", "Yes")) |>
  mutate(n_support = fct_relevel(n_support, "No", "Yes")) |>
  mutate(subject = as.character(subject))
```

Here is a simple summary of the 27 variables in my `data2` tibble, across all 2265 observations.

```
summary(data2)
```

```
   subject                age          female           race_eth
 Length:2265        Min.   :12   Min.   :0.0000   Hispanic: 267
 Class :character   1st Qu.:12   1st Qu.:0.0000   NH_White:1557
 Mode  :character   Median :12   Median :0.0000   NH_Black: 165
                    Mean   :12   Mean   :0.4892   NH_Other: 276
                    3rd Qu.:12   3rd Qu.:1.0000
                    Max.   :12   Max.   :1.0000

     hstatus       conditions          teeth          height
 Excellent:1444   Min.   : 0.000   Excellent:1000   Min.   :119.0
 VeryGood : 636   1st Qu.: 0.000   VeryGood : 805   1st Qu.:152.0
```

```
Good     : 151   Median : 1.000   Good     : 361   Median :157.0
Fair     :  25   Mean   : 1.149   Fair     :  81   Mean   :156.1
Poor     :   2   3rd Qu.: 2.000   Poor     :  14   3rd Qu.:163.0
NA's     :   7   Max.   :11.000   NA's     :   4   Max.   :191.0
                                                   NA's   :35
    bmi          bodyimage        sh_ideas        grades         activity
Over  : 692   NotAtAll:1555   VeryWell:1385   Min.   :0.500   Min.   :0.000
Normal:1398   Somewhat: 622   Somewhat: 766   1st Qu.:3.300   1st Qu.:1.000
Under : 129   VeryMuch:  73   NotWell :  85   Median :3.600   Median :3.000
NA's  :  46   NA's    :  15   NA's    :  29   Mean   :3.424   Mean   :3.502
                                              3rd Qu.:3.900   3rd Qu.:5.000
                                              Max.   :4.000   Max.   :7.000
                                              NA's   :133     NA's   :16
       income        ad_educ      mom_h       dad_h      smoking
1_belowFPL:290   4_CollG :1396   No  : 831   No  : 654   No  :1927
2_Low     :354   3_SomeC : 540   Yes :1146   Yes :1020   Yes : 297
3_Mid     :655   2_HS_GED: 300   NA's: 288   NA's: 591   NA's:  41
4_High    :966   1_Elem  :  29


  premature         bully           bullied       wf_system  forgone
Min.   :0.000   Min.   :0.0000   Min.   :0.0000   No :2192   No  :2166
1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   Yes:  73   Yes :  93
Median :0.000   Median :0.0000   Median :0.0000              NA's:   6
Mean   :0.109   Mean   :0.1838   Mean   :0.4786
3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :1.000   Max.   :1.0000   Max.   :1.0000
NA's   :36      NA's   :24       NA's   :25
n_support       momage         wt_samp              year
No  : 833   Min.   :18.00   Min.   :   21.92   Min.   :2022
Yes :1368   1st Qu.:26.00   1st Qu.:  350.76   1st Qu.:2022
NA's:  64   Median :30.00   Median :  795.64   Median :2022
            Mean   :29.89   Mean   : 1530.45   Mean   :2022
            3rd Qu.:34.00   3rd Qu.: 1733.85   3rd Qu.:2022
            Max.   :45.00   Max.   :51744.16   Max.   :2022
            NA's   :36
```

0.5.3 The `data3.xlsx` file (used in Questions 10-11)

We are interested here in studying a sample of 902 adults who participated in the 2019 National Health Interview Survey (NHIS) and were current smokers at that time[4].

The data in the Excel file `data3.xlsx` include the following variables:

Variable	Description
`subject`	subject ID
`ecig_s`	e-cigarette use status (Never, Former, Current)
`days_s`	days used tobacco (not an e-cigarette) in the past month
`cigs_m`	number of cigarettes smoked per month
`age`	age in years (top-coded at 85)
`female`	1 = female, 0 = not
`race_eth`	race_ethnicity (NH_White, NH_Black, Hispanic, Other)
`educ`	educational attainment (1_Elem = did not complete HS, 2_HS_GED = high school or GED, 3_College = Bachelor's or Associate's, 4 = GradDeg = post-college degree)
`asthma`	ever told by a health care provider that you have asthma? (1 = Yes, 0 = No)

Here is the code that I used to load these data into R. You are encouraged to use it, as well.

```
data3 <- read_excel(here("data/data3.xlsx"), na = c("", "NA")) |>
  clean_names() |>
  mutate(across(where(is.character), as_factor)) |>
  mutate(educ = fct_relevel(educ, "1_Elem", "2_HS_GED", "3_College",
                            "4_GradDeg")) |>
  mutate(subject = as.character(subject))
```

[4] Liz Stanley aggregated related data for her 432 Project B in 2024. Thanks, Liz!

Here is a simple summary of the 9 variables in my `data3` tibble, across all 902 observations.

```
summary(data3)
```

```
  subject              ecig_s        days_s          cigs_m
Length:902         Never  :482   Min.   : 0.00   Min.   :   0.0
Class :character   Former :279   1st Qu.: 5.00   1st Qu.:  10.0
Mode  :character   Current:141   Median :12.00   Median :  30.0
                                 Mean   :12.78   Mean   :  61.2
                                 3rd Qu.:20.00   3rd Qu.:  67.0
                                 Max.   :30.00   Max.   :1200.0
                                                 NA's   :10
     age        female        race_eth          educ         asthma
Min.   :18.00   No :477   NH_White:571   1_Elem   :107   Min.   :0.000
1st Qu.:32.00   Yes:425   NH_Black:148   2_HS_GED :469   1st Qu.:0.000
Median :43.00             Hispanic:118   3_College:262   Median :0.000
Mean   :45.14             NH_Other: 65   4_GradDeg: 61   Mean   :0.163
3rd Qu.:58.25                            NA's     :  3   3rd Qu.:0.000
Max.   :85.00                                            Max.   :1.000
NA's   :2
```

0.5.4 The `data4.dta` file (used in Questions 12-13 and in Questions 15-16)

We are interested here in studying a random sample of 2482 unique songs from those with information available at the TidyTuesday repository for 2020-01-21, describing songs on Spotify[5]. The data in the Stata file `data4.dta` include the following variables:

Variable	Description
`song`	song identifying code
`genre`	song's genre (5 levels: pop, rap, latin, rock, r&b (rhythm and blues))
`popularity`	popularity score (scaled from 0 to 100, higher = more popular)
`mode`	modality or derived scale of song (major or minor)
`dance`	danceability score (higher = more "danceable", maximum score is 1)
`energy`	energy score (higher = more intense/active, maximum score is 1)
`valence`	positiveness score (higher = more "positive" [happy, cheerful, euphoric], maximum score is 1)
`tempo`	average tempo of song (in beats per minute)
`track_id`	complicated Spotify code for the selected track (identifier)

Here is the code that I used to load these data into R. You are encouraged to use it, as well.

```
data4 <- read_stata(here("data/data4.dta")) |>
  clean_names() |>
  mutate(genre =
           fct_recode(factor(genre), "pop" = "1", "rap" = "2", "latin" = "3",
                      "rock" = "4", "r&b" = "5")) |>
  mutate(mode = fct_recode(factor(mode), "major" = "1", "minor" = "2"))
```

[5]Elizabeth Schultheis aggregated a superset of these data for her 432 Project B in 2024. Thanks, Elizabeth!

Here is a simple summary of the 9 variables in my `data4` tibble, across all 2482 observations.

```
summary(data4)
```

```
    song              genre       popularity       mode          dance        
 Length:2482        pop   :518   Min.   : 0.00   major:1440   Min.   :0.1730  
 Class :character   rap   :627   1st Qu.:24.00   minor:1042   1st Qu.:0.5500  
 Mode  :character   latin:443   Median :46.00                Median :0.6720  
                    rock  :454   Mean   :41.34                Mean   :0.6526  
                    r&b   :440   3rd Qu.:60.00                3rd Qu.:0.7648  
                                  Max.   :98.00                Max.   :0.9670  
    energy          valence           tempo          track_id        
 Min.   :0.0286   Min.   :0.0234   Min.   : 52.65   Length:2482       
 1st Qu.:0.5610   1st Qu.:0.3530   1st Qu.: 96.01   Class :character  
 Median :0.6990   Median :0.5310   Median :117.95   Mode  :character  
 Mean   :0.6801   Mean   :0.5282   Mean   :119.77                     
 3rd Qu.:0.8220   3rd Qu.:0.7000   3rd Qu.:139.41                     
 Max.   :0.9990   Max.   :0.9730   Max.   :214.52                     
```

0.5.5 The `data5.sav` file (used in Questions 17-18)

Here we study data on 2,787 adults (18 years or older) pulled from the National Inpatient Sample[6] for 2019-2020 with a principal diagnosis of shock[7] and who received mechanical ventilation[8]. The NIS is the largest publicly available all-payer inpatient care database in the United States, containing de-identified data on more than seven million hospital stays. Its large sample size allows for a good representation of the US population and generalization of findings

The data in the SPSS file `data5.sav` include the following variables:

Variable	Description
`subject`	subject identifying code
`age`	age in years
`sex`	Female or Male
`race`	White, Black or Other
`hosp_los`	days in hospital (length of stay)
`renal`	Yes = presence / No = absence of renal failure during admission based on ICD-10 diagnosis codes[9]
`chf`	Yes = presence/ No = absence of heart failure during admission based on ICD-10 diagnosis codes[10]
`vent_hrs`	mechanical ventilation (3 levels: below24, 24_to_96, above96)
`procs`	number of procedures during hospitalization (data range: 1 - 25)

The code that I used to load the data from this SPSS (`.sav`) file into R is shown on the top of the next page.

[6] Hala Nas aggregated these data for her 432 Project B in 2024. Thanks, Hala!

[7] ICD-10 diagnosis codes used for shock were: R570, R571, R578, R579 and R6521.

[8] ICD-10 procedure codes used for mechanical ventilation were: 5A09357, 5A1935Z, 5A09457, 5A1945Z, 5A09557 and 5A1955Z.

[9] The codes for renal failure were N170, N171, N172, N179, N178, N184, N185, and N186.

[10] The codes for heart failure were I5020, I5021, I5022, I5023, I5030, I5031, I5032, I5033, I5040, I5041, I5042, I5043, I50810, I50811, I50812, I50813, I50814, I5082, I5083, I5084, I5089, and I509.

Here is the code I used to ingest the data from the `data5.sav` file. You are encouraged to use this code to load the data for yourself.

