432 Quiz 2 for Spring 2024

Thomas E. Love, Ph.D.

2024-04-17 12:32 pm

Quiz Instructions

Quiz 2 includes material from the first 24 classes in 432, including all of Jeff Leek's *How to be a Modern Scientist*.

All necessary Quiz 2 elements will appear by 5 PM on 2024-04-18 at https://github.com/THOMASELOVE/432-quizzes-2024/tree/main/quiz2. There, we link to:

- the Main Document (this **47** page pdf) containing the instructions and all **34** questions,
- the Google Form Answer Sheet, and
- the seven data sets we are providing.

This is an open book, open notes quiz. You are welcome to consult the materials on the course website and that we've read for class, but you are not allowed to discuss these questions with **anyone** other than Professor Love and the teaching assistants. To submit your Quiz, you will have to affirm that you have obeyed these rules.

0.1 Deadline is Tuesday 2024-04-23 at NOON.

The deadline to submit the Google Form Answer Sheet is **Tuesday 2024-04-23 at Noon**. All of your answers must be submitted through the Google Form Answer Sheet by the deadline, without exception. Please do not wait until the last moment to submit your work.

0.2 Footnotes are hints.

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

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

0.3 The Google Form Answer Sheet

The Google Form Answer Sheet, found at https://bit.ly/432-2024-quiz2-answer-form, is where you will provide your responses to all 34 questions, and a final affirmation that you followed the Quiz rules. You must be logged into Google via CWRU to access the Answer Sheet. After you complete the form and hit submit, you will receive an emailed copy of your submission, with a link to edit your results, or complete your work, if needed.

0.4 Writing Code into the Answer Sheet

We may ask you to provide R code in your response on the Answer Sheet. Do not include the `library` command at any time. Assume in all questions that all relevant packages have been loaded in R. R packages that Dr. Love used in building the Quiz and its answer sketch are listed in the last section of these Instructions.

0.5 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.

0.6 Scoring

Four of the 34 items (Items Q02, Q21, Q29 and Q30) are worth 4 points while the rest are worth 3 each, adding to a total of **106** points[2].

0.7 When Will I Know How I Did?

Dr. Love will grade the Quiz, and results (including an answer sketch) will be available by class time on Thursday 2024-04-25.

0.8 How Long Should Quiz 2 Take?

Quiz 2 should take 6-7 hours to complete. I expect most students will take 4-9 hours, and some will take as little as 3 or as many as 12. It is **not** a good idea to spend a long time on any one 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".

[2] A score of 90 on the Quiz (out of 106 points) will be treated as if it were a score of 90 points out of 100, so in a sense there are six *extra* points available.

0.9 Asking for Help

If you need clarification on a Quiz question, you have exactly one way of getting help:

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

During the Quiz period (2024-04-18 through 2024-04-23) we will not answer questions about Quiz 2 except through the email listed above. We promise to respond to all questions received by 9 AM on 2024-04-23.

0.10 Taking the Quiz

- 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.
- Note that 15 minutes per question (which should be more time than you need for most questions) for 34 questions yields 8.5 hours of total time on the Quiz.

0.11 Seven Data Sets We Have Provided for Quiz 2

You have links to the **seven** data sets listed below on the Quiz 2 page. Each of them should be useful to you.

File Name	Used in Items
`dataB.rds`	Q09 - Q10
`dataD.csv`	Q16 - Q23
`dataE.rds`	Q25 - Q30
`dataH.csv`	Q01 - Q03
`dataN.rds`	Q31 - Q34
`dataS.rds`	Q04 - Q06
`dataT.csv`	Q11 - Q12

0.12 Packages and Settings used by Dr. Love

All packages and settings I used in writing the Quiz and its answer sketch are included in the list below[3]. You are permitted to use other packages to complete the Quiz if you like, but you shouldn't need to do so.

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

library(bayestestR); library(bestglm)
library(car); library(caret); library(countreg); library(cutpointr)
library(Epi)
library(GGally); library(glmnet); library(gt)
library(insight)
library(janitor)
library(lme4)
library(MASS); library(mice); library(mosaic)
library(naniar); library(nnet)
library(patchwork); library(pROC); library(pscl)
library(quantreg)
library(ROCR); library(rstanarm)
library(survey); library(survival); library(survminer)
library(topmodels) ## students do not need this but I did

library(conflicted)
library(rms)
library(tidymodels)
library(tidyverse)

conflicts_prefer(dplyr::filter, dplyr::select, dplyr::summarize, dplyr::count,
                 base::mean, base::sum, base::max,
                 car::vif, Matrix::update, rms::Predict)

options(dplyr.summarise.inform = FALSE)

theme_set(theme_bw())
```

This concludes the Quiz 2 instructions. Good luck!

[3] I also listed some other packages that I did not use.

Setting Up Q01-Q03: the `dataH` data

The `dataH.csv` file provided to you describes results from the General Social Survey on the happiness level of 1515 respondents, as well as the person's level of schooling and number of siblings. We will use these data for items Q01 through Q03. The six variables contained in the `dataH.csv` data are:

Name	Description
`subject`	meaningless subject identifier (S0001 - S1515)
`happiness`	3 levels: (1) Not too happy, (2) Pretty happy, (3) Very happy
`siblings`	count of siblings (observed range is 0-10)
`sib_cat`	sibling category based on `siblings` in 5 levels: (1) 0 or 1, (2) 2 or 3, (3) 4 or 5, (4) 6 or 7, (5) 8+
`sch_years`	years of schooling (observed range is 3-22)
`sch_cat`	schooling category based on `sch_years` in 4 levels: (1) < 12, (2) 12, (3) 13-16, (4) 17+

In reading in the `dataH.csv` file into an R tibble, retain the order of the factors `happiness`, `sib_cat`, and `sch_cat` from low to high as listed above, but you should only specify the `happiness` variable as being an ordered factor in R.

1 Q01

Fit an appropriate logistic regression model, which I'll call model `m01`, to predict the ordinal category `happiness` using the main effects (only) of the *numerical* versions of the schooling and siblings variables[4]. Use all 1515 observations in `dataH` to build your model.

Use this `m01` model to predict the happiness levels for all 1515 subjects in the `dataH` tibble.

Item Q01 has three parts, worth one point each.

a. What is the percentage (rounded to one decimal place) of the 1515 subjects in `dataH` for which model `m01` predicts the correct happiness category?

b. How many of the 1515 subjects in `dataH` are predicted by model `m01` to be in the "Not too happy" category?

c. How many of the subjects who are actually "Very happy" had their happiness level correctly predicted by model `m01`?

[4] The numerical versions are `siblings` and `sch_years`.

2 Q02 (4 points)

Fit an appropriate logistic regression model, which I'll call model `m02`, to predict the ordinal category `happiness` using the *categorical* versions of the schooling and siblings variables[5], along with an interaction effect of the two predictors on happiness level. Use all 1515 observations in `dataH` to build your model.

Item Q02 has two parts, each worth 2 points.

a. Which of the two models fit so far (`m01` or `m02`) has the better AIC value? Does that model also have the better BIC?

b. What is the p value for the likelihood ratio test[6] of the interaction term in model `m02`? Round your answer to two decimal places.

3 Q03

Fit a multinomial regression model to predict `happiness` using the same predictors that you used in `m01`. 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 `m01`.
b. The BIC of the multinomial model is an improvement over model `m01`.
c. A test of the proportional odds assumption made in `m01` suggests we should be comfortable with that assumption.
d. The multinomial model requires the estimation of two additional parameters, compared to model `m01`.
e. None of these statements are true.

[5] The categorical versions are `sch_cat` and `sib_cat`.

[6] Hint: you'll need to compare your `m02` to another model to obtain this result.

Setting Up Q04-Q06: the dataS data

The dataS.Rds file provided to you contains information on seven variables for 600 subjects, who entered the study at the moment when they were admitted to an intensive care unit. We will use these data in Items Q04 - Q06. The variables are:

Variable	Description
ptid	Subject Identifying Code[7]
study_time	Time (in days) for which the subject was followed in the study
death	Did the subject die while the study was going on? (Yes or No)
age	Age (in years, with one decimal place) at study entry
aps1	APACHE III score (ignoring coma) at study entry (quantitative: 0 to 150)
hrt1	Heart rate (in beats per minute) at study entry
dnr1	Did the subject have a do-not-resuscitate order in place at study entry? (Yes or No)

Now, for Item Q04, you should create a survival object describing study time until death (or censoring) using the **dataS** tibble, and then create a Kaplan-Meier curve comparing this survival object for the two dnr1 groups.

The output shown in Displays 1 and 2 for Item Q04 below may be helpful to you.

Display 1 of 2 for Item Q04

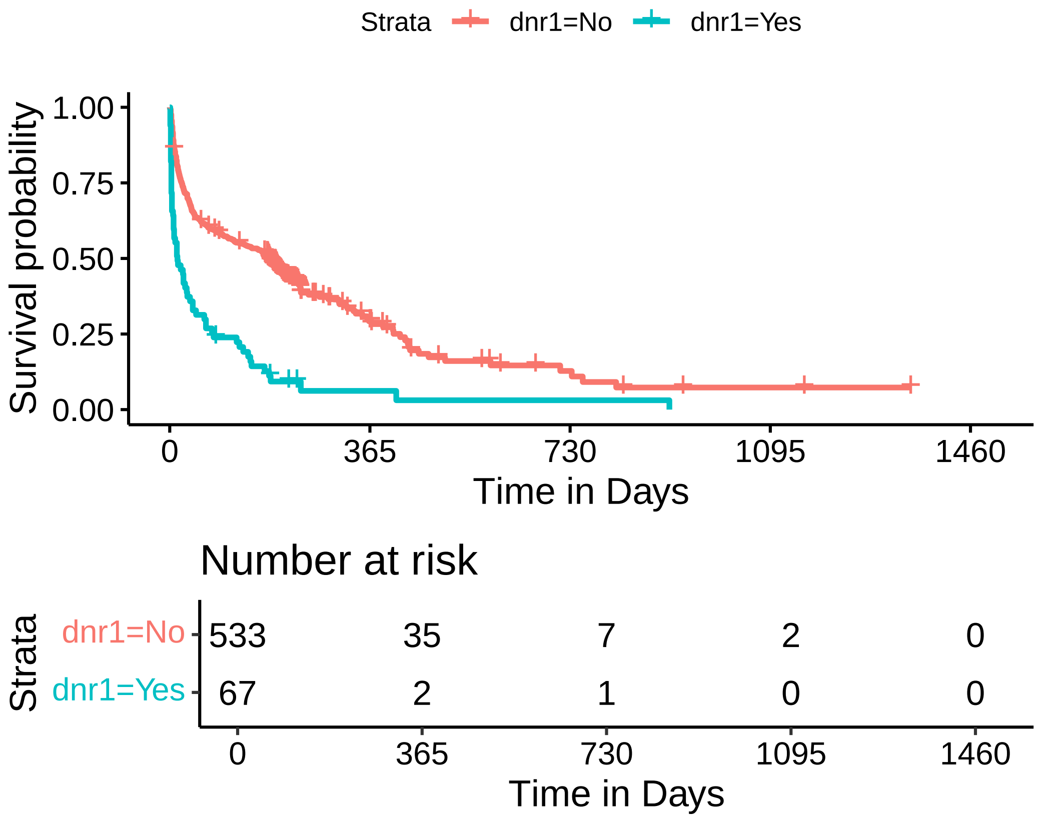
```
dataS <- read_rds("data/dataS.Rds")
dataS$surv04 <- Surv(time = dataS$study_time, event = dataS$death == "Yes")
kmfit04 <- survfit(dataS$surv04 ~ dataS$dnr1)
kmfit04
```

