431 Quiz 2 for Fall 2022

Thomas E. Love, Ph.D.

Due 2022-12-05 at 9 PM: Version 2022-11-29 11:35:13

Table of contents

Instructions for Students

There are **28** questions on this Quiz, and this PDF is **35** pages long. Be sure you have all **35** pages. It is to your advantage to answer all of the Questions. Your score is based on the number of correct responses, so there's no chance a blank response will be correct, while a guess might be, so you should definitely answer all of the questions.

0.1 The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, which will be found (once the Quiz starts) at https://bit.ly/431-2022-quiz2-answer-sheet. All of your answers must be submitted through that Google Form by 9 PM on Monday 2022-12-05, and no extensions will be made available, so do not wait until Monday evening to submit. We will only accept responses through the Google Form.

The Google Form's final question requires you to type in your full name to affirm that you followed the rules for the Quiz. You must complete that affirmation before you can submit your responses. After submission (like a Minute Paper) you will be emailed a copy of your submission, with a link allowing you to edit your responses.

If you wish to work on some of the quiz and then return later, do so by [1] completing the final question (the affirmation) by typing in your full name, and then [2] submitting the quiz. You will then receive a link at your CWRU email which will let you return to the quiz as often as you like without losing your progress.

0.2 Getting Help During the Quiz

This is an open book, open notes quiz. You are welcome to consult the materials provided on the course website and that we've been reading in the class, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants. You will be required to complete a short affirmation that you have obeyed these rules as part of submitting the Quiz.

If you need clarification on a Quiz question, you have exactly three ways of getting help:

1. You can ask your question in a private question to the instructors on Campuswire, or
2. You can ask your question via email to **431-help at case dot edu**.
3. You can ask your question during the "Ask Me Anything" session on Thursday 2022-12-01 from 1 to 2 PM with Dr. Love.

While the complete Quiz is available, we will not answer questions about the Quiz except through the three approaches listed above. We promise to respond to all questions received before 5 PM on 2022-12-05 at that time, if not sooner.

- Specific questions are more likely to get helpful answers.
- We will not review your code or your English for you.
- We will not tell you if your answer is correct, or if it is complete.
- We will email all students if we find an error in the Quiz that needs fixing.

When Should I ask for help?

We recommend the following process.

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

0.3 Scoring and Timing of the Quiz

Each question is worth 3 to 5 points, as indicated, and this sums to 104 points in total. The questions are not in any particular order, and range in difficulty from "things I expect everyone to get right" to "things that are deliberately tricky". The Quiz is meant to take **5-8** hours. I expect most students will take **4-10** hours, and some will take as little as 2 or as many as 12. It is not a good idea to spend a long time on any one question.

Dr. Love will grade all Quizzes, and you should have your result by Noon on Thursday 2022-12-08.

0.4 Writing Code into the Google Form

1. Occasionally, we ask you to provide a single line of code. If not otherwise specified, a single line of code in response should contain no more than two pipes, although you may or may not need the pipe in any particular setting.
2. You need not include the `library` command at any time in your responses on the Google Form. Assume in all questions that all relevant packages have been loaded in R.
3. If you are asked to complete a bootstrap method, use the default number of bootstrap replications (this is `B = 1000` for the Hmisc package's `smean.cl.boot` and `B.reps = 2000` for Dr. Love's `bootdif` within `Love-boost.R`.) Use these defaults by simply not setting a value for `B` or `B.reps` in calling the relevant function. Be sure in either case that you have set a seed properly immediately before running the bootstrap procedure.

4. When completing any procedure that requires random sampling, use the command `set.seed(4312021)` to set your random seed, and use `4312021` as that random seed. Do this at the start of the chunk of R code where you use the procedure that requires a set of random numbers, and do it again if you need a new set of random numbers later in the Quiz.

0.5 The Data Sets for this Quiz

We have provided four data sets that are mentioned in the Quiz. You will find them in our Shared Google Drive in the Quiz 2 folder. They may be helpful to you.

- `quiz2_surveyA.csv`, first mentioned in Question 2
- `quiz2_hospsim.csv`, first mentioned in Question 18
- `quiz2_ra.csv`, first mentioned in Question 22
- `quiz2_wtchg.csv`, first mentioned in Question 28

0.6 Packages loaded by Dr. Love when developing this Quiz

This doesn't mean you need to use all of these packages or even that I used them all, but it does mean that I did not add any additional packages to this list in building the quiz or the answer sketch.

```
source("data/Love-boost.R")

library(broom); library(car); library(Epi)
library(equatiomatic); library(GGally); library(ggrepel)
library(glue); library(Hmisc); library(janitor)
library(kableExtra); library(knitr); library(modelsummary)
library(mosaic); library(naniar); library(patchwork)
library(psych); library(pwr)
library(tidyverse)

theme_set(theme_bw())
```

1 Question 01 (4 points)

Once a confidence interval 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. We cannot tell whether the CI width will become wider, narrower or not change

Setup for Questions 02-15

Questions 02-15 all refer to the same data set (the `quiz2_surveyA.csv` data) which I have provided to you in our Shared Google Drive in the data subfolder within the Quiz 2 folder.

Consider the `quiz2_surveyA.csv` file Dr. Love has provided. The file contains (lightly edited) responses to nine items that I have asked of students during the first week of 431, since 2014. The `quiz2_surveyA.csv` file also includes a subject identification code called `student`, as well as the `year` (which ranges from 2014 through 2022) in which the survey was given.

- Ingest the data into R (name your tibble `surveyA` to match the answer sketch) and complete analyses to address Questions 02-15.
- Note that each of questions 02-15 contains two parts.

2 Question 02 (4 points)

The `height_in` variable stores responses to the item "what is your height?" that have been converted to measurements in inches.

a. How many of the students who responded to the `height_in` item gave responses that fell within two standard deviations of the sample mean?
b. What percentage of students with a response to `height_in` would we expect to see in your count in Question 2a if the Normal distribution was a perfect model for the distribution? Round your answer to zero decimal places.

3 Question 03 (4 points)

The `lastsleep` variable contains responses to the item "How many hours did you sleep last night?"

a. In which `year` did the sample of `lastsleep` values have the smallest mean number of hours?
b. After rounding to one decimal place, what was the sample mean of `lastsleep` in the `year` you identified in Question 3a?

4 Question 04 (4 points)

The most common response to the item "What is your favorite color?", which was gathered in the `favcolor` variable was blue, mentioned by nearly half of the respondents. The second most common response was green.

a. What color was the third most common choice, and how many subjects gave that response?
b. What color was the fourth most common choice, and how many subjects gave that response?

5 Question 05 (4 points)

All subjects in the `surveyA` data were assigned a `student` code and `year` value by Dr. Love. The other nine variables were collected from their responses to the items on the survey.

a. Which of the nine variables (excluding `student` and `year`) provided has the largest number of missing values?
b. How many values are missing in the variable you identified in Question 5a?

6 Question 06 (4 points)

a. Identify the `student` code for the subject with the largest number of missing values.
b. For which variables (excluding `student` and `year`) has the subject you identified in Question 6a provided data?