```
data5 <- read_sav(here("data/data5.sav")) |>
  clean_names() |>
  mutate(sex = fct_recode(factor(sex), "Female" = "1", "Male" = "2")) |>
  mutate(race_eth = fct_recode(factor(race_eth),
                               "White" = "1", "Black" = "2", "Other" = "3")) |>
  mutate(renal = fct_recode(factor(renal), "No" = "1", "Yes" = "2")) |>
  mutate(chf = fct_recode(factor(chf), "Yes" = "1", "No" = "2")) |>
  mutate(chf = fct_relevel(chf, "No", "Yes")) |>
  mutate(vent_hrs = fct_recode(factor(vent_hrs), "24_to_96" = "1",
                           "above96" = "2", "below24" = "3")) |>
  mutate(vent_hrs = factor(fct_relevel(vent_hrs,
                         "below24", "24_to_96", "above96"))) |>
  mutate(age = as.numeric(age), hosp_los = as.numeric(hosp_los),
         procs = as.numeric(procs), subject = as.character(subject))
```

Here is a simple summary of the 9 variables in my `data5` tibble, across all 2787 observations.

```
summary(data5)
```

```
   subject                age          sex         race_eth       hosp_los
 Length:2787        Min.   :18   Female:1281   White:1919   Min.   : 0.00
 Class :character   1st Qu.:61   Male  :1506   Black: 350   1st Qu.: 5.00
 Mode  :character   Median :70                 Other: 518   Median : 9.00
                    Mean   :69                              Mean   :12.18
                    3rd Qu.:79                              3rd Qu.:16.00
                    Max.   :90                              Max.   :72.00
 renal       chf           vent_hrs        procs
 No : 743   No :1449   below24 : 426   Min.   : 1.000
 Yes:2044   Yes:1338   24_to_96: 998   1st Qu.: 2.000
                       above96 :1363   Median : 5.000
                                       Mean   : 6.069
                                       3rd Qu.: 8.000
                                       Max.   :25.000
```

0.5.6 The `data6.csv` file (used in Questions 20-22, and Question 25)

Here we look at data from 1,362 respondents to the General Social Survey for 2018[11]. Visit https://gss.norc.org/us/en/gss/get-the-data.html to obtain similar data. The data in the SAS transport file `data6.csv` include the following variables:

Variable	Description
`subject`	subject identifying code
`satfin`	satisfaction with their financial situation (3 levels: Very = Very Satisfied, Sat = Satisfied, Not = Not Satisfied)
`female`	1 = female, 0 = not female
`partyid`	7 levels ranging from Strong Democrat to Strong Republican as: Strong_Dem, Dem, Lean_Dem, Indep, Lean_Rep, Rep, Strong_Rep
`fam_inc`	inflation-adjusted family income, in dollars
`wordsum`	Number of correct words in a 10-word vocabulary test
`kids`	Number of children subject has ever had

Here is the code that I used to load these data into R. You are encouraged to use it, as well.

```
data6 <- read_csv(here("data/data6.csv"), show_col_types = FALSE) |>
  clean_names() |>
  mutate(across(where(is.character), as_factor)) |>
  mutate(subject = as.character(subject)) |>
  mutate(satfin = factor(fct_relevel(satfin, "Not", "Satisfied", "Very"),
                         ordered = TRUE)) |>
  mutate(partyid = factor(fct_relevel(partyid, "Strong_Dem", "Dem", "Lean_Dem",
                                      "Indep", "Lean_Rep", "Rep", "Strong_Rep")))
```

[11] Samir Memic aggregated related data for his 432 Project B in 2024. Thanks, Samir!

Here is a simple summary of the 7 variables in my `data6` tibble, across all 1362 observations.

```
summary(data6)
```

```
  subject                satfin        female            partyid
Length:1362        Not      :337   Min.   :0.0000   Strong_Dem:228
Class :character   Satisfied:593   1st Qu.:0.0000   Dem       :231
Mode  :character   Very     :432   Median :1.0000   Lean_Dem  :182
                                   Mean   :0.5499   Indep     :239
                                   3rd Qu.:1.0000   Lean_Rep  :157
                                   Max.   :1.0000   Rep       :175
                                                    Strong_Rep:150
   fam_inc          wordsum           kids
Min.   :   168   Min.   : 0.000   Min.   :0.000
1st Qu.: 17373   1st Qu.: 5.000   1st Qu.:0.000
Median : 38527   Median : 6.000   Median :2.000
Mean   : 50054   Mean   : 5.963   Mean   :1.944
3rd Qu.: 70100   3rd Qu.: 7.000   3rd Qu.:3.000
Max.   :158402   Max.   :10.000   Max.   :8.000
```

1 Question 1

Here, I will create a survival object describing days (starting from the date of DBS surgery) until dementia diagnosis (or censoring) using the **data1** tibble, and then create a Kaplan-Meier curve comparing this survival object for the four `pdq_grp` groups.

The output shown in Displays 1 and 2 for Question 1 below may be helpful to you.

Display 1 of 2 for Question 1

```
data1$S <- Surv(time = data1$days_btw, event = data1$dem_dx == "1")
kmfit1 <- survfit(data1$S ~ data1$pdq_grp)
kmfit1
```

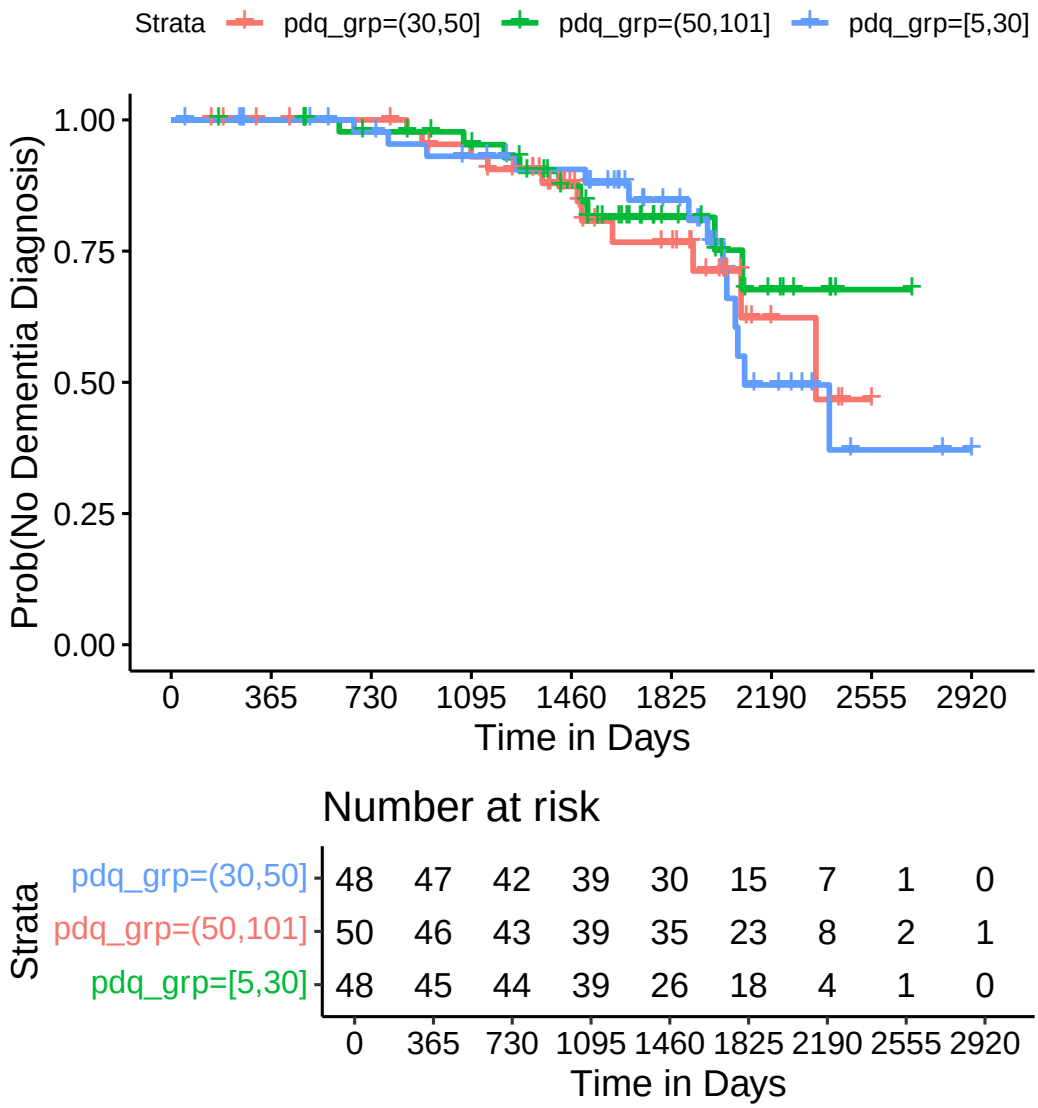
```
Call: survfit(formula = data1$S ~ data1$pdq_grp)

                         n events median 0.95LCL 0.95UCL
data1$pdq_grp=[5,30]    50     14   2092    2027      NA
data1$pdq_grp=(30,50]   48     11   2352    2079      NA
data1$pdq_grp=(50,101] 48      9     NA      NA      NA
```

```
survdiff(data1$S~data1$pdq_grp)
```

```
Call:
survdiff(formula = data1$S ~ data1$pdq_grp)

                         N Observed Expected (O-E)^2/E (O-E)^2/V
data1$pdq_grp=[5,30]    50       14     12.6    0.1652     0.263
data1$pdq_grp=(30,50]   48       11     10.3    0.0486     0.070
data1$pdq_grp=(50,101] 48        9     11.1    0.4138     0.617