```
Call: survfit(formula = dataS$surv04 ~ dataS$dnr1)

                  n events median 0.95LCL 0.95UCL
dataS$dnr1=No   533    330    183     141     205
dataS$dnr1=Yes   67     63     14       7      37
```

[7] Note that the dataS.Rds file is sorted from the lowest ptid (9195) up to highest ptid (10278).

Display 2 of 2 for Item Q04

```
ggsurvplot(kmfit04, data = dataS, risk.table = TRUE,
           xlab = "Time in Days", break.time.by = 365,
           risk.table.height = 0.35)
```



```
survdiff(dataS$surv04 ~ dataS$dnr1)
```

```
Call:
survdiff(formula = dataS$surv04 ~ dataS$dnr1)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
dataS$dnr1=No   533      330      367      3.73      57.6
dataS$dnr1=Yes   67       63       26     52.55      57.6

 Chisq= 57.6  on 1 degrees of freedom, p= 3e-14
```

4 Q04

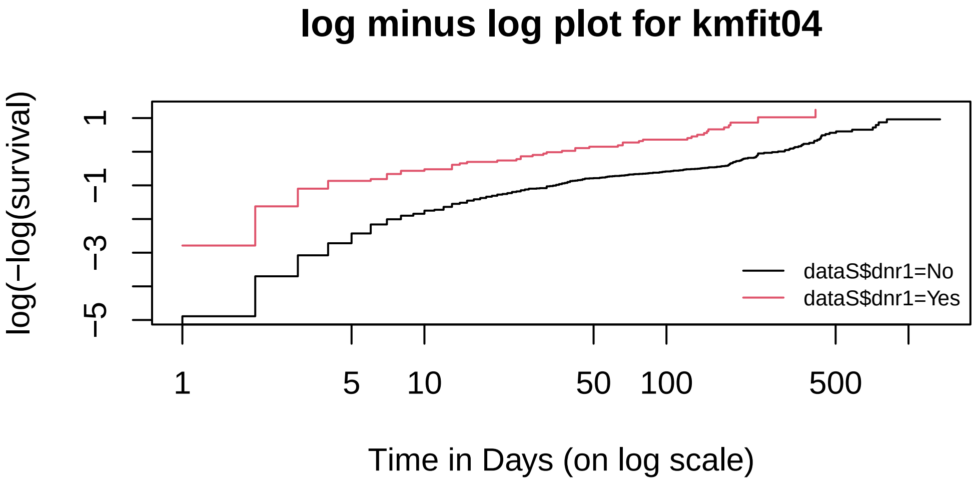
Consider the material on the previous two pages related to Q04, as well as any additional analyses you feel the need to run on the `dataS` tibble. Which of the following statements are true? More than one may be true.

CHOOSE EACH OF THE TRUE STATEMENTS.

a. The lowest numbered `ptid` for a censored subject in the `datS` data is 9195.
b. There are more subjects with a DNR order (i.e., `dnr1` = Yes) than without in the `datS` data.
c. The median survival time for subjects with a DNR order is higher than the median survival time for subjects without a DNR order.
d. The log rank test suggests that subjects without a DNR order and subjects with a DNR order have similar survival rates.
e. Neither of the groups (DNR or No DNR) contains more than 30 subjects who survived at least one year, according to the Kaplan-Meier curve.
f. None of the statements above are true.

5 Q05

Below I have provided a log minus log plot for the comparison in Item Q04. In a complete English sentence or two, what conclusion should you draw from this plot about our work in Item Q04?



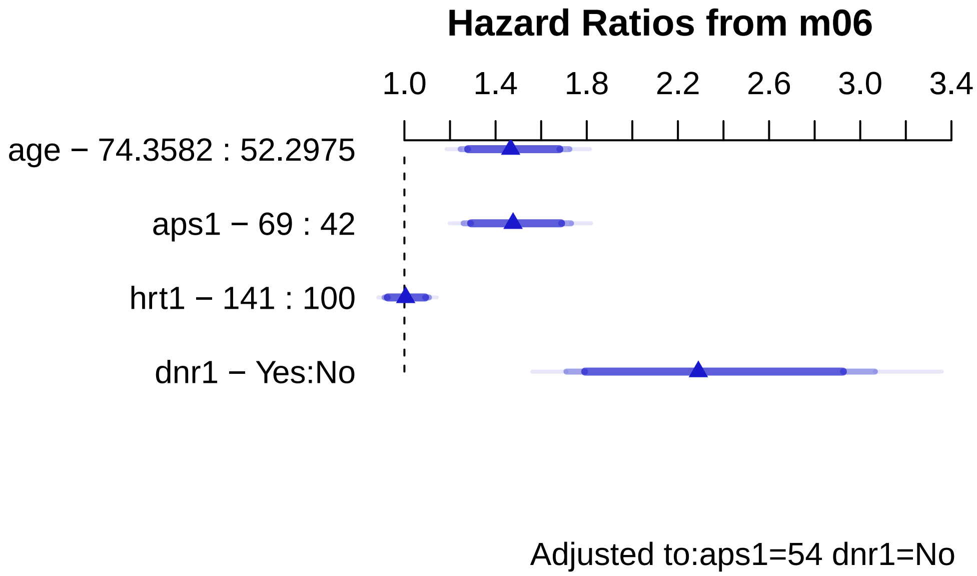
Setup for Q06

Fit a Cox proportional hazards model to predict the survival object from Item Q04 for all 600 subjects on the basis of four predictors: age, APACHE III score, heart rate and DNR status, including an interaction between DNR status and APACHE III score, and a restricted cubic spline with 3 knots in age.

Call that Cox model `m06`, although you'll need to figure out how to fit it, since I didn't show it in the output below. However, your result should match the effects summary shown below.

```
dataS <- read_rds("data/dataS.rds")

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

## model m06 fit here, hidden from you
```

```
plot(summary(m06), main = "Hazard Ratios from m06")
```



Additional output setting up Q06 is shown on the next page.

```
summary(m06)
```

```
             Effects              Response : Surv(study_time, death == "Yes") 

 Factor           Low     High    Diff.  Effect    S.E.     Lower 0.95 Upper 0.95
 age               52.297  74.358 22.061 0.3825800 0.082683  0.220520  0.54463   
  Hazard Ratio     52.297  74.358 22.061 1.4661000       NA  1.246700  1.72400   
 aps1              42.000  69.000 27.000 0.3898400 0.081081  0.230930  0.54876   
  Hazard Ratio     42.000  69.000 27.000 1.4767000       NA  1.259800  1.73110   
 hrt1             100.000 141.000 41.000 0.0054911 0.049446 -0.091422  0.10240   
  Hazard Ratio    100.000 141.000 41.000 1.0055000       NA  0.912630  1.10780   
 dnr1 - Yes:No      1.000   2.000     NA 0.8284300 0.148640  0.537100  1.11980   
  Hazard Ratio      1.000   2.000     NA 2.2897000       NA  1.711000  3.06410   

Adjusted to: aps1=54 dnr1=No  
```

6 Q06

Consider the information provided in the setup for Q06 shown above and on the previous page, along with any other analyses of the `dataS` 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 direction of the DNR effect is the same in this model as we observed in Q04.
b. The value of the APACHE score effect shown in this plot applies to subjects whose DNR status is Yes.
c. The value of the age effect shown in this plot applies to subjects whose DNR status is Yes.
d. The hazard ratio associated with a one beat-per-minute change in heart rate will be closer to 1 than the hazard ratio pictured in the plot above.
e. None of the statements above are true.

7 Q07

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.

8 Q08

Consider the Figure for Q08 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.

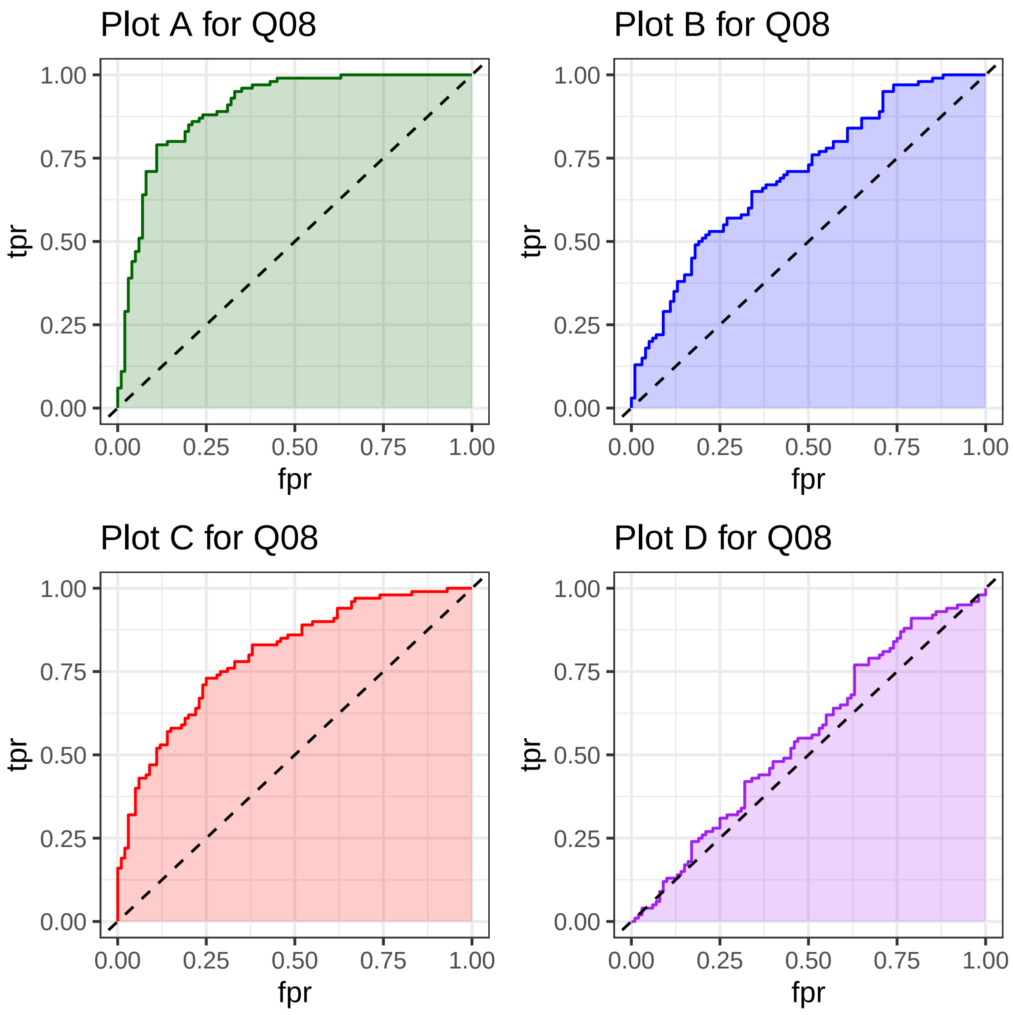
Rows:

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 Q08


Plot A for Q08

Plot B for Q08

Plot C for Q08

Plot D for Q08

This is the end of the output for Item Q08.

Setting Up Q09-Q10: the `dataB` data

A subset of the data from the BRFSS SMART study developed in Chapter 2 of the Course Notes are used in Q09 and Q10 with some modifications. The three variables of interest, collected in the `dataB.Rds` file provided to you 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)

Here's a table of the data contained in the `dataB` tibble.