7 Question 07 (4 points)

Ava and Ivy are two analysts working separately, who wish to compare the `haircut` prices across the survey years using an appropriate overall test of whether the average (mean or pseudo-median) `haircut` values are detectably different by year, using a 90% confidence level. Each analyst assumes all missing `haircut` prices are missing completely at random. Ava also assumes that each year's `haircut` prices follow a Normal distribution with similar variance across years. Ivy, on the other hand, assumes that the `haircut` prices are symmetric, but that they do not follow a Normal distribution due to problems in the tails of the distributions.

a. State the p values (rounded to three decimal places) for each of the two analysts (Ava and Ivy) under the assumptions stated above.
b. Which of the analysts will declare a statistically detectable difference between the average `haircut` prices across years?

    a. Ava alone
    b. Ivy alone
    c. Both Ava and Ivy
    d. Neither Ava nor Ivy

8 Question 08 (4 points)

The `lecture` variable describes agreement (1 = Strongly Disagree to 5 = Strongly Agree) with the statement "I prefer to learn from lectures over learning from activities." The `alone` variable described agreement (on the same 1-5 scale) with the statement "I prefer to work on projects alone over working on a team." Create a 5x5 table comparing lecture to alone (with `lecture` in the rows and `alone` in the columns) for the subjects in the data who have complete information on the `lecture` and `alone` variables, and also have a value in the `height_in` variable.

a. Following the Question 8 instructions to build a table, you should find one cell which contains exactly five subjects. Identify the `student` code for the shortest of those five subjects.
b. How many of your table's cells contain at least 10 subjects?

9 Question 09 (4 points)

Now use R to combine the responses 1 and 2 and 4 and 5 for the `lecture` and `alone` variables discussed in Question 8, and then create a new contingency table with `lecture` in the rows and `alone` in the columns. Your new table should have 3 rows and 3 columns. Now, use an appropriate hypothesis test to assess, with 90% confidence, whether the rows in your new table are independent of the columns.

a. What type of hypothesis test did you use to address this question?

    a. A test based on the t distribution
    b. A test based on the chi-square distribution
    c. A test based on a bootstrap
    d. A test based on the F distribution
    e. None of these

b. After rounding to five decimal places, what p value do you obtain from your test in Question 9a?

10 Question 10 (4 points)

Now consider the `english` variable, which specifies for each subject whether their most comfortable language is English, or some other language, and the `smoker` variable, which divides subjects into group 1 (non-smokers), or group 2 (former smoker), or group 3 (current smoker). For Question 10, including only the subjects with complete data on these two variables. Collapse the `smoker` responses of 2 and 3 into a single group, which we'll call the smokers for this question. Now, estimate the odds ratio for being a non-smoker for students who are most comfortable speaking in a language other than English, as compared to students who are most comfortable speaking in English. Use a Bayesian adjustment in Question 10 (adding two to each count) only if any of the observed counts are below 10.

a. Provide a point estimate for the odds ratio identified in Question 10, rounded to two decimal places.
b. Again rounding to two decimal places, specify a 90% confidence interval for the estimate you provided in Question 10a.

11 Question 11 (4 points)

Consider the responses to the item "How important do you think statistics will be in your future career?" as gathered in the `statfuture` variable. There have been five years so far where the class had no missing responses on this item. Use the data from those five years to estimate the proportion of subjects who responded with the value 7 (meaning extremely important), using the approach due to Agresti and Coull.

a. Rounding to three decimal places, specify a 90% confidence interval for the true value of the proportion described in Question 11.
b. Is the sample proportion of "7" values from the 2022 survey, ignoring any missing responses to the `statfuture` variable, inside the interval you reported in Question 11a?

    a. Yes, it is inside the interval
    b. No, the sample proportion is lower than all values in the interval
    c. No, the sample proportion is higher than all values in the interval

12 Question 12 (4 points)

Fit two linear models to predict `haircut` price on the basis of `height_in` and `english`, in each case using only those subjects with complete data on each of those three variables, as follows.

- In model **m12A**, use only `height_in` to predict `haircut`.
- In model **m12B**, use both `height_in` and `english` to predict `haircut`.

a. Specify the values of BIC for models m12A and m12B, rounded to zero decimal places.
b. According to the AIC and BIC, which of the two models (m12A or m12B) shows stronger performance?

    a. Model m12A shows stronger performance on both AIC and BIC
    b. Model m12B shows stronger performance on both AIC and BIC
    c. Model m12A is better on AIC, but m12B is better on BIC
    d. Model m12A is better on BIC, but m12B is better on AIC

13 Question 13 (4 points)

Using model m12B as described in Question 12, identify the standardized residuals for each subject, and identify the largest (in absolute value) of these standardized residuals.

a. Specify the `student` code for the subject with the largest (in absolute value) standardized residual in model m12B.
b. Specify the fitted and observed values of `haircut` for the subject you identified in Question 13a, in each case rounding to the nearest integer.

14 Question 14 (4 points)

Consider again the subject you identified by their `student` code in Question 13.

a. Rounded to three decimal places, what is the Cook's distance (using model m12B) for the subject identified in Question 13?
b. Does Cook's distance identify that subject as influential on model m12B?

    a. Yes
    b. No
    c. It is impossible to tell from the available information

15 Question 15 (4 points)

Now, let's create a new model, called model **m_15**, where you rerun model m12B, but this time excluding the `student` you identified in Question 13.

a. Specify the residual standard error (rounded to 1 decimal place) and the R-squared value (expressed as a proportion, and rounded to 3 decimal places) for model m_15.
b. Consider an appropriate plot of the standardized regression residuals for model m15. Is there a problem with the assumption of Normality?

    a. Yes, because there is substantial skew in the distribution of the standardized residuals
    b. Yes, because the residuals are symmetric, but there are several problematic outliers
    c. Yes, because the residuals are symmetric, but their distribution is light-tailed compared to a Normal model
    d. No, because the residuals are well-approximated by a Normal distribution

Note: Question 15 is the last question that has anything to do with the `quiz2_surveyA.csv` data.

16 Question 16 (4 points)

Data from a pilot study describe the pre-operative and post-operative creatinine clearance (in ml/minute) of six patients anesthetized with a new drug combination we want to test.

Subject	Pre-operative	Post-operative
1	110	137
2	101	93
3	71	99
4	73	91
5	133	112
6	118	134

Suppose we are willing to assume that the results of this study are an appropriate pilot for a larger study we are designing that will also compare pre-operative and post-operative creatinine clearance rates. Suppose we want to maintain a two-sided 5% significance level. Suppose that an effect that is at least half the effect size seen in the pilot study will be our minimum standard for clinical importance. We also assume that the pilot's standard deviation estimates are 25% too low, so we'll multiply them by 1.25 in developing our design.

As a hint, consider the following output:

```
subj <- 1:6; pre <- c(110, 101, 71, 73, 133, 118)
post <- c(137, 93, 99, 91, 112, 134)
pilot <- tibble(subj, pre, post)

favstats(~ pre, data = pilot)
```

```
min Q1 median  Q3 max mean       sd n missing
 71 80  105.5 116 133  101 24.81129 6       0
```

```
favstats(~ post, data = pilot)
```

```
min   Q1 median    Q3 max mean       sd n missing
 91 94.5  105.5 128.5 137  111 20.36664 6       0
```

```
favstats(~ pre-post, data = pilot)
```

```
min     Q1 median Q3 max mean    sd n missing
-28 -24.75    -17  2  21  -10 19.99 6       0
```

Question 16 (continued)

Under the conditions specified above, you have been asked to determine the smallest number of subjects we must include in our new study to obtain 90% power to detect the minimum clinically important effect.

a. What is the smallest number of subjects the new study will have to include?
b. Provide your R code (a single line will suffice) to justify your response to Question 16a.

Setup for Question 17

In this question, we're going to look at four sets of regression residual plots, for four separate models, called Model 1, Model 2, Model 3 and Model 4 The four models estimate different outcomes.