 Chisq= 0.6  on 2 degrees of freedom, p= 0.7
```

Display 2 of 2 for Question 1

```
ggsurvplot(kmfit1, data = data1, risk.table = TRUE,
           xlab = "Time in Days", ylab = "Prob(No Dementia Diagnosis)",
           break.time.by = 365, risk.table.height = 0.3)
```



Question 1 continues on the next page.

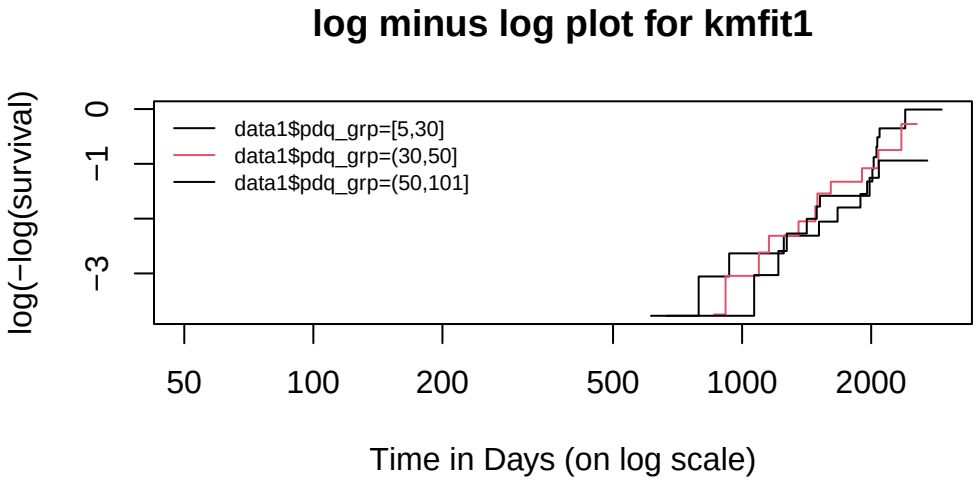
Consider the material on the previous two pages related to Question 1, as well as any relevant additional analyses you feel the need to run on the `data1` tibble. Which of the following statements are true? More than one may be true.

CHOOSE EACH OF THE TRUE STATEMENTS.

a. The lowest numbered `ptid` for a censored subject in the `data1` data is 6.
b. There are more subjects with `pdq_total` scores above 50 than there are subjects with `pdq_total` scores no higher than 30 in the `data1` data.
c. The median time without dementia diagnosis for subjects with `PDQ` in the lowest (5 to 30) group is higher than the median time without dementia diagnosis for subjects with `PDQ` in the middle (30 to 50) group.
d. The log rank test suggests that subjects across the three `PDQ` groups do not have similar rates of avoiding a dementia diagnosis.
e. None of the three PDQ groups contain more than 30 subjects who avoided a dementia diagnosis for at least four years, according to the Kaplan-Meier curve.
f. The colors shown on the risk table in Display 2 **do not** match those shown in the main plot in Display 2.
g. None of the statements above are true.

2 Question 2

Below I have provided a log minus log plot for the comparison shown in Question 1. In a complete English sentence or two, what conclusion should you draw from this plot about our work in Question 1?



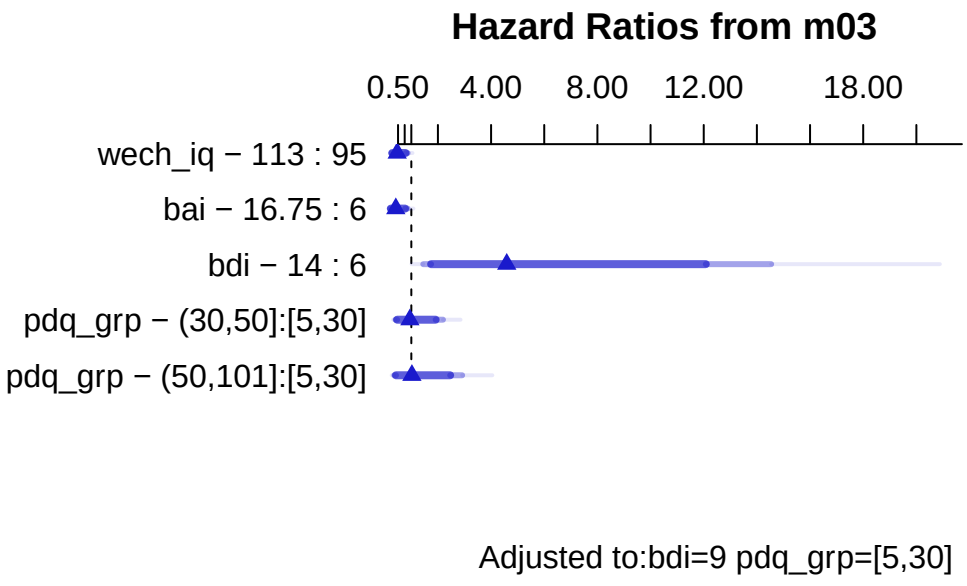
3 Question 3

Fit a Cox proportional hazards model to predict the time-to-event object from Question 1 for all 146 subjects on the basis of four predictors: `wech_iq`, `bai`, `bdi` and `pdq_grp`, including an interaction between PDQ group and the Beck Depression Inventory score, and a restricted cubic spline with 3 knots in the Wechsler IQ score.

Call that Cox model `m03`, although you'll need to figure out how to fit it, since I didn't show it in the output below[12]. Your results should match the effects summary shown below.

```
data1 <- data1 |> select(-S)

d <- datadist(data1)
options(datadist = "d")

## model m03 fit here, hidden from you
```

```
plot(summary(m03), main = "Hazard Ratios from m03")
```



Question 3 continues on the next page.

[12] In fitting the model, I **did not** use the `S` object that I created in Question 1, but instead incorporated the same time-to-event outcome that `S` represented as part of the actual model, using the original variable names from the `data1` tibble and the `Surv` function.

```
summary(m03)
```

```
            Effects              Response : Surv(days_btw, dem_dx == "1")

 Factor                     Low High   Diff. Effect    S.E.    Lower 0.95
 wech_iq                    95  113.00 18.00 -0.759190 0.31088 -1.36850
  Hazard Ratio              95  113.00 18.00  0.468050      NA  0.25449
 bai                         6   16.75 10.75 -0.878340 0.37090 -1.60530
  Hazard Ratio               6   16.75 10.75  0.415470      NA  0.20083
 bdi                         6   14.00  8.00  1.523700 0.58828  0.37070
  Hazard Ratio               6   14.00  8.00  4.589200      NA  1.44870
 pdq_grp - (30,50]:[5,30]    1    2.00    NA -0.056764 0.42903 -0.89764
  Hazard Ratio               1    2.00    NA  0.944820      NA  0.40753
 pdq_grp - (50,101]:[5,30]   1    3.00    NA  0.021131 0.53455 -1.02660
  Hazard Ratio               1    3.00    NA  1.021400      NA  0.35824
 Upper 0.95
 -0.14988
  0.86081
 -0.15138
  0.85952
  2.67670
 14.53700
  0.78411
  2.19050
  1.06880
  2.91200

Adjusted to: bdi=9 pdq_grp=[5,30]
```

Consider the information provided above, along with any other analyses of the `data1` data you wish to perform. Which of the following statements are true? More than one may be true.

CHOOSE EACH OF THE TRUE STATEMENTS

a. The value of the `bdi` score effect shown in this plot applies to subjects whose `pdq_total` score is over 50.
b. The value of the `wech_iq` effect shown in this plot does not apply to subjects whose `pdq_total` score is below 30.
c. The hazard ratio associated with a one point change in the Beck Anxiety Inventory will be closer to 1 than the hazard ratio pictured in the plot above.
d. None of the statements above are true.

4 Question 4

In *How To Be A Modern Scientist*, Jeff Leek includes numerous suggestions about scientific talks. Which two of the following are **NOT** part of Leek's suggestions?

CHECK BOTH RESPONSES NOT SUGGESTED BY LEEK.

a. The most important reasons to speak about your research are to meet people and to make people excited about your ideas and results.
b. Use SlideShare, SpeakerDeck or a similar service to share your slides with people attending your talk, and link your talks on your personal web page.
c. If you are asked a difficult question, don't get upset, and don't be afraid to say "I don't know."
d. Fonts in slides are often too big. Make sure your slides are legible, but don't make the fonts huge.
e. Each figure in your talk should be emphasized. Focus on explaining what the figure is supposed to communicate, what the axes mean, and point out what patterns the audience should look for.
f. When giving a talk to try to get a job, try to speak in as much detail as possible about multiple ideas you are working on.
g. Start off your talk with a brief statement of the problem you are studying that is understandable to everyone.

5 Question 5

This question uses the data in the `data2` tibble.

a. What percentage of the rows included in the `data2` tibble describe children who have described their health status (`hstatus`) as either "Excellent" or "Very Good"? Please express your response as a percentage between 0 and 100, including a single decimal place, and use a complete-case analysis to deal with missing data on the `hstatus` variable[13].

b. Suppose your intent is to create a tibble called `new5` to support a *complete-case analysis* across all of the `data2` variables, not just those discussed in part a of this question. How many different subjects should be included in your `new5` tibble?

[13] Note that your analysis should include all subjects with complete data on the `subject` and `hstatus` variables, regardless of whether or not they have complete data on other variables.

6 Question 6

Next, please answer the question about the `data2` tibble asked in Question 5 part a again, but this time accounting for the sampling weights used in `wt_samp`, again using a complete-case analysis to deal with missing `hstatus` values, as you did in Question 5 part a. What is the resulting estimate of the percentage of the US age 12 population who would describe their General Health as either "Excellent" or "Very Good". Again, express your response as a percentage, with a single decimal place.

7 Question 7

Once a confidence interval for the slope in a simple linear regression model is calculated, several design changes may be used by a researcher to make a confidence interval wider or narrower. For the changes listed in each of the rows below, indicate the impact of that change the width of the confidence interval by selecting the correct column.

Rows:

1. Increase the level of confidence.
2. Increase the sample size.
3. Increase the standard error of the estimate.
4. Use a bootstrap instead of a t-based approach to estimate the CI.

Columns:

a. CI will become wider
b. CI will become narrower
c. CI width will not change
d. It is impossible to tell

8 Question 8

Use single imputation, with a seed of `432`, and the **mice** package, to develop a version of the `data2` tibble that includes complete data (including imputed results) for the following 11 variables and all 2265 subjects.

- `subject`, `female`, `race_eth`, `height`, `bodyimage`, `sh_ideas`, `grades`,
- `activity`, `bullied`, `n_support`, and `wf_system`

Having accomplished that[14], fit a main effects logistic regression model to predict the `bullied` variable using the 9 other variables listed above excluding `subject`.

a. What is the area under the receiver operating characteristic curve[15] for this model? Round your answer to three decimal places, and express the statistic as a proportion, between 0 and 1.

b. What is Tjur's R^2 for this model? Show your result rounded to three decimal places, again expressing the statistic as a proportion, between 0 and 1.

9 Question 9

Here we will use a "best subsets" approach to identify the best model (on the basis of the BIC) using some of the 9 predictors you used in Question 8.

First, we need to turn our predictors which are non-numeric (factors) into numeric variables. That's `race_eth`, `bodyimage`, `sh_ideas`, `n_support` and `wf_system`. I would do this with the following code, assuming q8_si contains your singly imputed version of the data from Question 8…

```
q9_si <- q8_si |> select(-subject)
q9_si <- cobalt::splitfactor(data = q9_si, replace = TRUE, drop.first = TRUE)
glimpse(q9_si)
```

```
Rows: 2,265
Columns: 14
$ female             <dbl> 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, ~
$ race_eth_NH_White  <int> 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, ~
$ race_eth_NH_Black  <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, ~
$ race_eth_NH_Other  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

[14] Note that you should be imputing values for a total of 240 different subjects in the data2 tibble, across 7 of the 11 listed variables.