```
dataB <- readRDS("data/dataB.Rds")
dataB |> count(mmsa, binge, vax_pneumo) |> gt()
```

mmsa	binge	vax_pneumo	n
CIN	No	NoVax	617
CIN	No	Vax	587
CIN	Yes	NoVax	171
CIN	Yes	Vax	70
CLE	No	NoVax	371
CLE	No	Vax	442
CLE	Yes	NoVax	96
CLE	Yes	Vax	50
COL	No	NoVax	706
COL	No	Vax	762
COL	Yes	NoVax	156
COL	Yes	Vax	67
DAY	No	NoVax	212
DAY	No	Vax	231
DAY	Yes	NoVax	30
DAY	Yes	Vax	25

In all, there are 4,593 observations in the `dataB` tibble.

9 Q09

I used the `dataB` tibble described on the previous page to fit the following set of models to predict `mmsa` based on various combinations of the two predictors `binge` and `vax_pneumo`.

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

m09_1 <- multinom(mmsa ~ 1, data = dataB, trace = FALSE)
m09_B <- multinom(mmsa ~ binge, data = dataB, trace = FALSE)
m09_V <- multinom(mmsa ~ vax_pneumo, data = dataB, trace = FALSE)
m09_BV <- multinom(mmsa ~ binge + vax_pneumo, data = dataB, trace = FALSE)
m09_SAT <- multinom(mmsa ~ binge * vax_pneumo, data = dataB, trace = FALSE)
```

Here are some summary results[8] for these models:

			CLE	COL	DAY	Residual		
Model	df	nobs	intercept	intercept	intercept	deviance	AIC	BIC
`m09_1`	3	4593	-0.410	0.157	-1.065	11938.56	11944.6	11963.9
`m09_B`	6	4593	-0.393	0.198	-0.276	11925.27	11937.3	11975.9
`m09_V`	6	4593	-0.523	0.090	-1.180	11928.38	11940.4	11979.0
`m09_BV`	9	4593	-0.510	0.139	-1.101	11916.83	11934.8	11992.7
`m09_SAT`	12	4593	-0.509	0.135	-1.068	11912.76	11936.8	12014.0

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 Columbus.
c. The model which has the most negative intercept for Dayton.
d. The model with the second largest Bayes Information Criterion.
e. The model including the interaction of `binge` and `vax_pneumo`.

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

10 Q10

Consider the five displays for Q10 shown below this item. These five displays describe the five models we built in Q09, with each display corresponding to a different model. Which display describes the preferred model that you identified back in Q09?

a. Display A for Q10
b. Display B for Q10
c. Display C for Q10
d. Display D for Q10
e. Display E for Q10
f. It is impossible to tell which Display is correct.

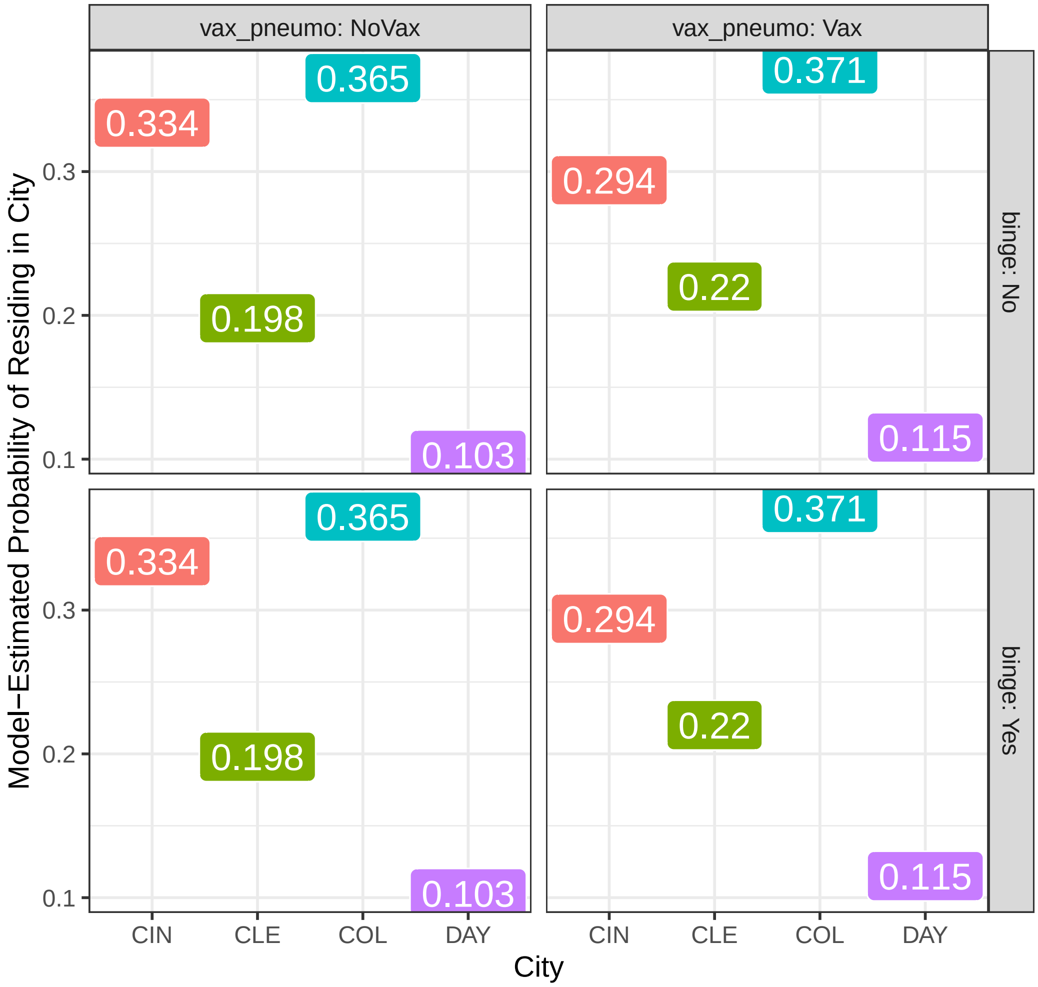
Display A for Q10

```
tidy(displayA, exponentiate = TRUE, conf.int = TRUE) |>
  gt() |> fmt_number(decimals = 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 B for Q10

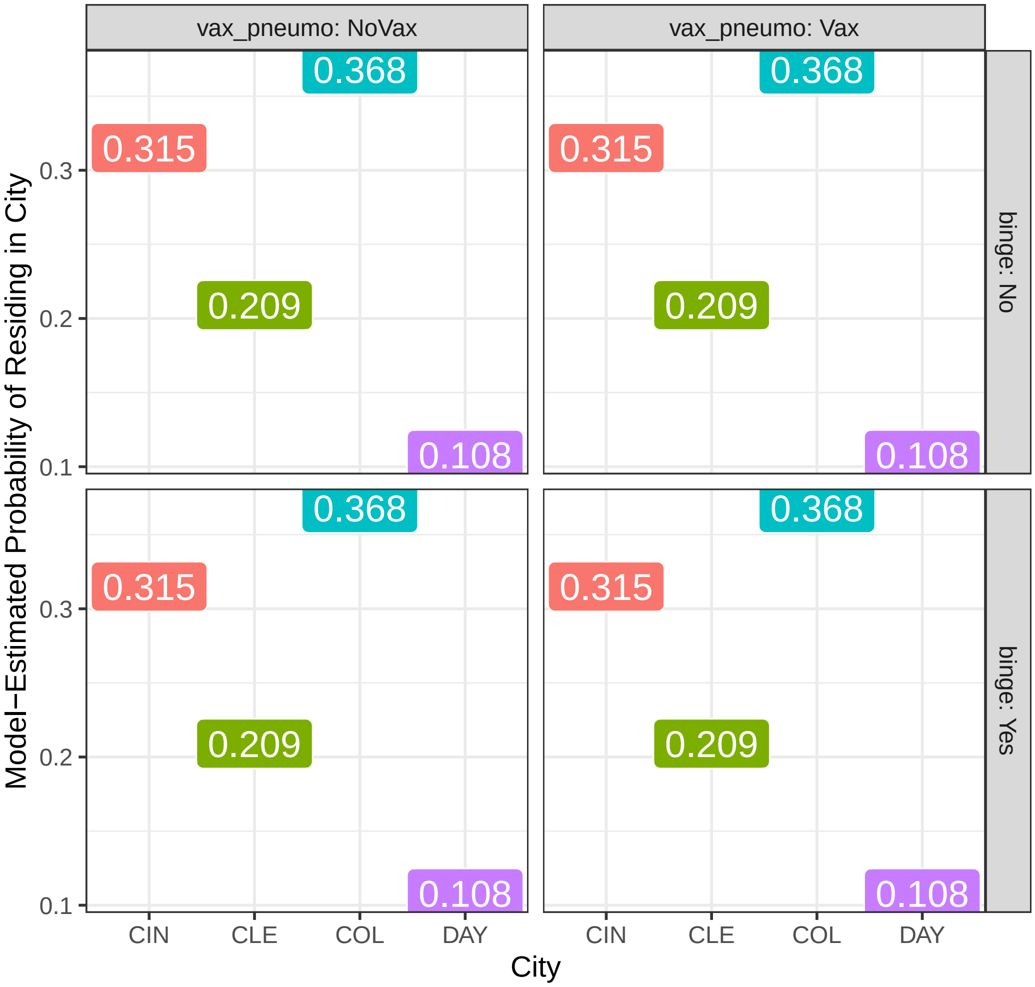
```
ggplot(displayB, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(binge ~ vax_pneumo, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Residing in City", x = "City")
```



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 C for Q10

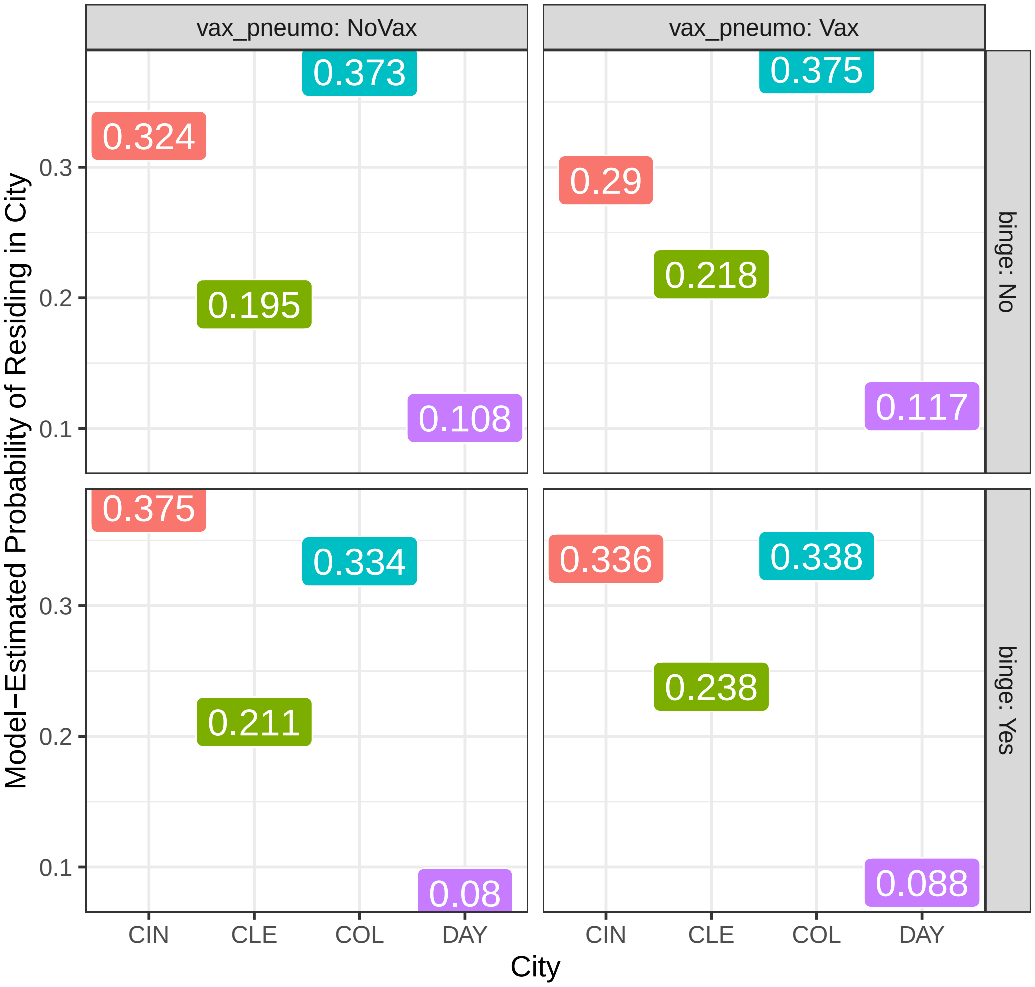
```r
ggplot(displayC, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(binge ~ vax_pneumo, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Residing in City", x = "City")
```



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 Q10

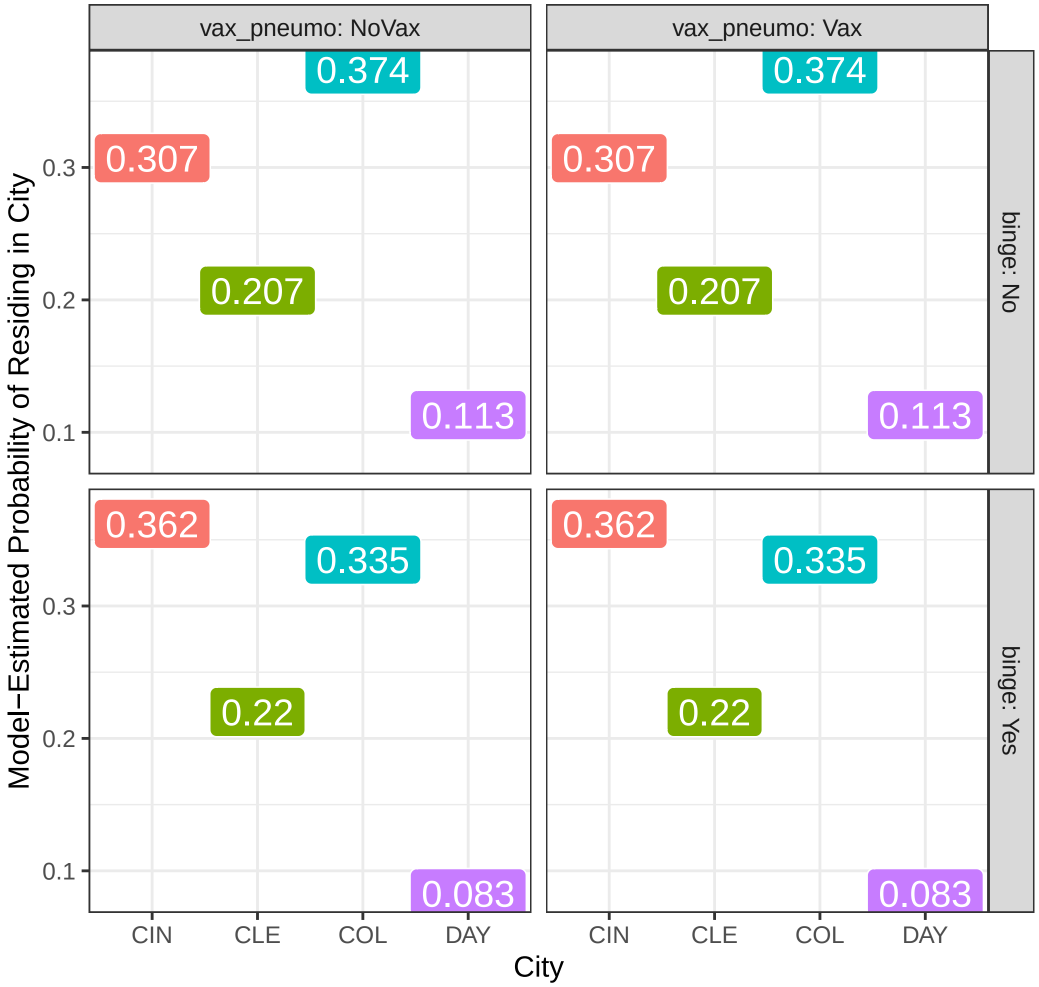
```
ggplot(displayD, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  guides(fill = "none") +
  facet_grid(binge ~ vax_pneumo, labeller = "label_both") +
  labs(y = "Model-Estimated Probability of Residing in City", x = "City")
```



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 Q10