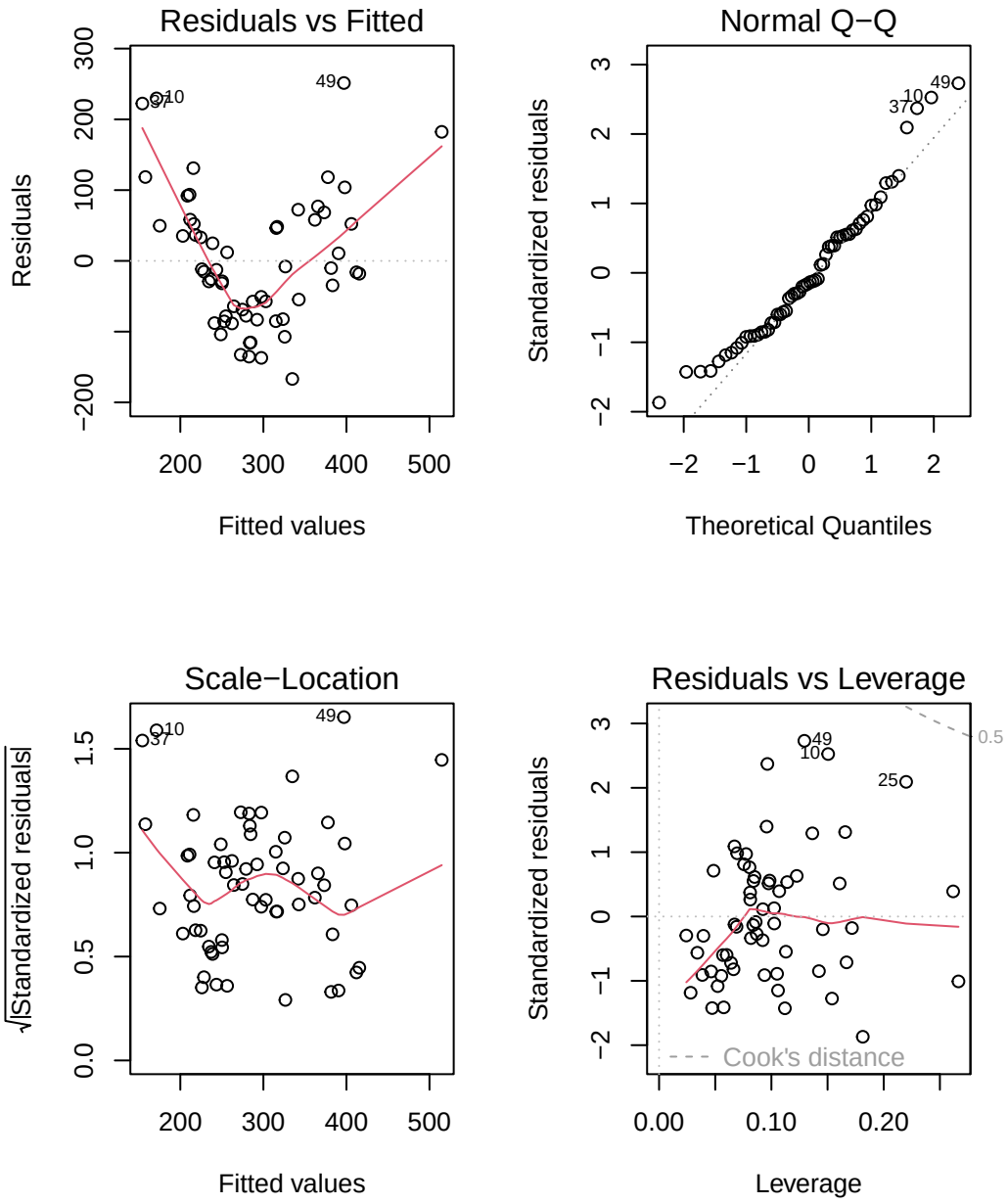
17 Question 17 (4 points)

For each of the four models whose residual plots are shown below, specify the most appropriate conclusion regarding its residual plots. For each model (1, 2, 3 and 4), you are asked to select (exactly) one of the following responses…

a. This model has a serious problem with the assumption of linearity
b. This model has a serious problem with the assumption of constant variance
c. This model has a serious problem with the assumption of Normality
d. There are no serious problems evident in this model's residual plots

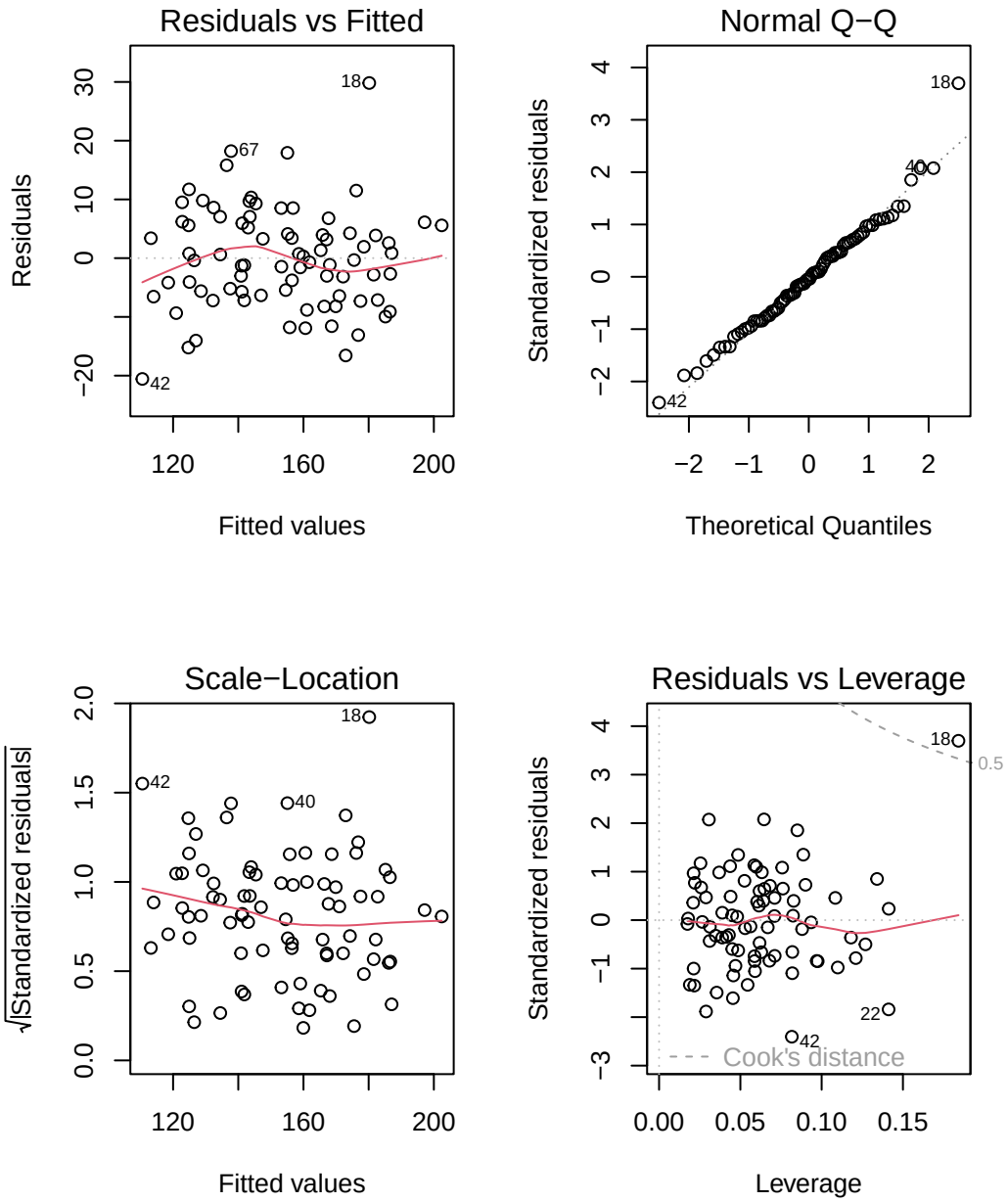
Question 17: Residual Plots - Model 1

Here are the plots for model 1, which includes 60 subjects.