[15] There is no reason to use validation in either part of Question 8.

```
$ height             <dbl> 155, 135, 152, 160, 152, 163, 157, 180, 147, 145, 1~
$ bodyimage_Somewhat <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ~
$ bodyimage_VeryMuch <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ~
$ sh_ideas_Somewhat  <int> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, ~
$ sh_ideas_NotWell   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ grades             <dbl> 3.7, 3.4, 4.0, 4.0, 3.8, 3.8, 3.6, 2.3, 2.4, 1.4, 3~
$ activity           <dbl> 1, 1, 0, 5, 7, 5, 7, 2, 1, 7, 2, 4, 1, 0, 1, 2, 2, ~
$ bullied            <dbl> 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, ~
$ n_support_Yes      <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, ~
$ wf_system_Yes      <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, ~
```

Now, you can use this `q9_si` data frame to create the necessary `Xy` data frame to allow us to search through our predictor set with a best subsets approach.

When using "best subsets", pick the top model in terms of BIC and use `nvmax = 5, TopModels = 1` as part of your call to the `bestglm()` function. The resulting model will use five of the predictors contained in the `q9_si` data set.

Now, we want you to specify which predictors from Question 8 are included in this set of five predictors. Since some of the predictors represent indicator variables, it is possible that fewer than five of our original predictors are included in this model. Which predictors from our original model `m8` should be included, according to this approach?

CHOOSE ALL VARIABLES THAT ARE INCLUDED IN THE NEW "BEST SUBSETS" MODEL.

a. `female`
b. `race_eth`
c. `height`
d. `bodyimage`
e. `sh_ideas`
f. `grades`
g. `activity`
h. `n_support`
i. `wf_system`

10 Question 10

Using the data from the `data3` tibble, build an appropriate pair of models to predict `ecig_s` for all of the subjects included in that tibble, one of which assumes proportional odds and one of which does not, using the following five predictors: `age`, `female`, `race_eth`, `educ`, and `asthma`.

Call the model that assumes proportional odds: `m10a` and the model that does not assume proportional odds: `m10b`. In each model, assume MCAR, and use an appropriate approach to dealing with missing data in light of that assumption to specify your models. Also, be sure to make `ecig_s` into an ordered factor before running model `m10a`, but it isn't necessary for you to rescale or transform any of the predictors.

Once you've fit the two models, use `plot(compare_performance(m10a, m10b))` to compare them. In a sentence or two, tell me which of the two models seems more appropriate in this setting, on the basis of what you see in that plot[16]. You are welcome to also use `compare_performance(m10a, m10b)` in preparing your response, should you find that helpful.

11 Question 11

Now, returning to the two models (`m10a` and `m10b`) that you fit in Question 10 to the data in the `data3` tibble, obtain a table of the predicted `ecig_s` results from each model. Use those two tables to answer the following questions.

a. What fraction of the subjects whose `ecig_s` was actually Never were correctly predicted by the `m10a` model?

b. What fraction of the subjects whose `ecig_s` was predicted by `m10b` to be Former were actually Former users of e-cigarettes?

c. In which model (`m10a` or `m10b`) do we see higher accuracy of predictions overall? In answering this question, specify the overall percentage of predictions that are accurate using each of the models, rounded to one decimal place. Does the difference in accuracy seem important?

[16] You are not providing me with your plot, so you'll have to explain what it indicates carefully to me in your response to this Question.

12 Question 12 (6 points)

Using the data from the `data4` tibble, I fit a linear model to predict the `popularity` of a song on the basis of its scores on danceability, positiveness, tempo and mode. Here is some output from the model.

a. Write a complete sentence or two to describe the meaning of the point estimate and the 99% confidence interval for the `mode` effect in model `m12`.

b. Write a complete sentence or two to describe the meaning of the point estimate and the 99% confidence interval for the danceability (`dance`) effect in model `m12`.

```
m12 <- lm(popularity ~ dance + valence + tempo + mode, data = data4)

model_parameters(m12, ci = 0.99)
```

```
Parameter    | Coefficient |   SE |          99% CI | t(2477) |      p
-------------------------------------------------------------------------
(Intercept)  |       31.17 | 3.28 | [ 22.73, 39.62] |    9.52 | < .001
dance        |       17.95 | 3.46 | [  9.03, 26.88] |    5.19 | < .001
valence      |       -8.23 | 2.26 | [-14.04, -2.41] |   -3.65 | < .001
tempo        |        0.03 | 0.02 | [ -0.02,  0.07] |    1.49 | 0.136
mode [minor] |       -0.55 | 0.97 | [ -3.04,  1.95] |   -0.57 | 0.572


Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a Wald t-distribution approximation.
```

```
model_performance(m12)
```

```
# Indices of model performance

AIC       |      AICc |       BIC |    R2 | R2 (adj.) |   RMSE |  Sigma
-----------------------------------------------------------------------------
22771.620 | 22771.654 | 22806.521 | 0.012 |     0.011 | 23.712 | 23.736
```

13 Question 13 (6 points)

Create and study an appropriate set of six plots using `check_model()` for the model `m12` fit in Question 12[17]. The plots you need to generate are:

- a posterior predictive check comparing model-predicted results to the actual outcome data,
- a plot of residuals vs. fitted values,
- a plot of the square root of the absolute value of the standardized residuals vs. the fitted values,
- a plot of standardized residuals vs. leverage values,
- a Normal Q-Q plot of the regression model residuals (I suggest you detrend this plot,) and
- a plot or (I suppose) a table listing the relevant variance inflation factors

Then, in one paragraph (limit yourself to at most six sentences and 500 characters[18]) describe whatever you feel are the three most important conclusions from your plots as they relate to the assumptions of the linear regression model `m12` and the data in `data4`.

14 Question 14

An investigator plans to replicate part of a study of the gut hormone fragment peptide YY_{3-36} (PYY) which reduces appetite and food intake when infused into subjects of normal weight. The original study is found at https://www.nejm.org/doi/full/10.1056/nejmoa030204, if you are curious.

> In common with the adipocyte hormone leptin, PYY reduces food intake by modulating appetite circuits in the hypothalamus. However, in obesity there is a marked resistance to the action of leptin, which greatly limits its therapeutic effectiveness.

The investigator wants to know whether obese subjects are also resistant to the anorectic effects of PYY. She intends to perform a randomized, placebo-controlled, double-blind crossover study on healthy obese subjects (including **an equal number of male subjects and female subjects**), with each subject studied on two occasions one week apart.

The subjects will be screened by a dietitian who will assess their eating behavior with (several established scales). The protocol specifies that:

[17] You need to create the plots, but you won't be showing the plots to me, or including them in any way in your response to Question 13.

[18] 500 characters is about half the length of Question 13

> (Study participants will) complete a three-day diet diary to permit us to assess their usual eating habits. Food preferences were assessed at screening (to) ensure that the food offered at the buffet lunch is acceptable. … The subjects' food intake for the 48 hours before each study day (will be) standardized, and during this period they (will complete) food diaries to confirm compliance. … (On each study day, cannulas will be) inserted into veins in both forearms, one for the collection of blood and the other for the infusion of PYY or saline. … Two hours after the infusion, the subjects (will be) offered a buffet lunch with food in such excess that all appetites (will be) satisfied. The amounts of food and water (will be) quantified preprandially and postprandially, and the caloric intake calculated.

On two consecutive Thursdays, we will measure caloric intake during a buffet lunch offered two hours after the infusion of that week's exposure. In one of the weeks, the subject will receive an infusion of PYY, and in the other week (with the order of the weeks determined at random) the subject will receive a placebo. The number of calories consumed at each lunch is measured and then converted to an appetite rating. Our primary outcome is the difference between the appetite rating after PYY and the appetite rating after placebo.

A clinically meaningful difference, the investigator tells you, would be one in which these comparisons would differ by 20 or more points on the appetite rating scale comparing the two infusions, which is 40% of the anticipated standard deviation of these results. The investigator then asks what the *smallest possible* number of patients is that she will have to enroll and gather data from in order to have at least 90% power to detect an effect of this size using a 5% two-tailed significance level, and to meet all other requirements described above.

a. Specify the number that the investigator asked for.

b. Specify a single line of R code (you may use up to two pipes) that yields the information necessary to complete Question 14 part a.

15 Question 15

I used the `data4` tibble to fit the following set of models to predict `genre` based on various combinations of the two predictors `energy` and `mode`.

```
options(contrasts = c("contr.treatment", "contr.poly"))

m15_1 <- multinom(genre ~ 1, data = data4, trace = FALSE)
m15_E <- multinom(genre ~ energy, data = data4, trace = FALSE)
m15_M <- multinom(genre ~ mode, data = data4, trace = FALSE)
m15_EM <- multinom(genre ~ energy + mode, data = data4, trace = FALSE)
m15_SAT <- multinom(genre ~ energy * mode, data = data4, trace = FALSE)
```

Here are some summary results[19] for these models:

Model	edf	nobs	`rap` int.	`latin` int.	`rock` int.	`r&b` int.	R^2	AIC	BIC
`m15_1`	4	2482	0.191	-0.156	-0.132	-0.163	0.000	7948.28	7971.54
`m15_E`	8	2482	1.101	-0.385	-1.106	1.916	0.022	7782.02	7828.56
`m15_M`	8	2482	0.037	-0.226	-0.256	-0.292	0.004	7923.88	7970.42
`m15_EM`	12	2482	0.955	-0.453	-0.991	1.786	0.026	7757.12	7826.92
`m15_SAT`	16	2482	0.876	-0.673	-0.800	1.798	0.027	7761.08	7854.15

Note that one of the five 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 which uses the fewest degrees of freedom.
b. The model which has the largest intercept for `latin`.
c. The model which has the most negative intercept for `r&b`.
d. The model with the smallest Bayes Information Criterion.
e. The model including the interaction of `energy` and `mode`.

[19] The (int. = intercept) terms shown in the table of summary results have not been exponentiated. If they had been, of course, they would all be positive.