```
ggplot(displayE, aes(x = city, y = prob, fill = city)) +
  geom_label(aes(label = prob), col = "white", size = 5) +
  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.

This is the end of the output for Item Q10.

Setting Up Q11-Q12: the `dataT` data

The `dataT.csv` file provided to you will be used in Q11 and Q12. Create a tibble that contains the information in the `dataT.csv` data and name it `dataT`.

The outcome of interest in the resulting `dataT` tibble, labeled `score`, is the number of standards (out of 25) met by subjects involved in an drug treatment program. Subjects are eligible for release from the program when they meet at least twenty of the 25 standards. The data in `score` describe the number of standards met after one week of treatment for 360 recent subjects.

Measures `entry`, `group` and `strength` are predictors of `score`, whose main effects (only) are of interest to us. `entry` and `strength` are quantitative measures, and `group` indicates whether or not the subject has completed a specific group of tasks. On the next two pages of this Quiz, I show the fit of a Poisson regression model to these data, and then show the fit of a negative binomial regression model to these data, in each case using only the main effects of the three predictors.

11 Q11

Consider the following three statements.

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 more scores of 1, 2 and 3 than we actually observed.

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

In light of the modeling results shown in Displays 1, 2 and 3 for Q11 on the next two pages, which of the above 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.

Q11 Display 1: Regression Models for the `dataT` data

```r
dataT <- read_csv("data/dataT.csv", show_col_types = FALSE)

m11_p <- glm(score ~ entry + group + strength, family = poisson(), data = dataT)
m11_nb <- glm.nb(score ~ entry + group + strength, link = log, data = dataT)

tidy(m11_p, exponentiate = TRUE, conf.int = TRUE, conf.level = 0.95) |>
  gt() |> fmt_number(decimals = 4)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.7931	0.2275	4.5146	0.0000	1.7856	4.3563
entry	1.0421	0.0021	19.7849	0.0000	1.0378	1.0463
group	1.1795	0.0570	2.8943	0.0038	1.0541	1.3183
strength	0.8629	0.0183	−8.0528	0.0000	0.8324	0.8944

```r
tidy(m11_nb, exponentiate = TRUE, conf.int = TRUE, conf.level = 0.95) |>
  gt() |> fmt_number(decimals = 4)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.8830	0.3499	3.0257	0.0025	1.4431	5.7555
entry	1.0459	0.0031	14.2658	0.0000	1.0393	1.0525
group	1.0892	0.0904	0.9449	0.3447	0.9121	1.3001
strength	0.8478	0.0291	−5.6727	0.0000	0.7999	0.8980

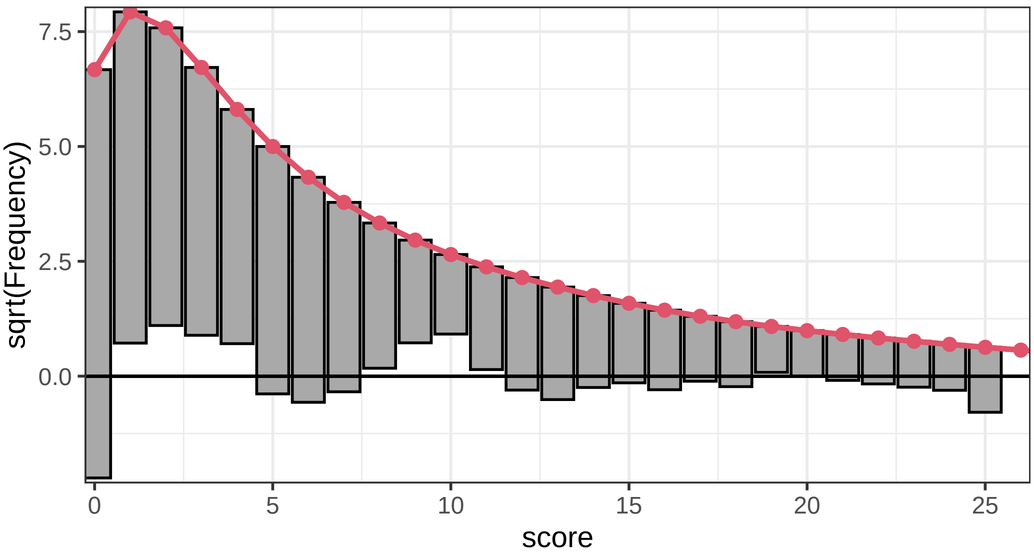
```r
glance(m11_p) |> gt()
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
1765.416	359	-812.0994	1632.199	1647.743	720.5416	356	360

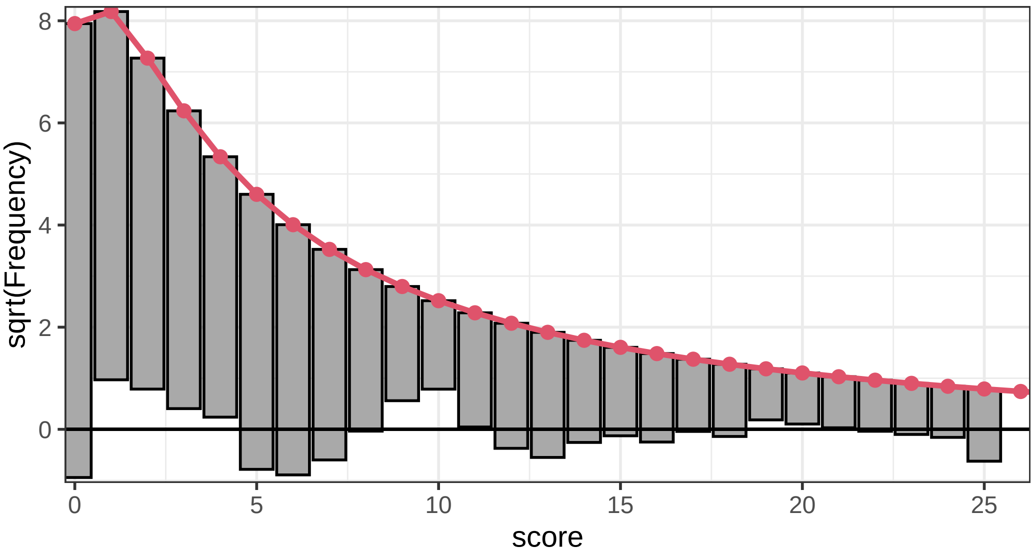
```r
glance(m11_nb) |> gt()
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
941.2164	359	-760.1901	1530.38	1549.811	409.8992	356	360

Q11 Display 2: A rootogram for the `m11_p` model



Q11 Display 3: A rootogram for the `m11_nb` model



This is the end of the output for Item Q11.