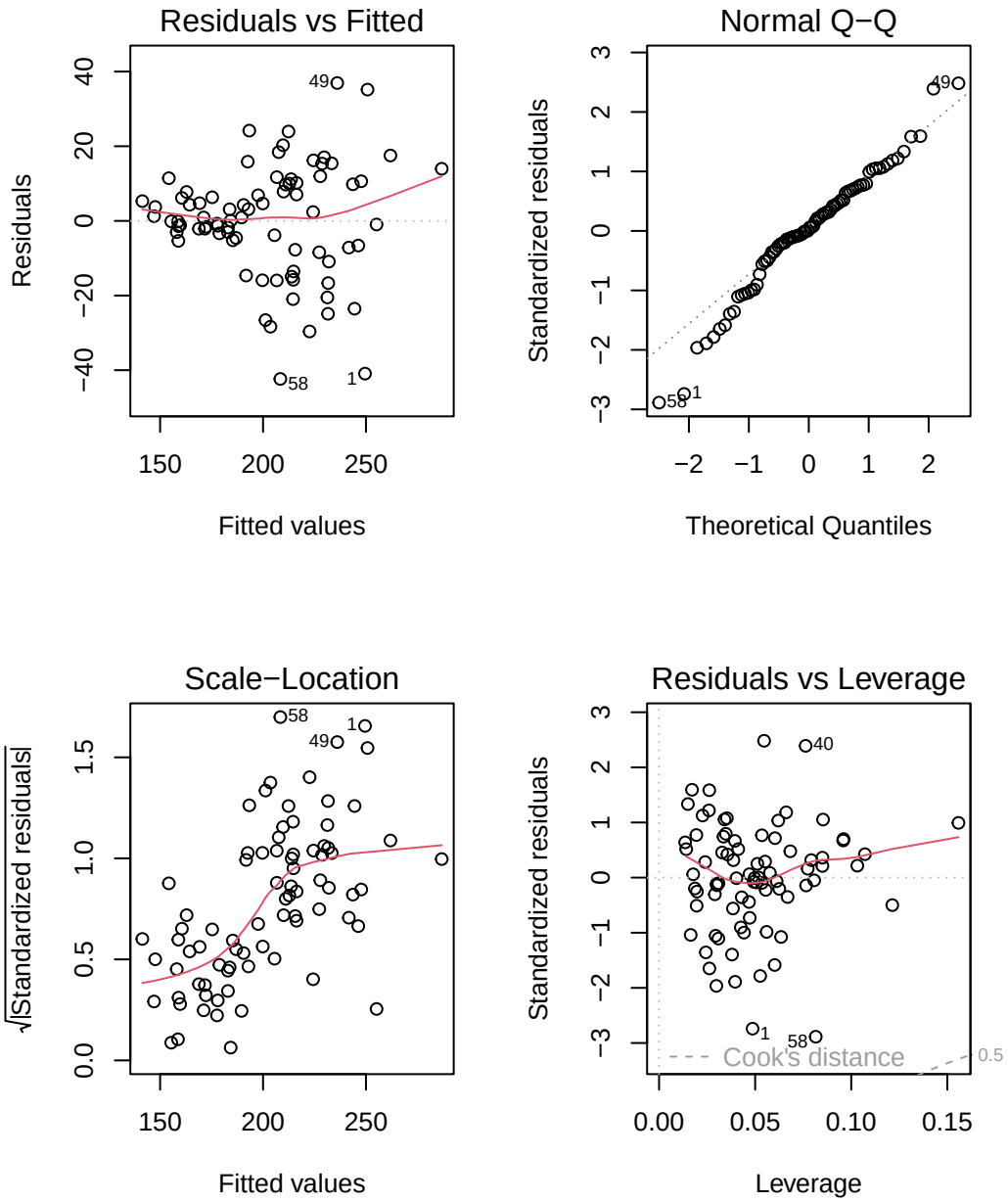
Question 17: Residual Plots - Model 2

Here are the plots for model 2, which includes 80 subjects.



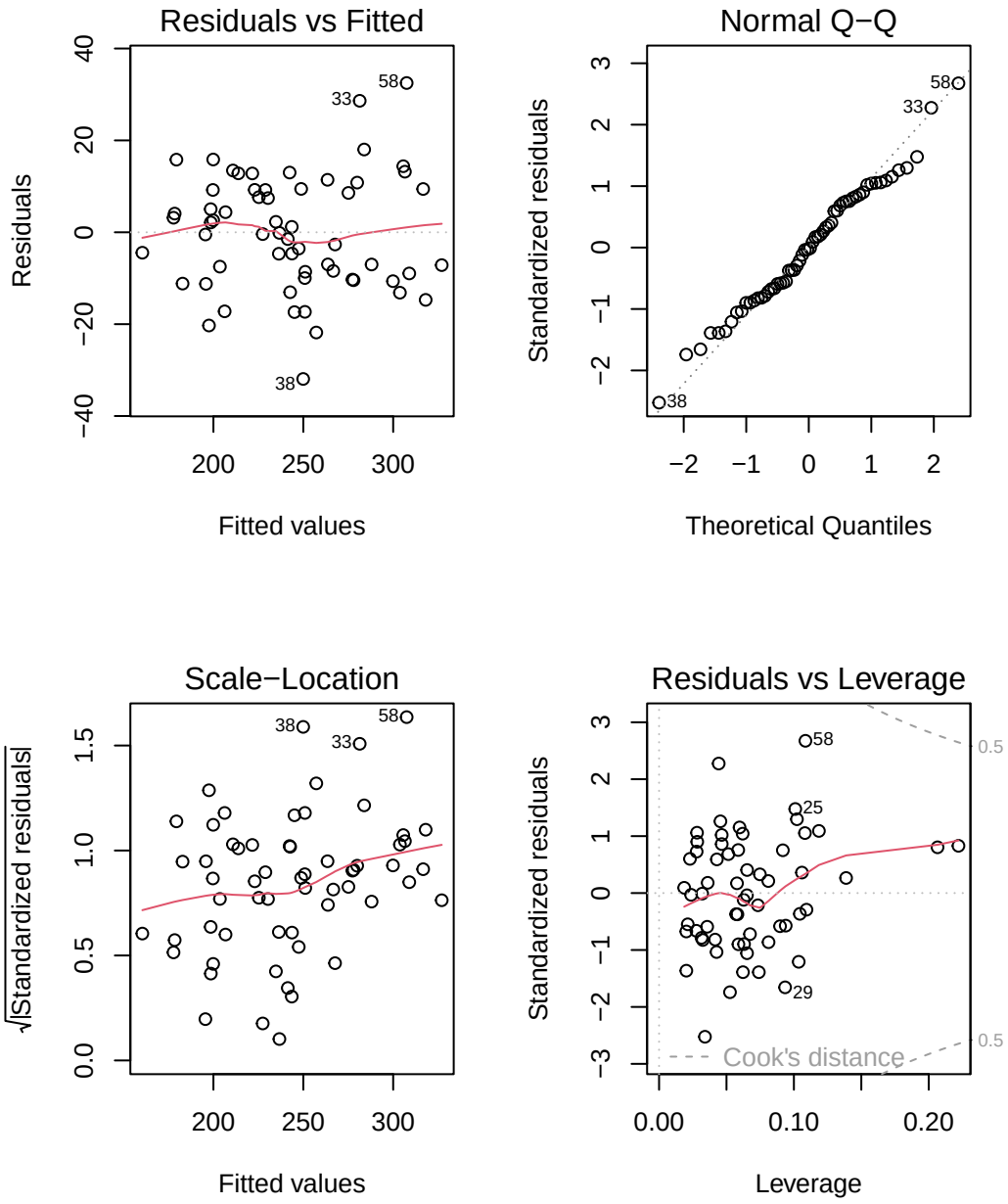
Question 17: Residual Plots - Model 3

Here are the plots for model 3, which includes 80 subjects.