16 Question 16

Consider the five displays for Question 16 shown below and on the next few pages. These five displays describe the five models we built in Question 15, with each display corresponding to a different model, based on the data in the `data4` tibble. Associate each display with the correct model for that display.

Rows: a. Display A for Question 16, b. Display B for Question 16, c. Display C for Question 16, d. Display D for Question 16, e. Display E for Question 16

Columns: 1. `m15_1`, 2. `m15_E`, 3. `m15_M`, 4. `m15_EM` and 5. `m15_SAT`

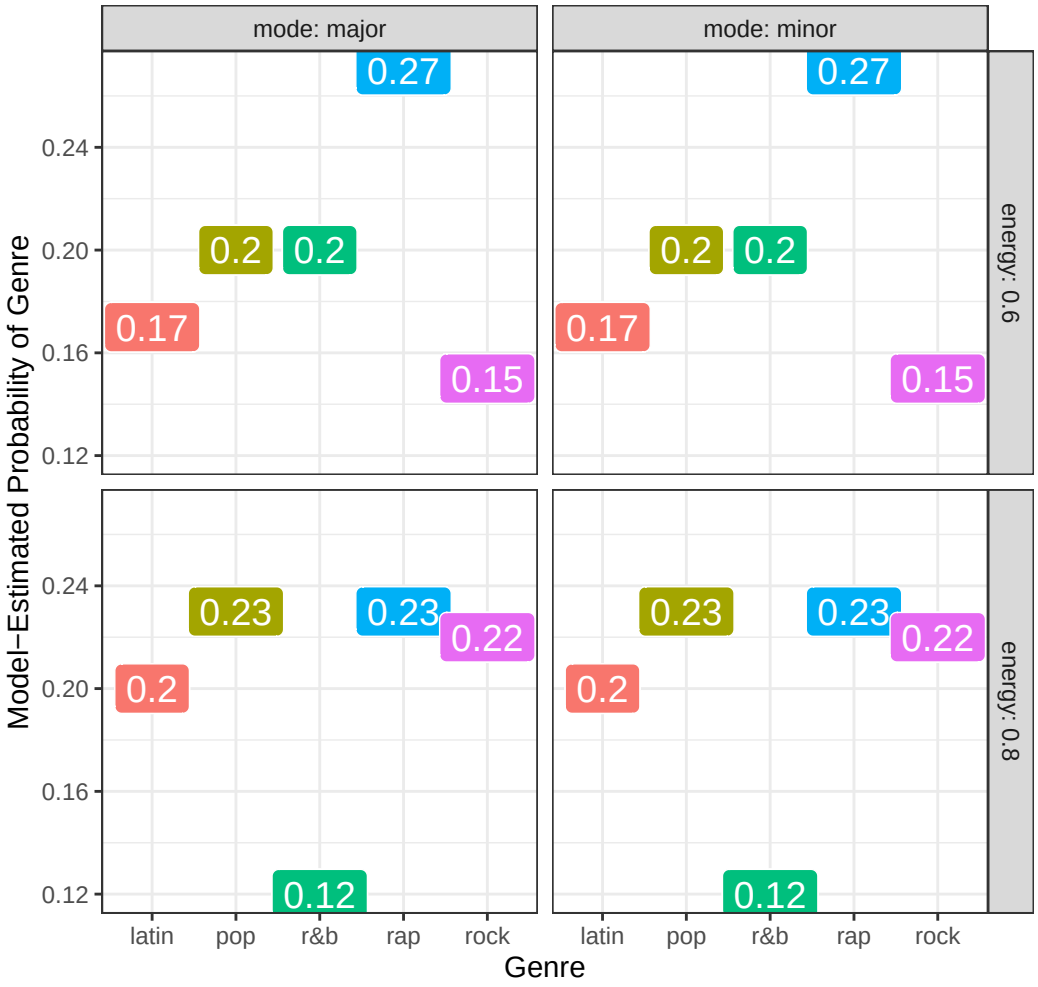
Display A for Question 16

```
tidy(displayA, exponentiate = TRUE) |> select(y.level, estimate, std.error, term)
```

```
# A tibble: 16 x 4
   y.level estimate std.error term
   <chr>      <dbl>     <dbl> <chr>
 1 rap        2.40      0.308 (Intercept)
 2 rap        0.291     0.438 energy
 3 rap        1.74      0.497 modeminor
 4 rap        0.749     0.703 energy:modeminor
 5 latin      0.510     0.360 (Intercept)
 6 latin      1.88      0.492 energy
 7 latin      2.04      0.573 modeminor
 8 latin      0.464     0.791 energy:modeminor
 9 rock       0.449     0.349 (Intercept)
10 rock       2.95      0.473 energy
11 rock       0.373     0.638 modeminor
12 rock       2.51      0.855 energy:modeminor
13 r&b        6.04      0.313 (Intercept)
14 r&b        0.0389    0.468 energy
15 r&b        1.38      0.512 modeminor
16 r&b        1.04      0.755 energy:modeminor
```

Display B for Question 16

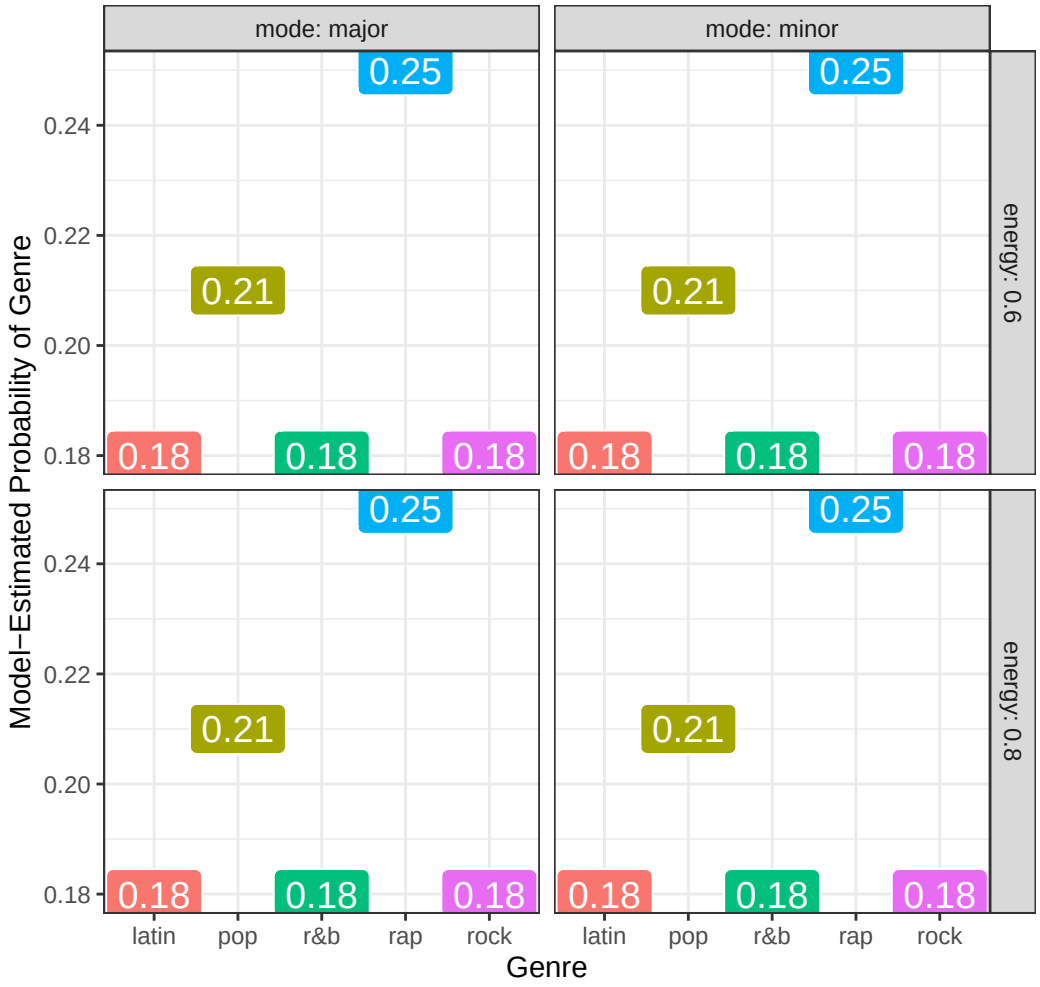
```
ggplot(displayB, aes(x = gen_pred, y = prob, fill = gen_pred)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(energy ~ mode, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Genre", x = "Genre") +
  theme_bw()
```



Note that the probability values plotted here show the fitted probability of the song belonging to each genre, according to the model fit for this display. Note the information in the labels for the four facets in the plot.

Display C for Question 16

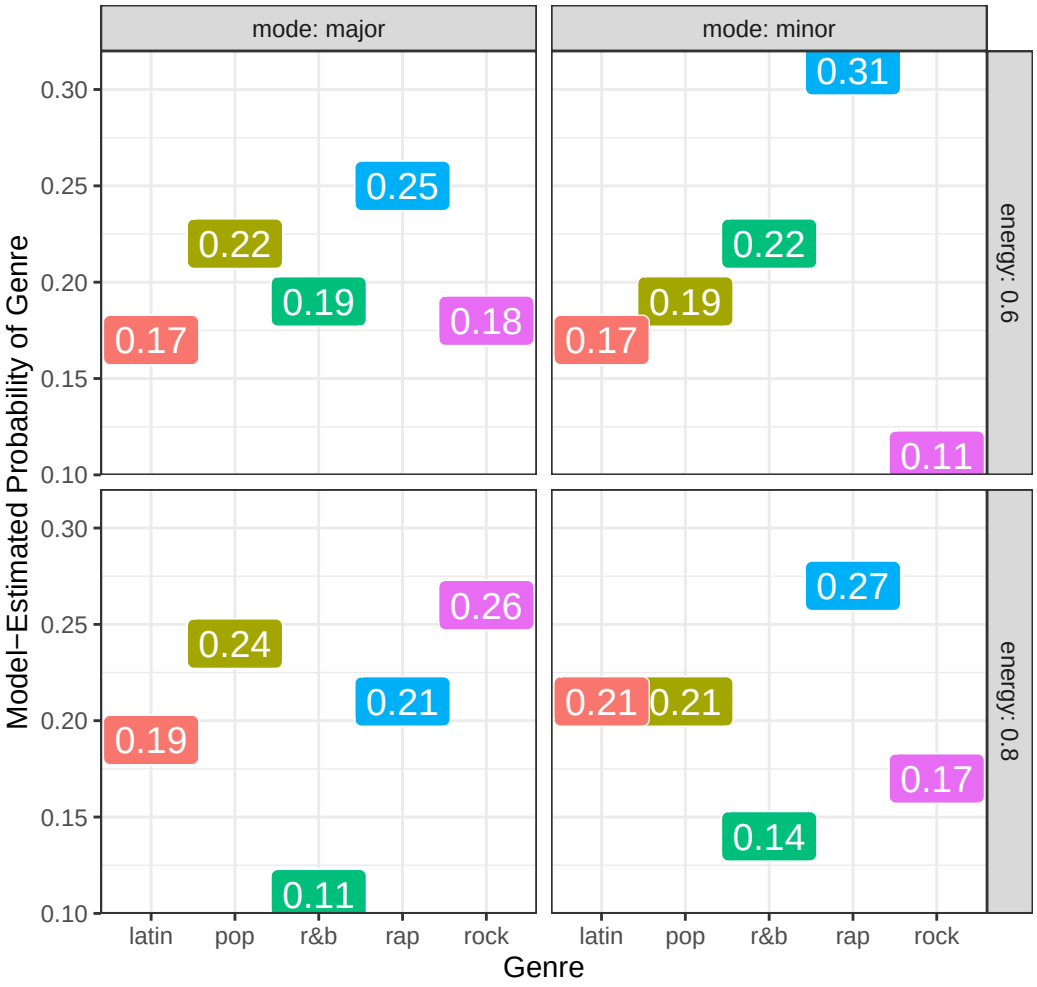
```
ggplot(displayC, aes(x = gen_pred, y = prob, fill = gen_pred)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(energy ~ mode, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Genre", x = "Genre") +
  theme_bw()
```