12 Q12

Fit the Poisson regression model (m11_p) that I fit for Item Q11, then use it to make a prediction for score for the three new subjects (named Amy, Bart and Chris) listed below.

Name	entry	group	strength
Amy	25	1	7.2
Bart	22	0	3.5
Chris	18	0	8.2

Which of the three new subjects (Amy, Bart or Chris) has the highest predicted score according to the m11_p model, and what is their predicted score? Please round your predicted score to zero decimal places.

13 Q13

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 at any time 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?

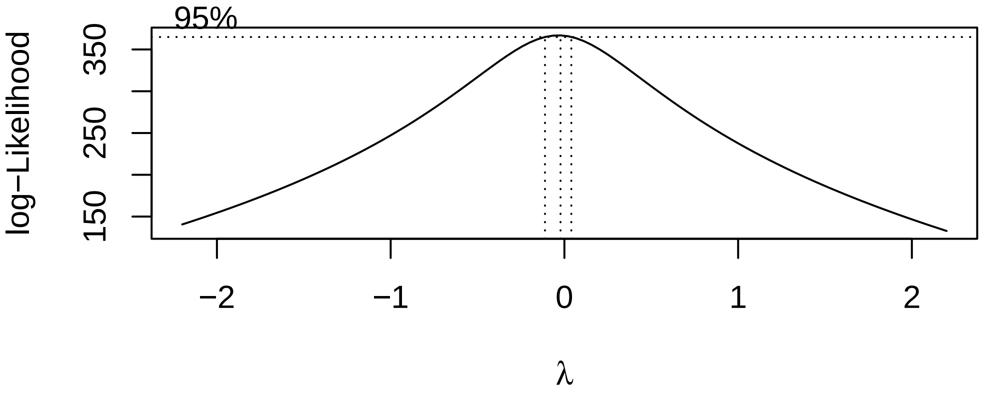
a. A multinomial logit model.
b. An ordinary least squares model.
c. A binary logistic regression model.
d. A Cox proportional hazards model.
e. None of these models would be appropriate.

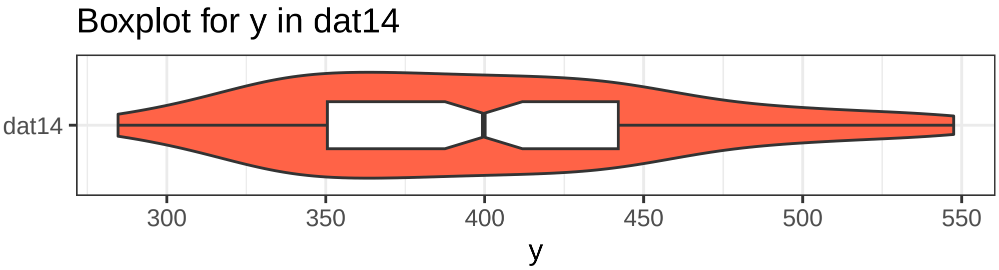
Setup for Q14

We are considering whether an outcome variable (y) should be transformed prior to fitting a model including main effects of three predictors called x1, x2 and x3 with 141 observations of data in the dat14 tibble. I have not provided dat14 to you. I have, however, provided a pair of figures for Q14 to you in the **Display for Q14**, which you'll find on top of the next page.

Display for Q14

```
par(mar = c(5, 4, 2, 2))
boxcox(y ~ x1 + x2 + x3, data = dat14, lambda = seq(-2.2, 2.2, 1/10))
```





14 Q14

Review the plots shown in the Display for Q14. Which of the following transformations of the outcome y is most appropriate?

a. Use the raw outcome, y, in the model.
b. Use the square root of y.
c. Use the logarithm of y.
d. Use the inverse of y.
e. Use the square of y.

15 Q15

In Q15, we focus on a tobacco cessation study 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 `dat15` tibble are shown in Display 2 for Q15, on the next page, but note that I have not provided the data in `data15` to you.

Suppose you want to add a survival object called `S` to the `dat15` 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 in Display 1 for Q15 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.

Display 1 for Q15

Chunk I

```
dat15$S = Surv(time = dat15$exitday - dat15$startday,
               event = dat15$exitreason %in% c("lost", "studyend"))
survdiff(S ~ treatment, data = dat15)
```

Chunk II

```
dat15$S <- Surv(time = dat15$exitday, event = dat15$exitreason)
survdiff(S ~ treatment, data = dat15)
```

Chunk III

```
dat15$S = Surv(time = dat15$exitday - dat15$startday,
                event = dat15$exitreason == "achieved")
survdiff(S ~ treatment, data = dat15)
```

Display 2 for Q15

```
dat15
```

```
# 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
```

```
dat15 |> 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
```

```
dat15 |> df_stats(~ startday + exitday) |>
  gt() |> fmt_number(min:sd, decimals = 2)
```

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

This is the end of the output for Item Q15.

Setting Up Q16-Q23: the `dataD` data

In Items Q16-Q23, we're interested in determining risk factors for high blood pressure (hypertension) using data from a sample of 1,111 women with diabetes. The data are provided to you in the `dataD.csv` file. Here's a glimpse.

```
dataD <- read_csv("data/dataD.csv", show_col_types = FALSE) |>
  mutate(smoke = as.factor(smoke))

glimpse(dataD)
```

```
Rows: 1,111
Columns: 4
$ ptid  <chr> "pt_0001", "pt_0002", "pt_0003", "pt_0004", "pt_0005", "pt_0006"~
$ hbp   <dbl> 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1~
$ age   <dbl> 27, 48, 39, 59, 43, 53, 59, 65, 65, 54, 46, 32, 49, 40, 42, 23, ~
$ smoke <fct> Yes, Yes, Yes, No, No, Yes, Yes, No, No, Yes, No, Yes, Yes, No, ~
```

```
summary(dataD)
```

```
     ptid                hbp              age         smoke
 Length:1111        Min.   :0.0000   Min.   :19.00   No :829
 Class :character   1st Qu.:0.0000   1st Qu.:54.00   Yes:282
 Mode  :character   Median :1.0000   Median :61.00
                    Mean   :0.7192   Mean   :59.82
                    3rd Qu.:1.0000   3rd Qu.:68.00
                    Max.   :1.0000   Max.   :75.00
```

- The `hbp` variable is 1 for patients with high blood pressure, and 0 otherwise.
- The `smoke` variable is Yes for smokers and No for non-smokers.
- As you can see, the median `age` across all subjects is 61 years, and the `age` values range from 19 to 75.

Fit an appropriate model (we'll call it Model X) to predict the log odds of high blood pressure on the basis of whether the patient smokes, the patients' age, and the interaction of `smoke` and `age`. Do not center or otherwise transform the `age` variable, and do not include any non-linear terms other than the specified interaction.

Later (in Item Q20), we'll fit an additional model (which we'll call Model Y) to the same 1111 women with diabetes.

Use this work and the `dataD` tibble to help you address Items Q16 through Q23.

16 Q16

According to Model X, what is a 57-year old non-smoker's predicted probability (rounded to two decimal places) of having high blood pressure?

17 Q17

Estimate the odds ratio comparing a 61-year old smoker to a 61-year old non-smoker, based on Model X. Provide both a point estimate and a 95% confidence interval. Round your final answers to two decimal places.

Hint: it's useful to first consider what the median age is in these data.

18 Q18

The results contained in the Display for Q18 show the result of using a bootstrap validation approach to estimate several summary statistics using Model X. What is the C statistic (area under the ROC curve; round your answer to 3 decimal places) that this output estimates would result from using Model X for prediction in new, but similar data to the sample used to fit the model?

Display for Q18

```
set.seed(43218); validate(modelX, B = 100)
```

```
          index.orig training   test optimism index.corrected   n
Dxy           0.2865   0.2883 0.2842   0.0041          0.2824 100
R2            0.0972   0.0993 0.0949   0.0045          0.0927 100
Intercept     0.0000   0.0000 0.0127  -0.0127          0.0127 100
Slope         1.0000   1.0000 0.9913   0.0087          0.9913 100
Emax          0.0000   0.0000 0.0042   0.0042          0.0042 100
D             0.0690   0.0708 0.0673   0.0035          0.0655 100
U            -0.0018  -0.0018 0.0000  -0.0018          0.0000 100
Q             0.0708   0.0726 0.0673   0.0053          0.0655 100
B             0.1867   0.1869 0.1874  -0.0006          0.1873 100
g             0.6454   0.6509 0.6361   0.0148          0.6306 100
gp            0.1277   0.1285 0.1259   0.0026          0.1251 100
```

19 Q19

Which of the following statements about Model X are true? More than one might be true.

CHECK ALL OF THE TRUE STATEMENTS.

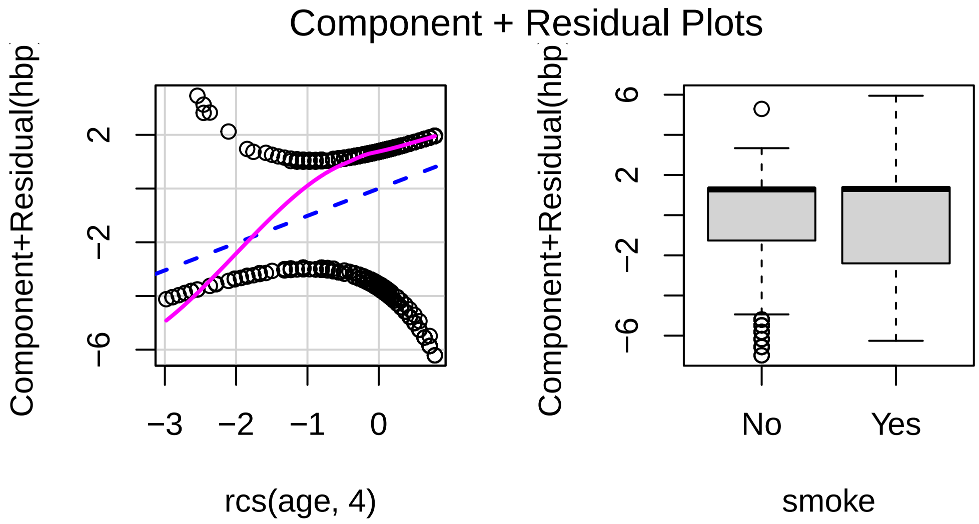
a. Cook's distance indicates a problem with the assumption of no extreme outliers.
b. The Brier score for Model X indicates an improvement over a model for the same outcome with a Brier score of 0.20.
c. Model X yields a p value smaller than 0.10 for the age-smoke interaction.
d. None of statements `a` through `c` are true.