Question 17: Residual Plots - Model 4

Here are the plots for model 4, which includes 60 subjects.



Background for Questions 18-21

For Questions 18-21, consider the data I have provided in the `quiz2_hospsim.csv` file, which describe 650 patients at a metropolitan hospital. They are simulated. Available are:

- `subj_id` = Subject Identification Number (not a meaningful code)
- `age` = the patient's age, in years (all subjects are between 21 and 75)
- `insurance` = the patient's insurance type (MEDICARE, COMMERCIAL, MEDICAID, UNINSURED)
- `hsgrads` = the percentage of adults in the patient's home neighborhood who have at least a high school diploma (this measure of educational attainment is used as an indicator of the socio-economic place in which the patient lives)
- `a1c` = the patient's hemoglobin A1c level (in %)
- `ldl` = the patient's LDL cholesterol level (in mg/dl)
- `bmi` = the patient's body mass index (in kg/square meter)
- `sbp` = the patient's systolic blood pressure (in mm Hg)
- `statin` = does the patient have a prescription for a statin medication (YES or NO)
- `clinic_type` = whether the patient goes to a newly built clinic or an old clinic
- `sex` = the patient's sex (FEMALE or MALE)

18 Question 18 (3 points)

Using the `hospsim` data, what is the point estimate and 95% confidence interval for the relative risk which compares the chance of receiving a statin if you are MALE divided by the chance of receiving a statin if you are FEMALE. Do **NOT** use a Bayesian augmentation here, and round your answers (for the point estimate and each endpoint) to two decimal places.

19 Question 19 (3 points)

Using the `hospsim` data, perform an appropriate analysis to determine whether insurance type is associated with the education (`hsgrads`) variable, ignoring all other information in the `hospsim` data. In developing your response, use the `hsgrads` outcome variable as it is, without applying a transformation or excluding any values. Which of the following conclusions is most appropriate based on your results?

a. The ANOVA F test does not meet the standard for a statistically detectable result with 95% confidence, so it doesn't make sense to compare insurance types pairwise.
b. The ANOVA F test shows a statistically detectable result at the 5% significance level, and a Tukey HSD comparison retaining a 95% family-wise confidence level reveals that Medicare shows detectably higher education levels than Uninsured.
c. The ANOVA F test shows a statistically detectable result at the 5% significance level, and a Tukey HSD comparison retaining a 95% family-wise confidence level reveals that Medicaid's education level is detectably lower than either Medicare or Commercial.
d. The ANOVA F test shows a statistically detectable result at the 5% significance level, and a Tukey HSD comparison retaining a 95% family-wise confidence level reveals that Uninsured's education level is detectably lower than Commercial or Medicare.
e. None of these conclusions is appropriate.

20 Question 20 (3 points)

Using the `hospsim` data, build a model to predict LDL cholesterol using the main effects of all of the other available variables except subject ID. After adjusting for all of the other variables, which of the following statements best describes your results? Do not transform your outcome, and use a 95% confidence level to motivate your conclusions.

a. Whether you were in an old or new clinic type doesn't seem to matter in a detectable way for predicting LDL.
b. Subjects at older clinics had detectably higher LDL levels, holding everything else constant, and the model accounts for less than 20% of the variation in LDL.
c. Subjects at older clinics had detectably lower LDL levels, holding everything else constant, and the model accounts for less than 20% of the variation in LDL.
d. Subjects at older clinics had detectably higher LDL levels, holding everything else constant, and the model accounts for 20% or more of the variation in LDL.
e. Subjects at older clinics had detectably lower LDL levels, holding everything else constant, and the model accounts for 20% or more of the variation in LDL.

21 Question 21 (3 points)

Using the `hospsim` data, now build a model using the main effects of sex and insurance type (and their interaction) to predict hemoglobin A1c. In this new model, identify the subject with the largest residual (in absolute value). Which of the following characteristics best describes this subject?

a. This is a female Medicare patient visiting a new clinic.
b. This is a female Medicare patient visiting an old clinic.
c. This is a male Medicare patient visiting a new clinic.
d. This is a male Medicare patient visiting an old clinic.
e. None of these accurately describe the subject in question.

Background for Questions 22-25

In a double-blind trial, 350 patients with active rheumatoid arthritis were randomly assigned to receive one of two therapy types: a cheaper one, or a pricier one, and went on to participate in the trial.

The primary outcome was the change in DAS28 at 48 weeks as compared to study entry. The DAS28 is a composite index of the number of swollen and tender joints, the erythrocyte sedimentation rate, and a visual-analogue scale of patient-reported disease activity. A decrease in the DAS28 of 1.2 or more (so a change of -1.2 or below) was considered to be a clinically meaningful improvement. Data are in the `quiz2_ra.csv` file. Here are a few potentially relevant summaries.

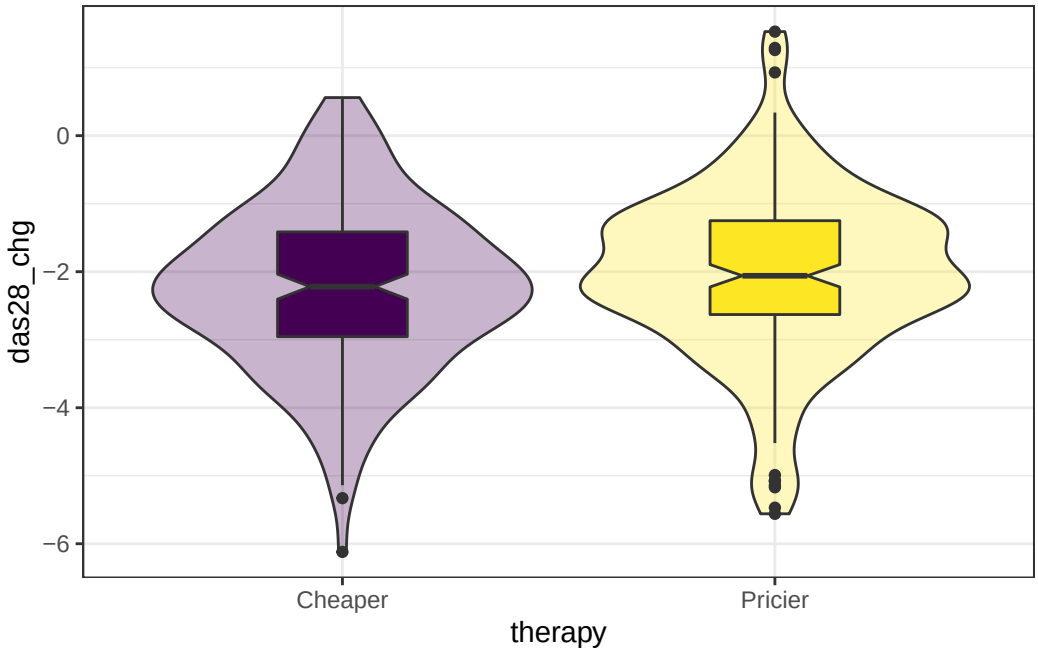
```
ra <- read_csv("data/quiz2_ra.csv", show_col_types = FALSE)
head(ra, 2)
```

```
# A tibble: 2 x 3
  subject das28_chg therapy
  <chr>       <dbl> <chr>
1 S-01        -6.12 Cheaper
2 S-02        -5.56 Pricier
```

```
favstats(das28_chg ~ therapy, data = ra)
```

```
  therapy   min     Q1 median     Q3  max      mean       sd   n missing
1 Cheaper -6.12 -2.955  -2.22 -1.415 0.56 -2.250857 1.208183 175       0
2 Pricier -5.56 -2.630  -2.06 -1.250 1.53 -2.027486 1.260694 175       0
```

```
ggplot(data = ra, aes(x = therapy, y = das28_chg, fill = therapy)) +
  geom_violin(alpha = 0.3) + geom_boxplot(width = 0.3, notch = TRUE) +
  guides(fill = "none") + scale_fill_viridis_d()
```