Note that the probability values plotted here show the fitted probability of the song belonging to each genre, according to the model fit for this display. Note the information in the labels for the four facets in the plot.

Display D for Question 16

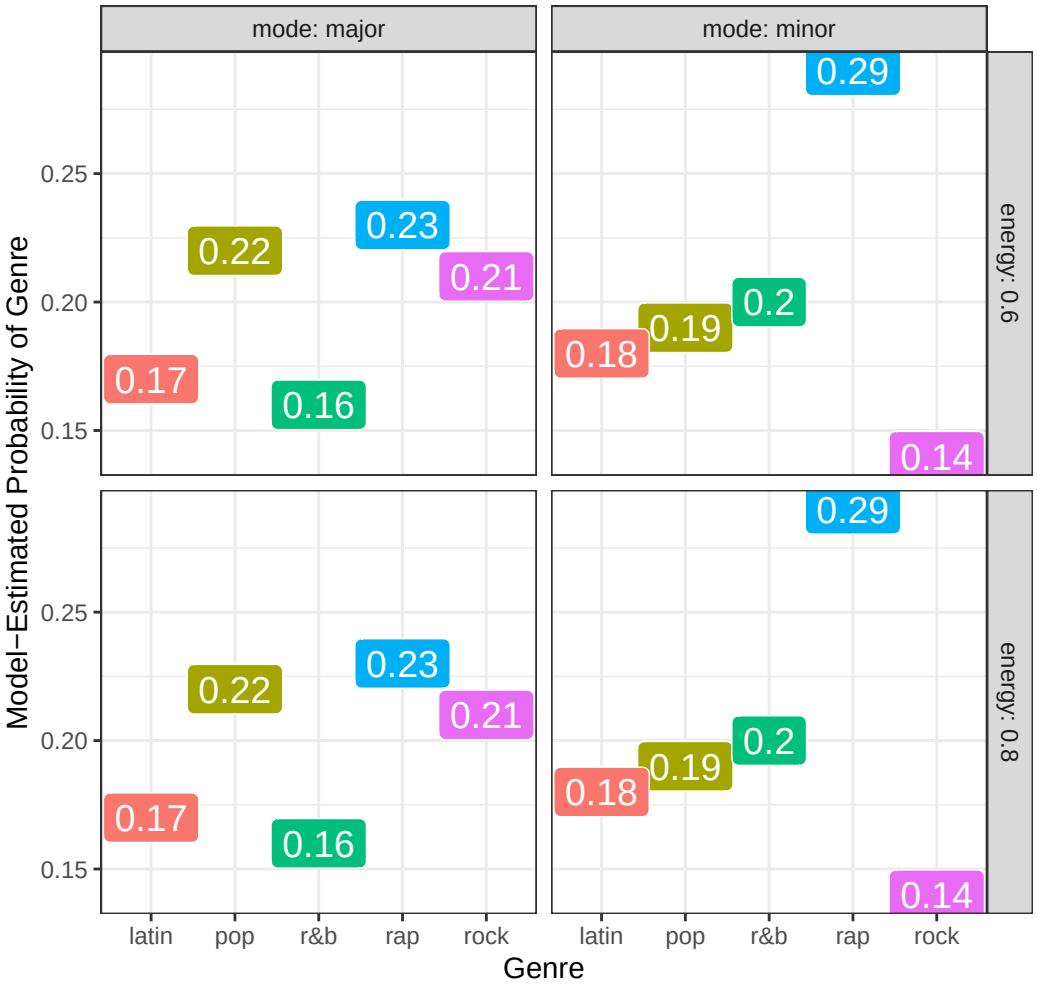
```
ggplot(displayD, aes(x = gen_pred, y = prob, fill = gen_pred)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(energy ~ mode, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Genre", x = "Genre") +
  theme_bw()
```



Note that the probability values plotted here show the fitted probability of the song belonging to each genre, according to the model fit for this display. Note the information in the labels for the four facets in the plot.

Display E for Question 16

```
ggplot(displayE, aes(x = gen_pred, y = prob, fill = gen_pred)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(energy ~ mode, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Genre", x = "Genre") +
  theme_bw()
```



Note that the probability values plotted here show the fitted probability of the song belonging to each genre, according to the model fit for this display. Note the information in the labels for the four facets in the plot.

This is the end of the output for Question 16.

17 Question 17

Consider the following three statements, and the output below (and on the next two pages) that describes some modeling using the `data5` tibble.

Statement I. The Poisson regression model provides a worse fit than the Negative Binomial regression, according to the Bayes information criterion.

Statement II The rootogram for the Poisson model indicates that the Poisson model predicts fewer lengths of stay of 0 to 5 days than we actually observed.

Statement III. The rootogram for the Negative Binomial model indicates a substantially better fit than the rootogram for the Poisson model.

In light of the modeling results shown in Displays 1, 2 and 3 for Question 17 below and on the next two pages, which of these statements are true?

a. I only.
b. II only.
c. III only.
d. I and II
e. I and III
f. II and III
g. All three statements.
h. None of these three statements.

Question 17 Display 1: Regression Models using `data5`

```
m17_p <- glm(hosp_los ~ procs + renal + chf + vent_hrs + age,
             family = poisson(), data = data5)
m17_nb <- glm.nb(hosp_los ~ procs + renal + chf + vent_hrs + age,
                 data = data5)

model_parameters(m17_p, exponentiate = TRUE, ci = 0.90)
```

```
Parameter           |   IRR |       SE |         90% CI |      z |      p
-----------------------------------------------------------------------------
(Intercept)         | 12.58 |     0.41 | [11.92, 13.28] |  76.80 | < .001
procs               |  1.07 | 9.26e-04 | [ 1.07,  1.07] |  81.80 | < .001
renal [Yes]         |  1.04 |     0.01 | [ 1.02,  1.06] |   3.06 | 0.002
chf [Yes]           |  0.92 |     0.01 | [ 0.91,  0.94] |  -7.28 | < .001
vent hrs [24_to_96] |  0.67 |     0.01 | [ 0.66,  0.69] | -25.67 | < .001
```

```
vent hrs [above96]  |  0.73 |     0.01 | [ 0.71,  0.75] | -21.74 | < .001
age                 |  1.00 | 3.97e-04 | [ 1.00,  1.00] |  -8.91 | < .001


Uncertainty intervals (profile-likelihood) and p-values (two-tailed)
  computed using a Wald z-distribution approximation.
```

```
model_parameters(m17_nb, exponentiate = TRUE, ci = 0.90)
```

```
Parameter           |   IRR |       SE |         90% CI |      z |      p
-------------------------------------------------------------------------
(Intercept)         | 12.41 |     0.97 | [10.90, 14.13] |  32.16 | < .001
procs               |  1.08 | 2.68e-03 | [ 1.08,  1.09] |  32.67 | < .001
renal [Yes]         |  1.02 |     0.03 | [ 0.97,  1.07] |   0.69 | 0.488
chf [Yes]           |  0.91 |     0.02 | [ 0.88,  0.95] |  -3.60 | < .001
vent hrs [24_to_96] |  0.64 |     0.02 | [ 0.60,  0.68] | -11.73 | < .001
vent hrs [above96]  |  0.71 |     0.03 | [ 0.67,  0.75] |  -9.53 | < .001
age                 |  1.00 | 9.29e-04 | [ 0.99,  1.00] |  -3.96 | < .001


Uncertainty intervals (profile-likelihood) and p-values (two-tailed)
  computed using a Wald z-distribution approximation.
```

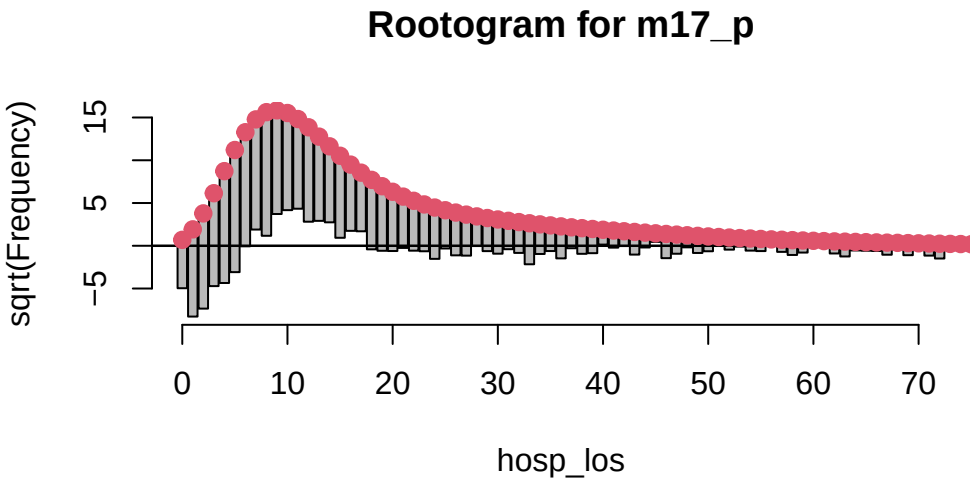
```
model_performance(m17_p, metrics = "common")
```

```
# Indices of model performance

AIC       |       BIC | Nagelkerke's R2 |  RMSE
-----------------------------------------------
25208.547 | 25250.076 |           0.920 | 8.806
```