20 Q20

Now consider a second model for these data, which we'll call Model Y. Model Y also uses `age` and `smoke` as predictors, but it does not include an interaction term, and instead includes a restricted cubic spline in `age` with 4 knots.

In a complete English sentence or two, please tell us what the set of partial residual plots[9] for Model Y shown below indicates about the assumption of a linear relationship between the spline in `age` and the log odds of having high blood pressure.

```
crPlots(modelY)
```



[9] These are also sometimes called Component + Residual plots.

21 Q21 (4 points)

Identify the optimal decision rule's cut point for fitted values that maximizes the sum of sensitivity and specificity for Model Y's confusion matrix applied to the subjects in `dataD`. Summarize your confusion matrix using the decision rule incorporating your cut point **after** you round your cut point to exactly two decimal places.

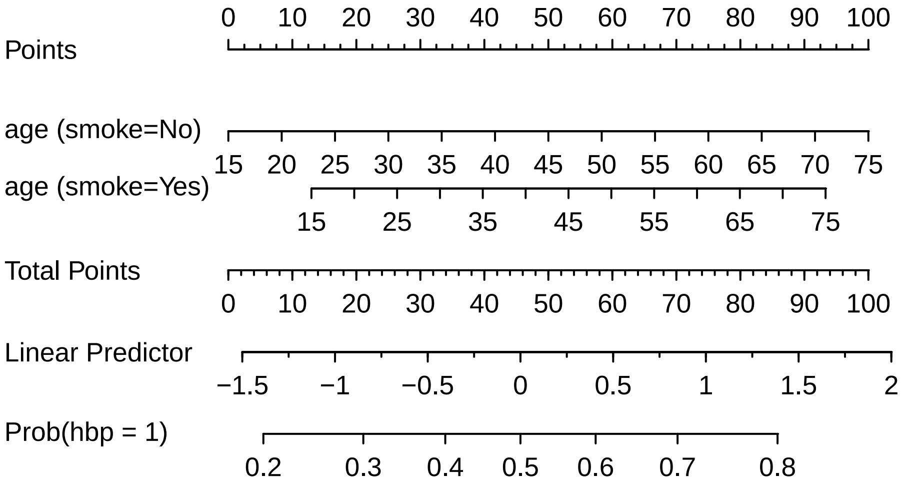
Item Q21 has two parts, each worth 2 points.

a. What is the cut point used in your optimal decision rule? Express your response as a proportion, rounded to two decimal places.

b. What is the positive predictive value exhibited by your confusion matrix with this decision rule? Express this response as a percentage, rounded to one decimal place.

22 Q22

Which model (X or Y) yields the nomogram shown below? In a complete English sentence or two, tell us how you know[10].



[10] Hint: You should be able to answer this question without running any R code.

23 Q23

Which model (Model X or Model Y) shows the better predictive ability based on each of the following summaries?

Note: Make your decisions for Q23 based on raw, unvalidated summaries across all 1111 observations in the `dataD` tibble.

Rows:

a. The Brier score
b. The Nagelkerke R^2
c. The Bayes Information Criterion

Columns:

1. Model X looks better
2. Model Y looks better
3. Neither model looks better

24 Q24

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.

Setting Up Q25-Q30: the `dataE` data

The `dataE` tibble used in Items Q25 through Q30 contains 432 observations on the eight variables tabulated below.

Name	Description
`person`	Identifier of the subject
`charge_k`	total charges, in 1000s of US dollars, in the past year (our outcome[11])
`age`	subject's age in years
`sex`	subject's sex (female/male)
`bmi`	subject's body mass index in kg/m^2
`kids`	# of children covered by subject's insurance (0-5)
`smoke`	whether the subject smokes tobacco (yes/no)
`region`	where the subject resides (4 categories)

Note that the four `region` categories are

- NE = northeast, NW = northwest, SE = southeast, SW = southwest

If you explore the `dataE.Rds` file we provided to you, you'll note (as we demonstrate below) that it includes some missing values.

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

dataE |> slice(8:12) # rows 8-12 of the dataE tibble
```

```
# A tibble: 5 x 8
  person charge_k   age sex      bmi  kids smoke region
  <chr>     <dbl> <dbl> <fct>  <dbl> <dbl> <fct> <fct>
1 A0029      2.78    23 male    17.4    NA no    NW
2 A0042      4.95    31 female  36.6     2 no    SE
3 A0048      3.56    28 female  34.8     0 no    NW
4 A0051      2.21    18 female  NA       0 no    NE
5 A0052      3.58    21 female  33.6     2 <NA>  NW
```

```
prop_miss_case(dataE)
```

```
[1] 0.1458333
```

[11] The outcome of interest to us is total health-related charges billed to an insurance company for the subject, which is gathered in the `charge_k` variable in thousands of dollars, so `charge_k` = 2.78 means the total charges for this subject in the past year were $2,780.

Using mice to create a singly imputed tibble called dataE_s

We will use the mice package to create a set of 20 imputations for the dataE missing values after setting a seed of 25. Then we will form a new data set, called dataE_s (standing for a single imputation of dataE), which contains the 17th of those 20 imputations, as follows...

```
## first we remove the subject ID codes from dataE

dataE_noid <- dataE |> select(-person)

## we next set our seed then perform 20 imputations

set.seed(25)
dataE_20imps <- mice(dataE_noid, m = 20, printFlag = FALSE)
```

```
## use the 17th of those imputations to form dataE_s

dataE_s <- complete(dataE_20imps, 17) |> tibble()

## add back in the subject ID codes to finish dataE_s

dataE_s$person <- dataE$person

## dataE_s should have 432 rows, 8 columns, no missing values
## and these subjects are in rows 8-12...

dim(dataE_s); n_miss(dataE_s); dataE_s |> slice(8:12)
```

```
[1] 432   8
```

```
[1] 0
```

```
# A tibble: 5 x 8
  charge_k   age sex      bmi  kids smoke region person
     <dbl> <dbl> <fct>  <dbl> <dbl> <fct> <fct>  <chr>
1     2.78    23 male    17.4     0 no    NW     A0029
2     4.95    31 female  36.6     2 no    SE     A0042
3     3.56    28 female  34.8     0 no    NW     A0048
4     2.21    18 female  30.2     0 no    NE     A0051
5     3.58    21 female  33.6     2 no    NW     A0052
```

25 Q25

We're going to transform our outcome when building regression models in this work, so add a variable called `logcharges` to your `dataE_s` tibble which contains the natural logarithm of `charge_k`.

Now, using this `dataE_s` tibble, build a linear model (using the `lm()` function) for the natural logarithm of `charge_k` using the main effects (and only the main effects) of the six predictors: `age`, `sex`, `bmi`, `kids`, `smoke` and `region`. Call this model `m25`.

Having run the model `m25`, use it to predict the total charges, in thousands of dollars, for two new subjects (Alice and Jacob) whose information on our predictors is tabulated below.

Subject	age	sex	bmi	kids	smoke	region
Alice	39	female	31.2	2	no	NE
Jacob	44	male	34.5	0	yes	SW

a. Which of the two new subjects, Alice or Jacob, has the higher estimated **total charges** estimated by model `m25`?

b. What are the estimated **total charges** for the subject you identified in Q25a? Report your answer in **dollars**, rounded to the nearest dollar.

26 Q26

I developed the output shown in the Displays for Q26 below using the linear model `m25` we fit in Q25. In light of this output, which of the conclusions below are appropriate?

CHECK ALL OF THE APPROPRIATE CONCLUSIONS.

a. There is a serious problem with collinearity in `m25`.
b. There is a serious problem with the assumption of Normality in `m25`.
c. There is a serious problem with the assumption of non-constant variance in `m25`.
d. Subject `A1096` has a problematic level of influence over `m25`.
e. None of the conclusions above are appropriate.

The Display for Q26 is found on the next **three** pages of this PDF.

Display for Q26

```
vif(m25)
```

```
           GVIF Df GVIF^(1/(2*Df))
age    1.027394  1        1.013604
sex    1.023715  1        1.011788
bmi    1.158817  1        1.076483
kids   1.015720  1        1.007829
smoke  1.010331  1        1.005152
region 1.172455  3        1.026871
```

```
outlierTest(m25)
```

```
    rstudent unadjusted p-value Bonferroni p
174 4.571336         6.3752e-06    0.0027541
128 4.229018         2.8802e-05    0.0124420
```

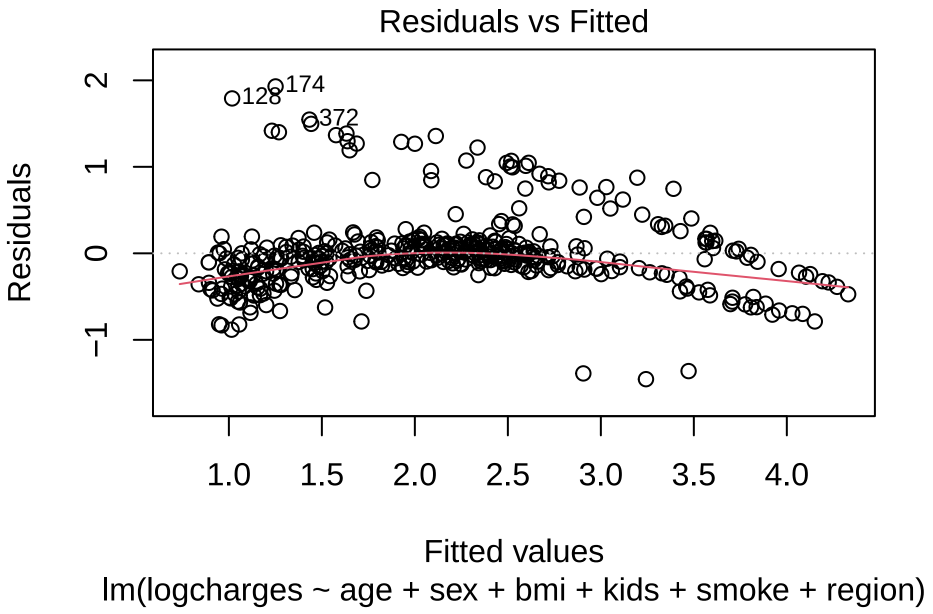
```
res25 <- augment(m25) |> mutate(person = dataE$person) |>
  slice(c(174, 128, 361))