22 Question 22 (3 points)

A student completed four analyses of the data described in the Background for Questions 22-25, and those analyses are shown below. Which of the following 90% confidence intervals for the change in DAS28 at 48 weeks most appropriately compares the pricier therapy to the cheaper one?

d. Analysis D
e. Analysis E
f. Analysis F
g. Analysis G

Analysis D for Question 22

```
t.test(das28_chg ~ therapy, data = ra, var.equal = TRUE, conf.level = 0.90) |>
  tidy(conf.int = TRUE) |>
  mutate(estimate = estimate1 - estimate2) |>
  select(estimate, conf.low, conf.high, method)
```

```
# A tibble: 1 x 4
  estimate conf.low conf.high method
     <dbl>    <dbl>     <dbl> <chr>
1   -0.223   -0.441  -0.00568 Two Sample t-test
```

Analysis E for Question 22

```
t.test(das28_chg ~ therapy, data = ra, paired = TRUE, conf.level = 0.90) |>
  tidy(conf.int = TRUE) |>
  select(estimate, conf.low, conf.high, method)
```

```
# A tibble: 1 x 4
  estimate conf.low conf.high method
     <dbl>    <dbl>     <dbl> <chr>
1   -0.223   -0.245    -0.201 Paired t-test
```

Analysis F for Question 22

```
wilcox.test(das28_chg ~ therapy, data = ra, paired = TRUE,
                  conf.int = TRUE, conf.level = 0.90) |>
  tidy() |>
  select(estimate, conf.low, conf.high, method)
```

```
# A tibble: 1 x 4
  estimate conf.low conf.high method
     <dbl>    <dbl>     <dbl> <chr>
1   -0.230   -0.245    -0.215 Wilcoxon signed rank test with continuity correct~
```

Analysis G for Question 22

```
wilcox.test(das28_chg ~ therapy, data = ra, conf.int = TRUE, conf.level = 0.90) |>
  tidy() |>
  select(estimate, conf.low, conf.high, method)
```

```
# A tibble: 1 x 4
  estimate conf.low conf.high method
     <dbl>    <dbl>     <dbl> <chr>
1   -0.240   -0.450   -0.0300 Wilcoxon rank sum test with continuity correction
```

23 Question 23 (3 points)

Referring again to the study initially described in Question 22, which of the following analyses provides an appropriate 90% confidence interval for the difference (cheaper - pricier) in the proportion of participants who had a clinically meaningful improvement (DAS28 change of -1.2 or below) at 48 weeks?

j. Analysis J
k. Analysis K
l. Analysis L
m. Analysis M
n. None of the above.

Analysis J for Question 23

```
ra <- read_csv("data/quiz2_ra.csv", show_col_types = FALSE)
ra <- ra |>
  mutate(improved = das28_chg < -1.2) |>
  mutate(improved = fct_relevel(factor(improved), "FALSE"))
ra |> tabyl(improved, therapy)
```

```
 improved Cheaper Pricier
    FALSE      31      41
     TRUE     144     134
```

```
twobytwo(31, 41, 144, 134, "improved", "didn't improve",
         "cheaper", "pricier")
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome    : cheaper
Comparing : improved vs. didn't improve

               cheaper pricier    P(cheaper) 95% conf. interval
improved            31      41        0.4306    0.3217   0.5466
didn't improve     144     134        0.5180    0.4593   0.5762

                                   95% conf. interval
             Relative Risk:  0.8312    0.6227   1.1096
         Sample Odds Ratio:  0.7036    0.4173   1.1864
Conditional MLE Odds Ratio:  0.7043    0.4019   1.2246
    Probability difference: -0.0874   -0.2100   0.0416

             Exact P-value: 0.2339
        Asymptotic P-value: 0.1872
------------------------------------------------------
```

Analysis K for Question 23

```
ra <- read_csv("data/quiz2_ra.csv", show_col_types = FALSE)
ra <- ra |>
  mutate(improved = das28_chg <= -1.2) |>
  mutate(improved = fct_relevel(factor(improved), "TRUE"))
ra |> tabyl(improved, therapy)
```

```
 improved Cheaper Pricier
     TRUE     144     134
    FALSE      31      41
```

```
twobytwo(144, 134, 31, 41, "improved", "didn't improve",
         "cheaper", "pricier")
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : cheaper
Comparing : improved vs. didn't improve

               cheaper pricier    P(cheaper) 95% conf. interval
improved           144     134        0.5180    0.4593   0.5762
didn't improve      31      41        0.4306    0.3217   0.5466

                                  95% conf. interval
             Relative Risk: 1.2031    0.9013   1.6059
         Sample Odds Ratio: 1.4213    0.8429   2.3965
Conditional MLE Odds Ratio: 1.4198    0.8166   2.4880
    Probability difference: 0.0874   -0.0416   0.2100

             Exact P-value: 0.2339
        Asymptotic P-value: 0.1872
------------------------------------------------------
```

Analysis L for Question 23

```
ra <- read_csv("data/quiz2_ra.csv", show_col_types = FALSE)
ra <- ra |>
  mutate(improved = das28_chg < -1.2) |>
  mutate(improved = fct_relevel(factor(improved), "FALSE"))
ra |> tabyl(improved, therapy)
```

```
 improved Cheaper Pricier
    FALSE      31      41
     TRUE     144     134
```

```
twobytwo(31, 41, 144, 134, conf.level = 0.90,
         "improved", "didn't improve", "cheaper", "pricier")
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : cheaper
Comparing : improved vs. didn't improve

               cheaper pricier    P(cheaper) 90% conf. interval
improved            31      41        0.4306    0.3383   0.5279
didn't improve     144     134        0.5180    0.4687   0.5669

                                    90% conf. interval
             Relative Risk:  0.8312    0.6523   1.0592
         Sample Odds Ratio:  0.7036    0.4538   1.0908
Conditional MLE Odds Ratio:  0.7043    0.4379   1.1271
    Probability difference: -0.0874   -0.1914   0.0212

             Exact P-value: 0.2339
        Asymptotic P-value: 0.1872
------------------------------------------------------
```