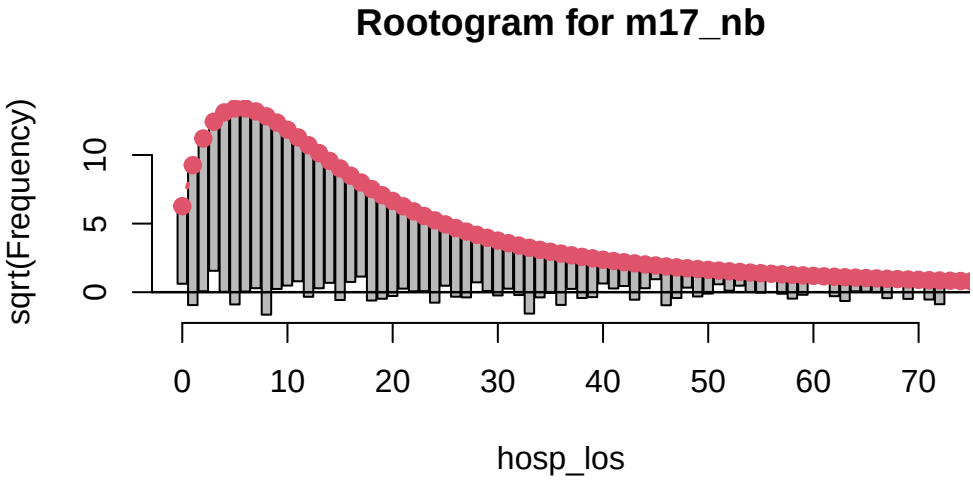
```
model_performance(m17_nb, metrics = "common")
```

```
# Indices of model performance

AIC       |       BIC | Nagelkerke's R2 |  RMSE
-----------------------------------------------
18256.175 | 18303.637 |           0.483 | 9.065
```

Question 17 Display 2: A rootogram for the `m17_p` model



Question 17 Display 3: A rootogram for the `m17_nb` model



This is the end of the output for Question 17.

18 Question 18

Fit the Poisson regression model (`m17_p`) that I fit for Question 17 to the data in the `data5` tibble, then use the model to make a prediction for `hosp_los` for the three new subjects (named Amy, Bart and Chris) listed below.

Name	`procs`	`renal`	`chf`	`vent_hrs`	`age`
Amy	2	Yes	No	below24	63
Bart	5	No	Yes	above96	70
Chris	8	No	No	24_to_96	79

a. Which of the three new subjects (Amy, Bart or Chris) has the highest predicted length of stay in the hospital according to the `m17_p` model?

b. What is that subject's predicted `hosp_los`? Please round your predicted `hosp_los` to zero decimal places.

19 Question 19

In Question 19, we focus on a tobacco cessation study[20] that began on day 0, and we have available the `startday` and `exitday` for each subject. The study compares three `treatments` (called A, B and usual care). 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`). Some summaries of the `dat19` tibble are shown in the Display for Question 19, on the next page, but note that I have not provided the data in `data19` to you.

Suppose you want to add a survival object called `S` to the `dat19` tibble, 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 below will accomplish this?

a. Chunk I only.
b. Chunk II only.
c. Chunk III only.
d. Chunks I and II.
e. Chunks I and III.
f. Chunks II and III.
g. All three Chunks.
h. None of these Chunks.

Chunk I

```
dat19$S = Surv(time = dat19$exitday - dat19$startday,
               event = dat19$exitreason %in% c("lost", "studyend"))
survdiff(S ~ treatment, data = dat19)
```

Chunk II

```
dat19$S <- Surv(time = dat19$exitday, event = dat19$exitreason)
survdiff(S ~ treatment, data = dat19)
```

Chunk III

```
dat19$S = Surv(time = dat19$exitday - dat19$startday,
                event = dat19$exitreason == "achieved")
survdiff(S ~ treatment, data = dat19)
```

Question 19 continues on the next page.

[20] We have not provided a data set for question 19.

response	min	Q1	median	Q3	max	mean	sd	n	missing
startday	0.00	0.00	24.00	29.00	41.00	19.47	13.18	140	0
exitday	17.63	42.75	53.53	69.62	93.45	55.53	18.12	140	0

Display for Question 19

```
dat19 |> df_stats(~ startday + exitday) |>
  gt() |> fmt_number(min:sd, decimals = 2)
```

```
# note: results shown at the top of this page
```

```
dat19 |> tabyl(treatment, exitreason) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_title()
```

```
           exitreason
 treatment   achieved lost studyend Total
         A         13    7       13    33
        UC         26   15       27    68
         B         19    8       12    39
     Total         58   30       52   140
```

```
dat19
```

```
# A tibble: 140 x 4
   startday exitday exitreason treatment
      <dbl>   <dbl> <fct>      <fct>
 1        0    34.2 lost       B
 2        0    23.2 lost       UC
 3        0    38.3 lost       A
 4        0    24.8 achieved   UC
 5        0    31.1 achieved   B
 6        0    32.0 achieved   UC
 7        0    53.2 achieved   B
 8        0    42.5 achieved   UC
 9        0    27.9 achieved   A
10        0    38.2 achieved   UC
# i 130 more rows
```

This is the end of the output for Question 19.

20 Question 20

This question uses the `data6` tibble. Fit an appropriate logistic regression model, which I'll call model `m20`, to predict the ordinal category `satfin`, using the main effects (only) of four predictors: `female`, `partyid`, `kids` and `wordsum`. Assume that the proportional odds assumption is reasonable for this question. Use this `m20` model to predict the happiness levels for all 1362 subjects in the `data6` tibble.

a. What is the percentage (rounded to one decimal place) of the 1362 subjects in `data6` for which model `m20` predicts the correct happiness category?

b. How many of the 1362 subjects in `data6` are predicted by model `m20` to be in the "Very Satisfied" category?

c. How many of the subjects who are actually "Not Satisfied" had their satisfaction level correctly predicted by model `m20`?

21 Question 21

Now fit another appropriate logistic regression model to the `data6` tibble, which I'll call model `m21`, to predict the ordinal category `satfin` using only two of the original predictors: `partyid` and `wordsum`, along with an interaction effect of the two predictors, on financial satisfaction. Again use all 1362 observations in `data6` to build your model.

a. Which of the two models fit so far (`m20` or `m21`) has the better AIC value? Does that model also have the better BIC?

b. What is the p value for the likelihood ratio test[21] of the interaction term in model `m21`? Round your answer to two decimal places.

[21] Hint: you'll need to compare your `m21` to another model to obtain this result.

22 Question 22

Again using all 1362 observations in the `dat6` tibble, now fit a multinomial regression model to predict `satfin` using the same predictors that you used in `m20`. Which of the following statements are true? More than one may be true.

CHECK ALL OF THE TRUE STATEMENTS.

a. The AIC of the multinomial model is an improvement over that of model `m20`.
b. The BIC of the multinomial model is an improvement over that of model `m20`.
c. A test of the proportional odds assumption made in `m20` suggests we should be comfortable with that assumption.
d. The multinomial model requires the estimation of six additional parameters, compared to model `m20`.
e. None of these statements are true.

23 Question 23

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? More than one can be 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. Codebooks 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.

24 Question 24

Consider the Figure for Q24 on the next page, which contains plots associated with four logistic regression models for the same outcome. The C statistics are 0.551 0.701, 0.801 and 0.901. For each of the four plots, select the correct C statistic from the list provided.

Rows:

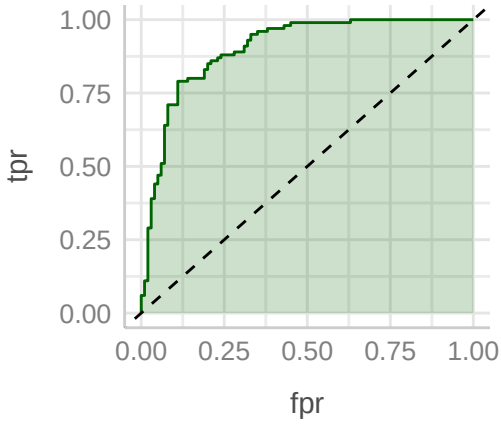
a. Plot A
b. Plot B
c. Plot C
d. Plot D

Columns:

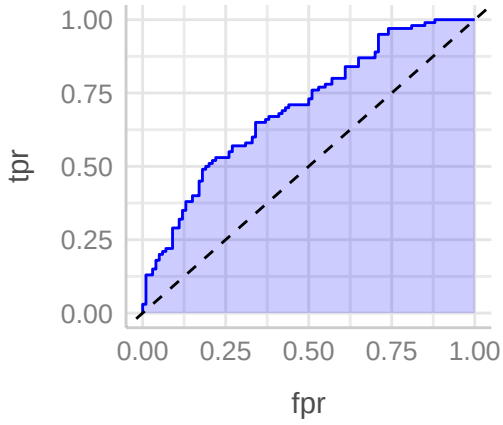
1. 0.551
2. 0.701
3. 0.801
4. 0.901

Figure for Q24

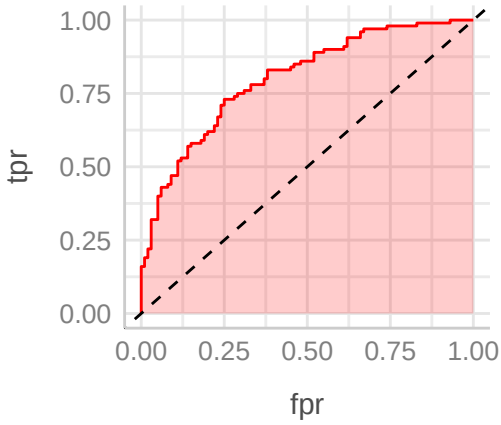
Plot A for Q24



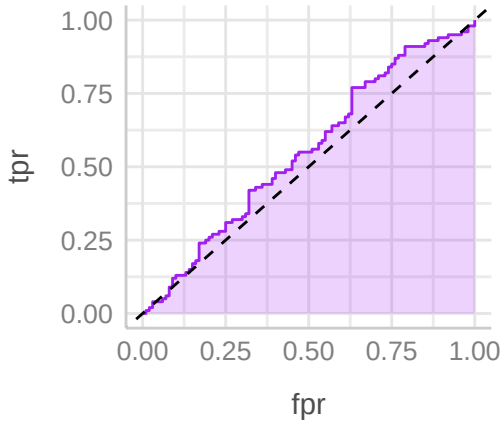
Plot B for Q24



Plot C for Q24

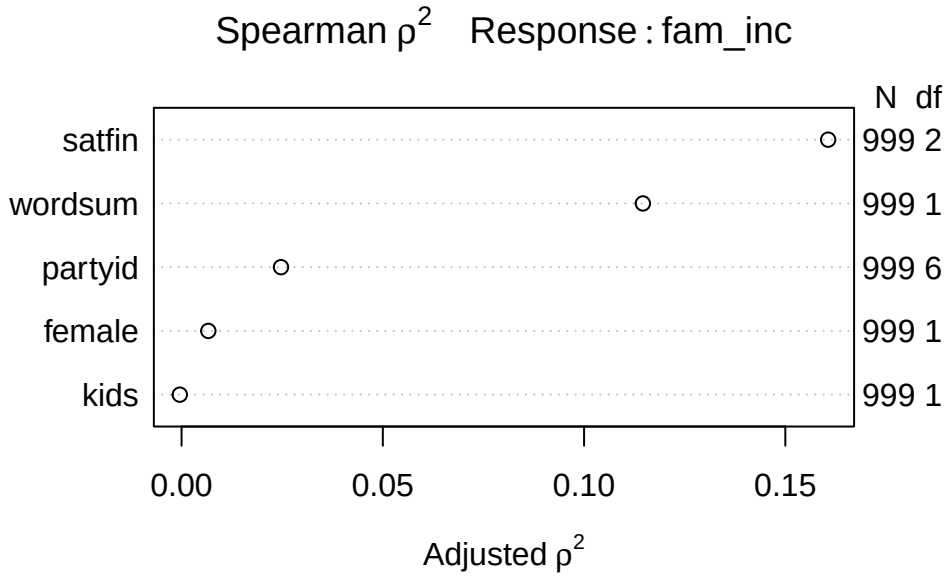


Plot D for Q24



25 Question 25

The Spearman ρ^2 plot below describes a random sample of 999 observations drawn from the `data6` tibble[22]. For this question, assume I am using `satfin` as an unordered factor, rather than an ordered one. Suppose we fit a linear regression model to predict family income using the five predictors listed in the figure along with a single non-linear term.



Suppose you were to add the single non-linear term recommended by the Spearman ρ^2 plot to the "main effects" model[23].

a. What is the non-linear term you would add to the main effects model?

b. How many **additional** degrees of freedom will be required, beyond those used in the main effects model?

[22] I haven't told you which random sample.

[23] A "main effects" model for `fam_inc` using these five predictors (and an intercept term) spends 11 degrees of freedom.

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. Session information follows.

Session Information

```
session_info()
```

```
R version 4.4.3 (2025-02-28 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)

Locale:
  LC_COLLATE=English_United States.utf8
  LC_CTYPE=English_United States.utf8
  LC_MONETARY=English_United States.utf8
  LC_NUMERIC=C
  LC_TIME=English_United States.utf8

Package version:
  abind_1.4-8            askpass_1.2.1          backports_1.5.0
  base64enc_0.1-3        bayestestR_0.15.2      bestglm_0.37.3
  bigD_0.3.1             bit_4.6.0              bit64_4.6.0-1
  bitops_1.0.9           blob_1.2.4             boot_1.3-31
  broom_1.0.8            bslib_0.9.0            cachem_1.1.0
  callr_3.7.6            car_3.1-3              carData_3.0-5
  caret_7.0-1            caTools_1.18.3         cellranger_1.1.0
  checkmate_2.3.2        chk_0.10.0             class_7.3-23
  cli_3.6.4              clipr_0.8.0            clock_0.7.3
  cluster_2.1.8          cobalt_4.6.0           coda_0.19-4.1
  codetools_0.2-20       colorspace_2.1-1       commonmark_1.9.5
  compiler_4.4.3         conflicted_1.2.0       correlation_0.8.7
  corrplot_0.95          countreg_0.3-0         cowplot_1.1.3
  cpp11_0.5.2            crayon_1.5.3           curl_6.2.2
  cutpointr_1.2.0        data.table_1.17.0      datasets_4.4.3
  datawizard_1.0.2       DBI_1.2.3              dbplyr_2.5.0
  Deriv_4.1.6            diagram_1.6.5          digest_0.6.37
  distributions3_0.2.2   doBy_4.6.26            dplyr_1.1.4
```

dtplyr_1.3.1	e1071_1.7.16	easystats_0.7.4
effectsize_1.0.0	emmeans_1.11.0	estimability_1.5.1
evaluate_1.0.3	exactRankTests_0.8.35	fansi_1.0.6
farver_2.1.2	fastmap_1.2.0	fontawesome_0.5.3
forcats_1.0.0	foreach_1.5.2	foreign_0.8-90
Formula_1.2-5	fs_1.6.6	furrr_0.3.1
future_1.40.0	future.apply_1.11.3	gargle_1.5.2
gdata_3.0.1	generics_0.1.3	GGally_2.2.1
ggformula_0.12.0	ggplot2_3.5.2	ggpubr_0.6.0
ggrepel_0.9.6	ggridges_0.5.6	ggsci_3.2.0
ggsignif_0.6.4	ggstats_0.9.0	ggtext_0.1.2
glmnet_4.1-8	globals_0.17.0	glue_1.8.0
gmodels_2.19.1	goftest_1.2-3	googledrive_2.1.1
googlesheets4_1.1.1	gower_1.0.2	gplots_3.2.0
graphics_4.4.3	grDevices_4.4.3	grid_4.4.3
gridExtra_2.3	gridtext_0.1.5	grpreg_3.5.0
gt_1.0.0	gtable_0.3.6	gtools_3.9.5
hardhat_1.4.1	haven_2.5.4	here_1.0.1
highr_0.11	Hmisc_5.2-3	hms_1.1.3
htmlTable_2.4.3	htmltools_0.5.8.1	htmlwidgets_1.6.4
httpuv_1.6.16	httr_1.4.7	ids_1.0.1
insight_1.1.0	ipred_0.9-15	isoband_0.2.7
iterators_1.0.14	janitor_2.2.1	jomo_2.7-6
jpeg_0.1.11	jquerylib_0.1.4	jsonlite_2.0.0
juicyjuice_0.1.0	KernSmooth_2.23.26	km.ci_0.5-6
KMsurv_0.1-5	knitr_1.50	labeling_0.4.3
labelled_2.14.0	later_1.4.2	lattice_0.22-7
lava_1.8.1	leaps_3.2	lifecycle_1.0.4
listenv_0.9.1	litedown_0.7	lme4_1.1-37
lubridate_1.9.4	magrittr_2.0.3	markdown_2.0
MASS_7.3-65	Matrix_1.7-2	MatrixModels_0.5-4
maxstat_0.7.25	memoise_2.0.1	methods_4.4.3
mgcv_1.9.3	mice_3.17.0	microbenchmark_1.5.0
mime_0.13	minqa_1.2.8	mitml_0.4-5
mitools_2.4	modelbased_0.10.0	ModelMetrics_1.2.2.2
modelr_0.1.11	mosaic_1.9.1	mosaicCore_0.9.4.0
mosaicData_0.20.4	multcomp_1.4-28	munsell_0.5.1
mvtnorm_1.3-3	naniar_1.1.0	nlme_3.1-167
nloptr_2.2.1	nnet_7.3-20	norm_1.0.11.1
nortest_1.0-4	numDeriv_2016.8.1.1	olsrr_0.6.1
openssl_2.3.2	ordinal_2023.12.4.1	pan_1.9
parallel_4.4.3	parallelly_1.43.0	parameters_0.24.2
patchwork_1.3.0	pbkrtest_0.5.3	performance_0.13.0

pillar_1.10.2 pkgconfig_2.0.3 pls_2.8-5
plyr_1.8.9 png_0.1.8 polspline_1.1.25
polynom_1.4.1 prettyunits_1.2.0 pROC_1.18.5
processx_3.8.6 prodlim_2024.06.25 progress_1.2.3
progressr_0.15.1 promises_1.3.2 proxy_0.4.27
ps_1.9.1 pscl_1.5.9 purrr_1.0.4
pwr_1.3-0 quantreg_6.1 R6_2.6.1
ragg_1.4.0 rappdirs_0.3.3 rbibutils_2.3
RColorBrewer_1.1-3 Rcpp_1.0.14 RcppArmadillo_14.4.1.1
RcppEigen_0.3.4.0.2 Rdpack_2.6.4 reactable_0.4.4
reactR_0.6.1 readr_2.1.5 readxl_1.4.5
recipes_1.2.1 reformulas_0.4.0 rematch_2.0.0
rematch2_2.1.2 report_0.6.1 reprex_2.1.1
reshape2_1.4.4 rlang_1.1.6 rmarkdown_2.29
rms_8.0-0 ROCR_1.0-11 rpart_4.1.24
rprojroot_2.0.4 rsample_1.3.0 rstatix_0.7.2
rstudioapi_0.17.1 rvest_1.0.4 sandwich_3.1-1
sass_0.4.10 scales_1.3.0 see_0.11.0
selectr_0.4.2 shape_1.4.6.1 shiny_1.10.0
slider_0.3.2 snakecase_0.11.1 sourcetools_0.1.7.1
SparseM_1.84-2 sparsevctrs_0.3.3 splines_4.4.3
SQUAREM_2021.1 stats_4.4.3 stats4_4.4.3
stringi_1.8.7 stringr_1.5.1 survey_4.4-2
survival_3.8-3 survminer_0.5.0 survMisc_0.5.6
sys_3.4.3 systemfonts_1.2.2 tableone_0.13.2
textshaping_1.0.0 TH.data_1.1-3 tibble_3.2.1
tidyr_1.3.1 tidyselect_1.2.1 tidyverse_2.0.0
timechange_0.3.0 timeDate_4041.110 tinytex_0.57
tools_4.4.3 topmodels_0.3-0 tzdb_0.5.0
ucminf_1.2.2 UpSetR_1.4.0 utf8_1.2.4
utils_4.4.3 uuid_1.2.1 V8_6.0.3
vctrs_0.6.5 viridis_0.6.5 viridisLite_0.4.2
visdat_0.6.0 vroom_1.6.5 warp_0.2.1
withr_3.0.2 xfun_0.52 xml2_1.3.8
xplorerr_0.2.0 xtable_1.8-4 yaml_2.3.10
yardstick_1.3.2 zoo_1.8-14