res25 |> select(1:7, 14) |> gt()
```

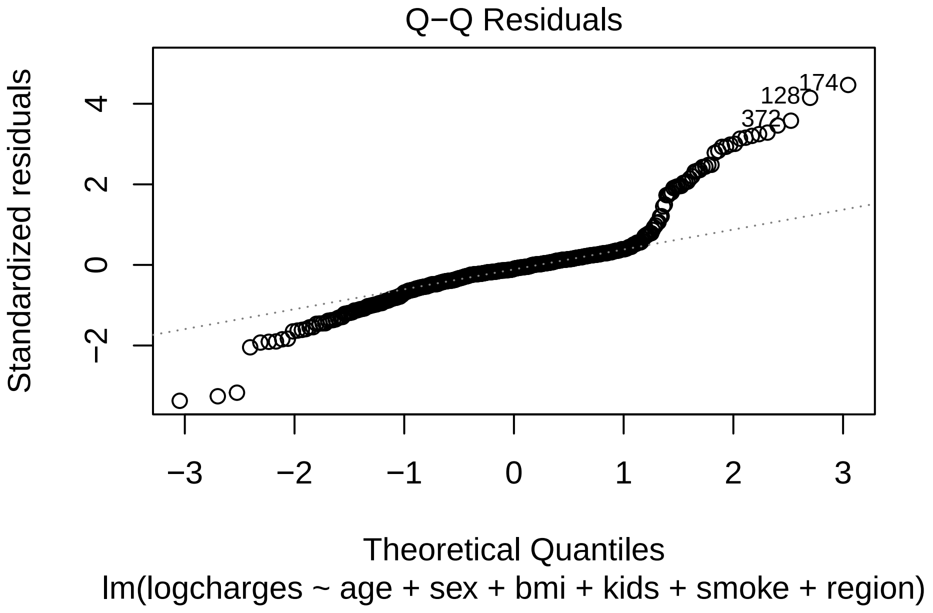
logcharges	age	sex	bmi	kids	smoke	region	person
3.180551	19	female	30.59	2	no	NW	A0527
2.808559	21	male	31.02	0	no	SE	A0398
1.517542	18	female	31.35	4	yes	NE	A1096

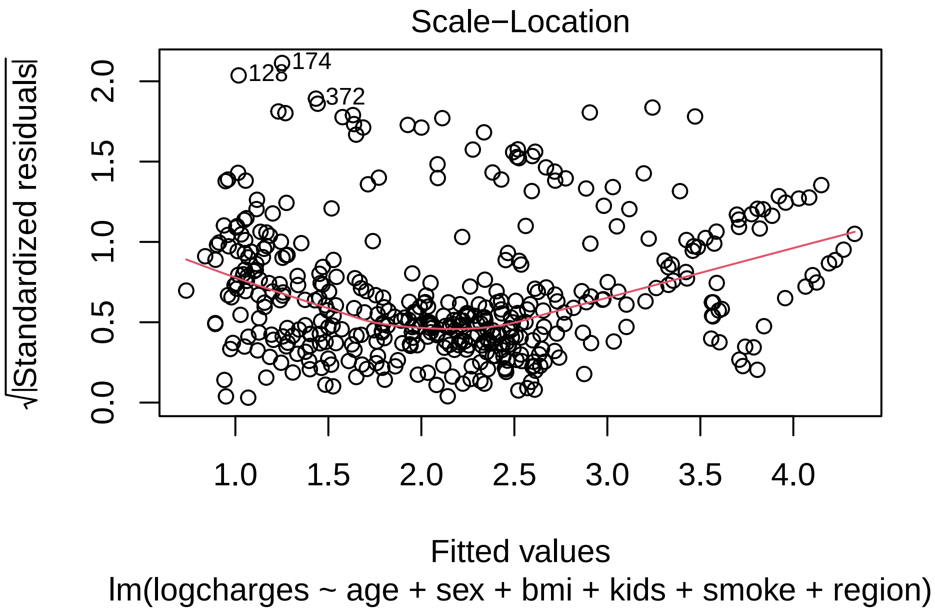
```
res25 |> select(8:14) |> gt() |> fmt_number(decimals = 3)
```

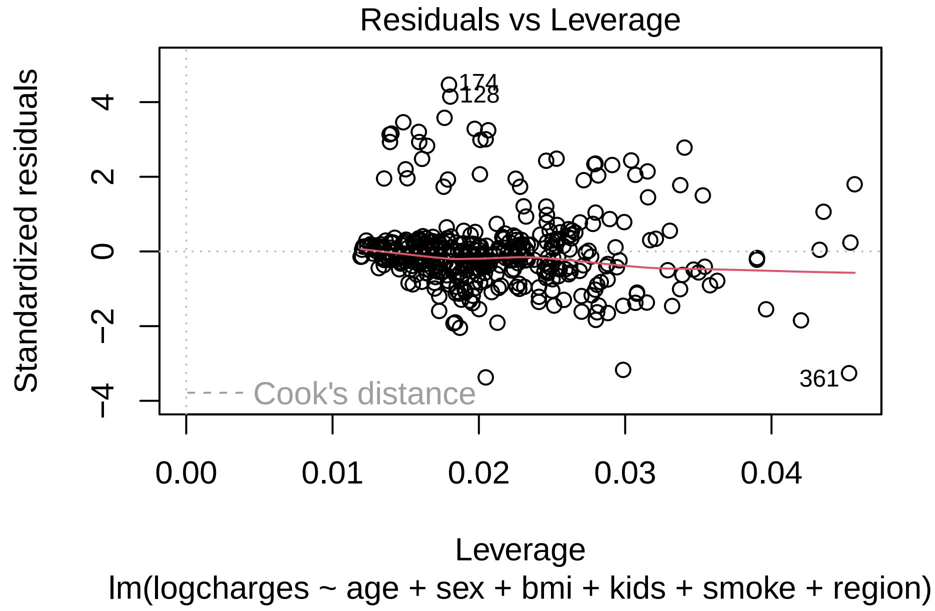
.fitted	.resid	.hat	.sigma	.cooksd	.std.resid	person
1.251	1.929	0.018	0.426	0.041	4.467	A0527
1.018	1.791	0.018	0.427	0.035	4.147	A0398
2.906	−1.388	0.045	0.431	0.056	−3.260	A1096

The Display for Q26 continues with the set of four residual plots (obtained using `plot(m25)`) found on the next two pages of this PDF.


Residuals vs Fitted
Residuals
2
1
0
−1
128
174
372
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(logcharges ~ age + sex + bmi + kids + smoke + region)


Q−Q Residuals
Standardized residuals
4
2
0
−2
174
128
372
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(logcharges ~ age + sex + bmi + kids + smoke + region)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
128
174
372
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(logcharges ~ age + sex + bmi + kids + smoke + region)


Residuals vs Leverage
Standardized residuals
−4
−2
0
2
4
174
128
361
- - - Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(logcharges ~ age + sex + bmi + kids + smoke + region)

27 Q27

For this item, we will again start with the singly imputed data contained in the dataE_s tibble, but we will be using the LASSO approach[12] to develop a new linear regression model.

To set up this work, we will first create our data matrix for the predictors, and a matrix of the outcomes, as follows (using the model m25 that you built in Item Q25.)

```
pred_x <- model.matrix(m25)
out_y <- dataE_s |> select(logcharges) |> as.matrix()
```

Having done that, please use the LASSO approach to fit a model for your transformed outcome, with the following specifications:

- First set a random seed of 273, please.
- In the cross-validation step, use type.measure = "mse" and nfolds = 10.
- As for the fitting step, because this is a LASSO model, you'll want to set alpha = 1.
- Use as your lambda value in the fitting step the minimum lambda value obtained in the cross-validation step.
- Call the LASSO model m27.

Which predictors (of those included in m25) remain in the m27 model for the logarithm of total charges in dataE_s?

CHECK EACH PREDICTOR THAT IS INCLUDED IN MODEL m27.

a. age
b. sex
c. bmi
d. kids
e. smoke
f. region
g. None of these predictors remain in model m27.

[12] Hint: The Class 11 slides and Chapter 32 in the Course Notes should help.

Setup for Item Q28, part 1 of 2

The output shown below and on the next page demonstrates the fit of a model I'll call `m28`.

```r
set.seed(28)
m28 <- stan_glm(logcharges ~ age + sex + bmi + kids + smoke + region,
                data = dataE_s, refresh = 0)

describe_prior(m28) |> print_md()
```

Parameter	Prior_Distribution	Prior_Location	Prior_Scale
(Intercept)	normal	2.14	2.33
age	normal	0.00	0.16
sexmale	normal	0.00	4.67
bmi	normal	0.00	0.39
kids	normal	0.00	1.97
smokeyes	normal	0.00	5.79
regionNW	normal	0.00	5.49
regionSE	normal	0.00	5.23
regionSW	normal	0.00	5.54

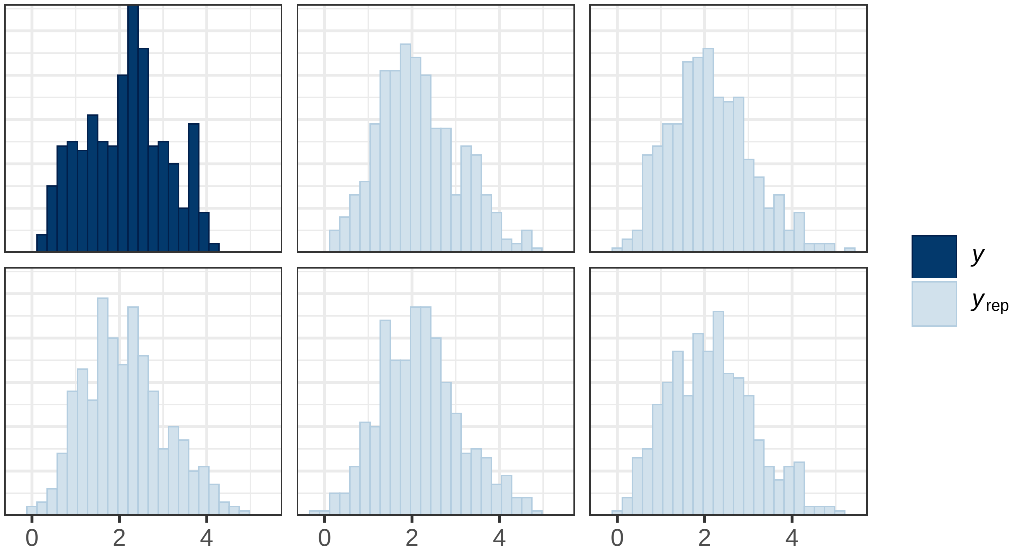
```r
describe_posterior(m28) |> print_md()
```

Table 9: Summary of Posterior Distribution

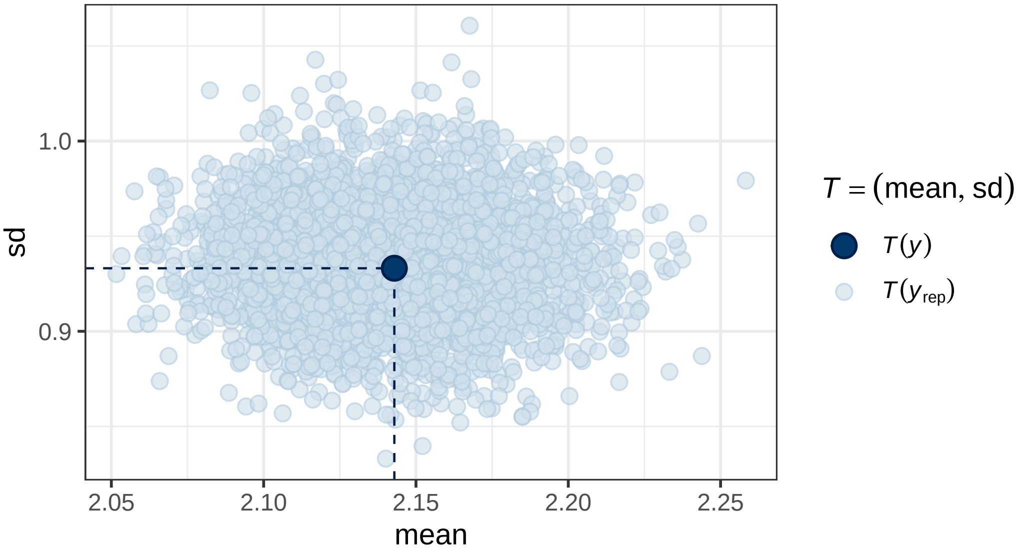
					% in		
Parameter	Median	95% CI	pd	ROPE	ROPE	Rhat	ESS
(Intercept)	-0.19	[-0.44, 0.07]	92.60%	[-0.09, 0.09]	21.58%	1.000	4431.00
age	0.04	[0.03, 0.04]	100%	[-0.09, 0.09]	100%	1.000	4272.00
sexmale	-1.95e-03	[-0.08, 0.08]	51.62%	[-0.09, 0.09]	100%	1.000	4683.00
bmi	0.02	[0.01, 0.02]	100%	[-0.09, 0.09]	100%	1.000	3996.00
kids	0.10	[0.06, 0.13]	100%	[-0.09, 0.09]	36.84%	1.000	4658.00
smokeyes	1.51	[1.41, 1.61]	100%	[-0.09, 0.09]	0%	0.999	4106.00
regionNW	0.03	[-0.09, 0.15]	69.90%	[-0.09, 0.09]	85.97%	1.000	3382.00
regionSE	-0.08	[-0.20, 0.04]	90.33%	[-0.09, 0.09]	58.97%	1.000	3053.00
regionSW	-0.03	[-0.15, 0.09]	67.62%	[-0.09, 0.09]	88.47%	1.001	3060.00

Setup for Item Q28, part 2 of 2