Analysis M for Question 23

```
ra <- read_csv("data/quiz2_ra.csv", show_col_types = FALSE)
ra <- ra |>
  mutate(improved = das28_chg <= -1.2) |>
  mutate(improved = fct_relevel(factor(improved), "TRUE"))
ra |> tabyl(improved, therapy)
```

```
 improved Cheaper Pricier
     TRUE     144     134
    FALSE      31      41
```

```
twobytwo(144, 31, 134, 41, conf.level = 0.90,
         "cheaper", "pricier", "improved", "not improved")
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : improved
Comparing : cheaper vs. pricier

        improved not improved    P(improved) 90% conf. interval
cheaper      144           31         0.8229    0.7703   0.8655
pricier      134           41         0.7657    0.7090   0.8142

                                  90% conf. interval
             Relative Risk: 1.0746    0.9824   1.1756
         Sample Odds Ratio: 1.4213    0.9168   2.2034
Conditional MLE Odds Ratio: 1.4198    0.8872   2.2838
    Probability difference: 0.0571   -0.0141   0.1278

             Exact P-value: 0.2339
        Asymptotic P-value: 0.1872
------------------------------------------------------
```

24 Question 24 (3 points)

In response to unexpectedly low enrollment, the protocol was amended part-way through the trial described in Questions 22 and 23 to change the primary outcome from a binary outcome to a continuous outcome in order to increase the power of the study.

Originally, the proposed primary outcome was the difference in the proportion of participants who had a DAS28 of 3.2 or less at week 48. The original power analysis established a sample size target of 225 completed enrollments in each therapy group, based on a two-sided 10% significance level, and a desire for 90% power. In that initial power analysis, the proportion of participants with a DAS28 of 3.2 or less at week 48 was assumed to be 0.27 under the less effective of the two therapies.

What value was used in the power calculation for the proportion of participants with DAS28 of 3.2 or less at week 48 for the more effective therapy? State your answer rounded to two decimal places.

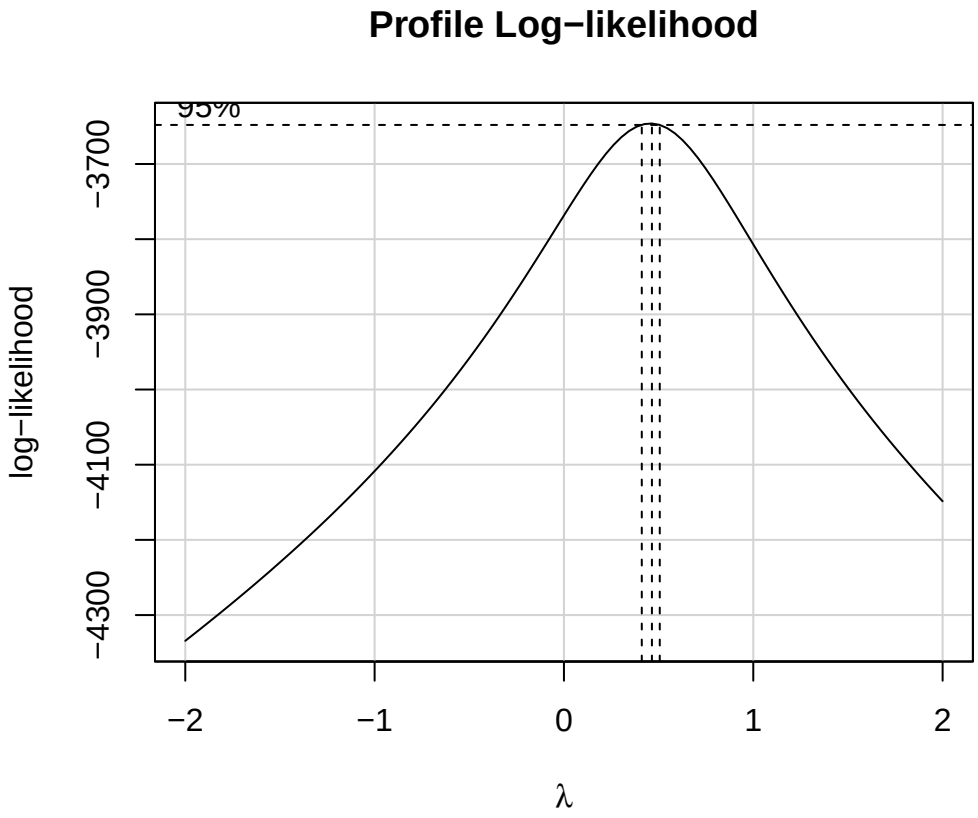
25 Question 25 (4 points)

In the trial we have been discussing since Question 22, 21 of the 222 subjects originally assigned to receive the cheaper therapy and 35 of the 219 subjects originally assigned to receive the pricier therapy experienced a serious adverse event (which included infections, gastrointestinal, renal, urinary, cardiac or vascular disorders, as well as surgical or medical procedures).

Suppose you wanted to determine whether or not there was a statistically detectable difference in the rates of serious adverse events in the two therapy groups at the 95% confidence level. Specify a single line of R code that would do this, appropriately.

26 Question 26 (3 points)

Consider the Box-Cox plot below, which was developed using a model to predict CD4 count (CD4 cells are the cells that the HIV virus kills; a normal range is about 500 - 1,500) using several predictors related to genetic makeup and several exposures of interest for a study involving 400 young men.



Which of the following is the most promising strategy for fitting a linear regression model to describe the relationship between the CD4 counts and the predictors of interest?

a. Model the inverse of CD4 count: 1/CD4 count.
b. Model the logarithm of CD4 count: $log(\text{CD4 count})$.
c. Model the square root of CD4 count: $\sqrt{\text{CD4 count}}$.
d. Model CD4 count without transformation.
e. Model the square of CD4 count: $(\text{CD4 count})^2$.
f. None of the above.
g. We cannot tell from the information provided.

27 Question 27 (3 points)

On 2019-09-25, Maggie Koerth-Baker at FiveThirtyEight published "We've Been Fighting the Vaping Crisis Since 1937." In that article, she quotes a 2019-09-06 article at the *New England Journal of Medicine* by Jennifer E. Layden et al. entitled "Pulmonary Illness Related to E-Cigarette Use in Illinois and Wisconsin: A Preliminary Report." Quoting that report:

> E-cigarettes are battery-operated devices that heat a liquid and deliver an aerosolized product to the user. … In July 2019, the Wisconsin Department of Health Services and the Illinois Department of Public Health received reports of pulmonary disease associated with the use of e-cigarettes (also called vaping) and launched a coordinated public health investigation…. We defined case patients as persons who reported use of e-cigarette devices and related products in the 90 days before symptom onset and had pulmonary infiltrates on imaging and whose illnesses were not attributed to other causes.

In the study, 53 case patients were identified, but some patients gave no response to the question of whether or not "they had used THC (tetrahydrocannabinol) products in e-cigarette devices in the past 90 days." 33 of the 41 reported THC use. Assume those 41 subjects are a random sample of all case patients that will appear in Wisconsin and Illinois in 2019.

Estimate an appropriate 90% confidence interval for the **PERCENTAGE** of case patients in Illinois and Wisconsin in 2019 that used THC in the 90 days prior to symptom onset using the Agresti-Coull method. Note that I've emphasized the word **PERCENTAGE** here, so as to stop you from instead presenting a proportion. Specify your point estimate of this **PERCENTAGE**, and then the lower and upper bound for your confidence interval, in each case rounded to a single decimal place.