```
pp_check(m28, plotfun = "hist", nreps = 5, bins = 25)
```



```
pp_check(m28, plotfun = "stat_2d", stat = c("mean", "sd"))
```



28 Q28

Which of the following statements are true, based on the output provided on the previous two pages[13] as part of the **Setup for Item Q28**?

CHECK ALL OF THE TRUE STATEMENTS.

a. According to model `m28`, there is a greater than 90% probability that the effect of having more children on total charges is positive.
b. Each additional kg/m^2 of BMI is associated with an increase in the natural logarithm of total charges in this model, and the median of the posterior distribution for that effect is 0.02.
c. Model `m28` allows us to conclude that there is a 95% chance that the difference between smokers and non-smokers is between 1.41 and 1.61 in terms of the natural logarithm of total charges.
d. The posterior predictive checks provided suggest that model `m28` provides a poor fit to the data.
e. The Rhat values indicate a serious problem with the fitting of model `m28` with regard to the `smokeyes` regression coefficient.
f. This model incorporates a highly informative prior.
g. None of these statements are true.

29 Q29 (4 points)

Build a robust linear model using Huber weights, which I'll call model `m29`, to predict the logarithm of total charges in the `dataE_s` tibble using all six predictors which you included in model `m25`.

Item Q29 has two parts, each worth 2 points.

a. Comparing models `m25` and `m29`, which has the better predictive value as evaluated by the Bayes information criterion?

b. Is the 95% confidence interval for the `smoke` coefficient in model `m29` wider or narrower than the same confidence interval from model `m25`? Answer this question in a clear English sentence, which should specify both of the endpoints for each of the confidence intervals.

[13] I'd look at the Slides from Class 23 and the Course Notes Chapter 33.

30 Q30 (4 points)

Below, I fit model `m25` for the natural logarithm of total charges again, but this time incorporating all 20 of the multiple imputations created earlier as `dataE_20imps`.

```r
fitimp <- with(dataE_20imps,
               lm(log(charge_k) ~ age + sex + bmi + kids + smoke + region))

tidy(pool(fitimp), conf.int = TRUE, conf.level = 0.95) |>
  select(term, estimate, std.error, conf.low, conf.high, df, fmi) |>
  gt() |> fmt_number(decimals = 3)
```

term	estimate	std.error	conf.low	conf.high	df	fmi
(Intercept)	−0.182	0.132	−0.442	0.077	413.194	0.019
age	0.036	0.002	0.033	0.039	405.697	0.029
sexmale	−0.015	0.044	−0.102	0.072	350.551	0.079
bmi	0.017	0.004	0.010	0.024	413.110	0.019
kids	0.102	0.018	0.066	0.138	394.596	0.041
smokeyes	1.510	0.053	1.406	1.613	415.075	0.016
regionNW	0.030	0.063	−0.093	0.153	337.538	0.089
regionSE	−0.061	0.062	−0.184	0.061	387.217	0.048
regionSW	−0.037	0.062	−0.159	0.086	376.639	0.058

```r
glance(pool(fitimp)) |> gt() |>
  fmt_number(columns = r.squared:adj.r.squared, decimals = 3)
```

nimp	nobs	r.squared	adj.r.squared
20	432	0.786	0.782

Item Q30 has two parts, each worth 2 points.

a. How many of the nine coefficients fit in the table have **larger** estimates (after rounding to three decimal places) after multiple imputation than they did in model `m25` fit after single imputation?

b. Specify the raw R^2 value from model `m25` and compare it to the raw `r.squared` value after multiple imputation shown in the table above. How large is the difference between the two estimates, and what does that suggest about the use of single vs. multiple imputation here, in terms of the proportion of variation explained by the model?

Setting Up Q31-Q34: the `dataN` data

We have provided the `dataN.Rds` file to you, and will use this in Q31 through Q34. Dr. Love gathered these data from NHANES 2011-12 Demographics and Questionnaire data, specifically the `DEMO_G` (Demographics), `HSQ_G` (Current Health Status) and `PAQ_G` files[14].

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
`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
`FEMALE`	Sex	1 = Female, 0 = Male (based on `RIAGENDR`)

- `RIDRETH3` categories and their counts are
  - 1 = Mexican American (n = 312)
  - 2 = Other Hispanic (n = 240)
  - 3 = Non-Hispanic White (n = 919)
  - 4 = Non-Hispanic Black (n = 634)
  - 6 = Non-Hispanic Asian (n = 440)
  - 7 = Other Race including Multi-Racial (n = 95)
- `HSD010` Would you say your health in general is
  - 1 = Excellent (n = 269)
  - 2 = Very Good (n = 690)
  - 3 = Good (n = 927)
  - 4 = Fair (n = 322)
  - 5 = Poor (n = 41)

[14] See https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2011.

- `HSQ571` During the past 12 months have you donated blood?
  - 1 = Yes (n = 107)
  - 2 = No (n = 2139)
  - 7 = Refused (n = 0)
  - 9 = Don't Know (n = 3)
- `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 (n = 1244)
  - 2 = No (n = 1396)

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

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

glimpse(dataN)
```

```
Rows: 2,640
Columns: 9
$ SEQN     <dbl> 62161, 62164, 62169, 62172, 62176, 62180, 62184, 62189, 62195~
$ WTINT2YR <dbl> 102641.41, 127351.37, 14391.78, 26960.77, 53830.60, 20457.61,~
$ RIDAGEYR <dbl> 22, 44, 21, 43, 34, 35, 26, 30, 35, 42, 36, 28, 35, 38, 22, 3~
$ 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   <dbl> 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1~
```

```
miss_var_summary(dataN) |> filter(n_miss > 0)
```

```
# A tibble: 2 x 3
  variable n_miss pct_miss
  <chr>     <int>    <num>
1 HSD010      391     14.8
2 HSQ571      391     14.8
```

31 Q31

What percentage of the rows included in the `dataN` data describe subjects who have described their General Health 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 General Health variable.

32 Q32

Next, please answer the question asked in Q31 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". Again, express your response as a percentage, with a single decimal place.

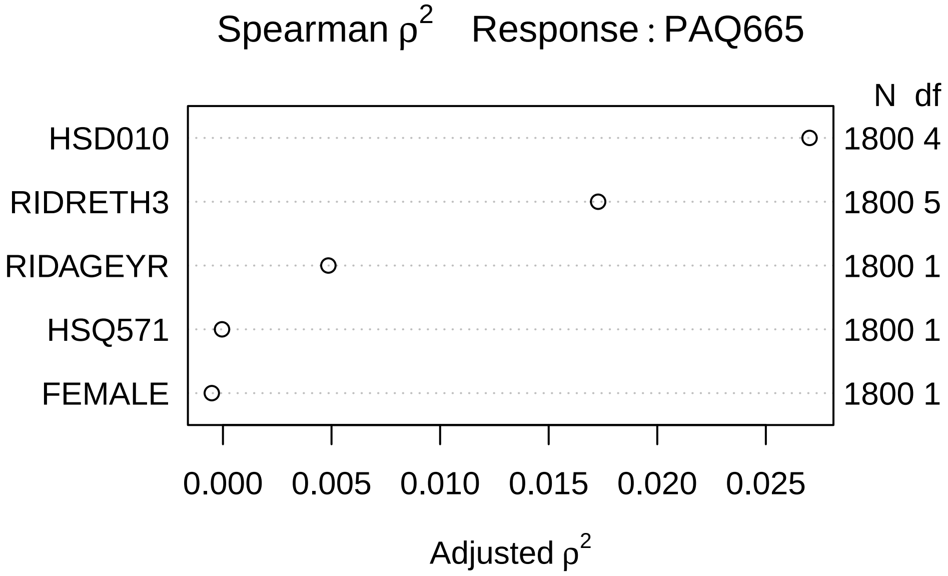
33 Q33

Suppose your intent is to create a tibble called `dat33` to support a *complete-case analysis* across all of the `dataN` data. Suppose that you have decided to treat as missing the data for any subjects who have either refused to answer or given the answer "don't know" to any one of the questions included in the data set as part of developing that complete-case analysis.

How many different subjects should be included in your new `dat33` tibble?

Setup for Q34

The Spearman ρ^2 plot below describes a random sample of 1800 observations drawn from `dataN`, after removing missing values. Suppose we fit a logistic regression model to predict the log odds of engaging in "moderate-intensity activities for at least 10 minutes continuously" for subjects in NHANES (without weighting) using the five predictors listed in the figure along with a single non-linear term.



34 Q34

A "main effects" model for `PAQ665` using these five predictors (and an intercept term) spends 12 degrees of freedom. If you were to add the single non-linear term recommended by the Spearman ρ^2 plot to the "main effects" model, how many **additional** degrees of freedom will be required?

This is the end of the Quiz.

Be sure to complete the Affirmation at the end of the Answer Sheet, and then submit your Answer Sheet, and verify that you have received your copy in your CWRU email by the deadline for submitting the Quiz.