28 Question 28 (5 points)

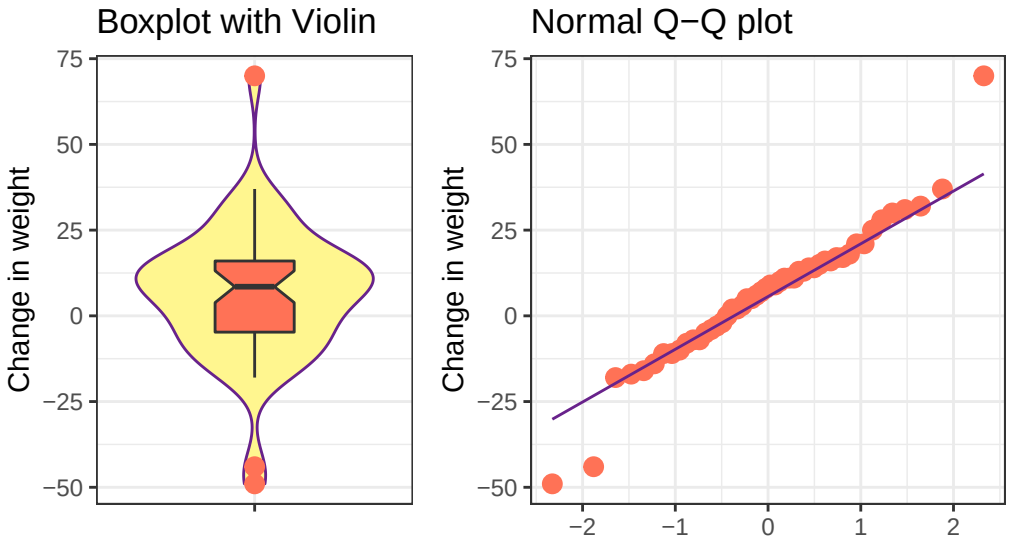
Suppose we compare the change in weight (before - after, in pounds) for 50 overweight male adult subjects who enter into a strict nutritional regimen, where subjects drink nothing other than water, and eat nothing but a variety of potatoes for two weeks, then spend four weeks eating only high-nutrition vegetables, and still drinking only water. An analyst prepares the output below, and you also have the `quiz2_wtchg` data file. We have decided in this case to use a 99% confidence level for our work.

Question 28 has three parts.

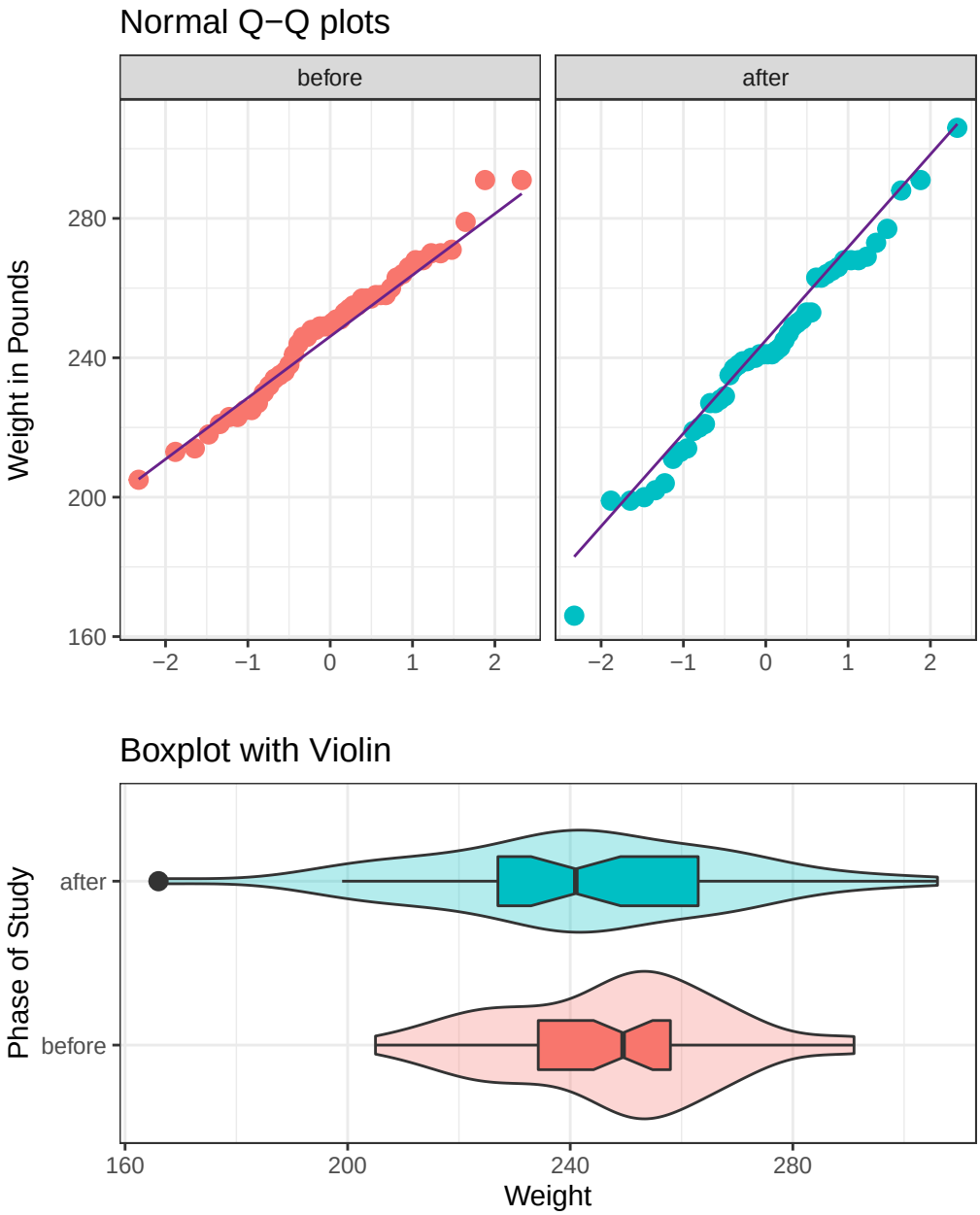
a. Specify which of the five Results prepared by the analyst and shown below (Result 1, 2, 3, 4 or 5) is most appropriate in this setting.
b. Write a sentence or two to tell us why your selection in Question 28a is the most appropriate choice.
c. Write a sentence or two to tell us what your selected result (from Question 28a) indicates about what can be said based on the available confidence intervals about the mean weight loss in a population represented by this sample, and in particular that population value's relationship to zero, justifying your decision by referring to the output provided.

A few plots of the data to help you get started

Plot 28A. Weight Change (Before − After diet) in pounds



Plot 28B. Weights Before and After Diet in pounds



Result 1 for Question 28

Method: One Sample t-test on `diffs`

estimate	p.value	conf.low	conf.high	conf.level
6.80	0.0116	-0.149	13.749	0.99

Result 2 for Question 28

Method: `Hmisc::smean.cl.boot` on `diffs`

mean	conf.low	conf.high	conf.level
6.800	0.059	13.402	0.99

Result 3 for Question 28

Method: Two Sample t-test for `before - after`

estimate	p.value	conf.low	conf.high	conf.level
6.80	0.134	-5.034	18.634	0.99

Result 4 for Question 28

Method: Welch Two Sample t-test for `before - after`

estimate	p.value	conf.low	conf.high	conf.level
6.80	0.134	-5.036	18.637	0.99

Result 5 for Question 28

Method: `bootdif` from `Love-boost.R` for `before - after`

mean	conf.low	conf.high	conf.level
6.80	-3.94	18.46	0.99

THIS IS THE END OF THE QUIZ