431 Quiz 1 for Fall 2024

Deadline: Wednesday 2024-10-16 at Noon

Thomas E. Love

2024-10-10

Table of contents

Instructions for Students 3

0.1 The Google Form Answer Sheet . . . 3

0.2 The Data Sets . . . 3

0.3 Getting Help . . . 4

0.4 When Should I Ask For Help? . . . 4

0.5 Scoring and Timing . . . 4

0.6 Writing Code into the Google Form . . . 5

0.7 R Packages and Love-431.R script . . . 5

1 Question 1 (4 points) 6

2 Question 2 (3 points) 6

3 Question 3 (4 points) 7

4 Question 4 (3 points) 8

5 Question 5 (3 points) 9

6 Question 6 (3 points) 9

7 Question 7 (4 points) 10

8 Question 8 (4 points) 11

9 Question 9 (4 points) 12

10 Question 10 (3 points) 13

11 Question 11 (4 points) **14**

12 Question 12 (4 points) **14**

13 Question 13 (3 points) **15**

14 Question 14 (3 points) **16**
Table for Question 14 16

15 Question 15 (3 points) **17**

16 Question 16 (4 points) **18**

17 Question 17 (3 points) **19**
Figure for Question 17 19

18 Question 18 (4 points) **20**

19 Question 19 (4 points) **21**

20 Question 20 (3 points) **21**

21 Question 21 (4 points) **22**

Setup for Questions 22-28 **23**

22 Question 22 (4 points) **23**

23 Question 23 (3 points) **23**

24 Question 24 (4 points) **24**

Setup for Questions 25-28 **25**
First Model: `fit25A` 25
Second Model: `fit25B` 27

25 Question 25 (4 points) **29**

26 Question 26 (3 points) **29**

27 Question 27 (4 points) **29**

28 Question 28 (4 points) **30**

This is the end of the Quiz. Congratulations! **30**

Session Information **30**

Instructions for Students

There are **28** questions on this Quiz and this PDF is **32** pages long. Be sure you have all **32** pages. It is to your advantage to answer all **28** questions. Your score is based on the number of correct responses, so there's no chance a blank response will be correct, and a guess might be, so you should definitely answer all of the questions.

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 (not even the teaching assistants.) You will be required to complete a short affirmation that you have obeyed these rules as part of submitting the Quiz.

0.1 The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, located at...

- https://bit.ly/431-2024-quiz1-form

All of your answers must be submitted through the Google Form by noon on Wednesday 2024-10-16, without exception. The form will close at 1 PM on that date, and no extensions will be made available, so do not wait until late in the morning on Wednesday to submit your work. We will only accept responses through the Google Form.

The Google Form contains places to provide your responses to each question, and a final affirmation where you'll type in your name to tell us that you followed the rules for the Quiz. You must complete that affirmation before you can submit your responses. When you submit your results (in the same way you submit a Minute Paper) you will receive an email copy of your submission, with a link that will allow you to edit your work.

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

0.2 The Data Sets

Along with the other materials, I have provided **three** data sets (called **fastfood.csv**, **hosp680.csv** and **oscar.csv**) that are mentioned in the Quiz. They may be helpful to you.

0.3 Getting Help

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

1. Ask your quiz question via email to **Thomas dot Love at case dot edu**.

During the Quiz period (5 PM 2024-10-10 through 1 PM 2024-10-16) we will not answer questions about the Quiz through Campuswire or in TA office hours or in person. Instead, Dr. Love will respond to questions sent to the email address listed above. We promise to respond to all questions received before 10 AM on 2024-10-16 in a timely fashion.

A few cautions:

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

0.4 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.

0.5 Scoring and Timing

16 of the 28 questions are worth **4** points, and the remaining 12 are worth **3** points, adding to a total of 100 points. Available points are specified at the start of each question. 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 hours. I expect most students will take 4-6 hours, and some will take as little as 2 or as many as 8. It is not a good idea to spend a long time on any one question.

0.6 Writing Code into the Google Form

Occasionally, we ask you to provide a single line of code. If not otherwise specified, a single line of code in response can contain **at most** two pipes, although you may or may not need the pipe in any particular setting. Note that I exclusively used the `|>` pipe, and not the `%>%` pipe, in developing this Quiz, but you may use either.

Moreover, you need not include the `library` command at any time for any of your code. Assume in all questions that all of the packages listed below have been loaded in R.

0.7 R Packages and Love-431.R script

This doesn't mean you need to use all of these packages, or indeed, that I used all of them in building the Quiz and its answer sketch.

```r
library(car)
library(Epi)
library(ggdist)
library(ggrepel)
library(glue)
library(infer)
library(janitor)
library(knitr)
library(MKinfer)
library(naniar)
library(patchwork)
library(rstanarm)
library(xfun)
library(easystats)
library(tidyverse)

source("data/Love-431.R")

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

Tip

Statistics is a "getting the details right" business. Read each question carefully, to be sure you get the details right in your answer.

1 Question 1 (4 points)

On our 431-data page, in the data subfolder, you'll find a file called `nnyfs.Rds`. Import that file into R, and then use it to specify answers to the following four questions:

a. How many variables are there in the `nnyfs` tibble?
b. Which variable has the most missingness of the variables in the `nnyfs` tibble?
c. How many missing observations exist in the variable you identified in part b of this question?
d. How many observations in the `nnyfs` tibble have complete data on all listed variables?

2 Question 2 (3 points)

Suppose you have a tibble with two variables. One is a factor called Exposure with levels High, Low and Medium, arranged in that order, and the other is a quantitative outcome. You want to rearrange the order of the Exposure variable so that you can then use it to identify for ggplot2 a way to split histograms of outcomes up into a series of smaller plots, each containing the histogram for subjects with a particular level of exposure (Low then Medium then High.)

Which of the pairs of `tidyverse` functions identified below should be used to accomplish such a plot?

a. `fct_reorder` and `facet_wrap`
b. `fct_relevel` and `facet_wrap`
c. `fct_collapse` and `facet_wrap`
d. `fct_reorder` and `group_by`
e. `fct_collapse` and `group_by`

3 Question 3 (4 points)

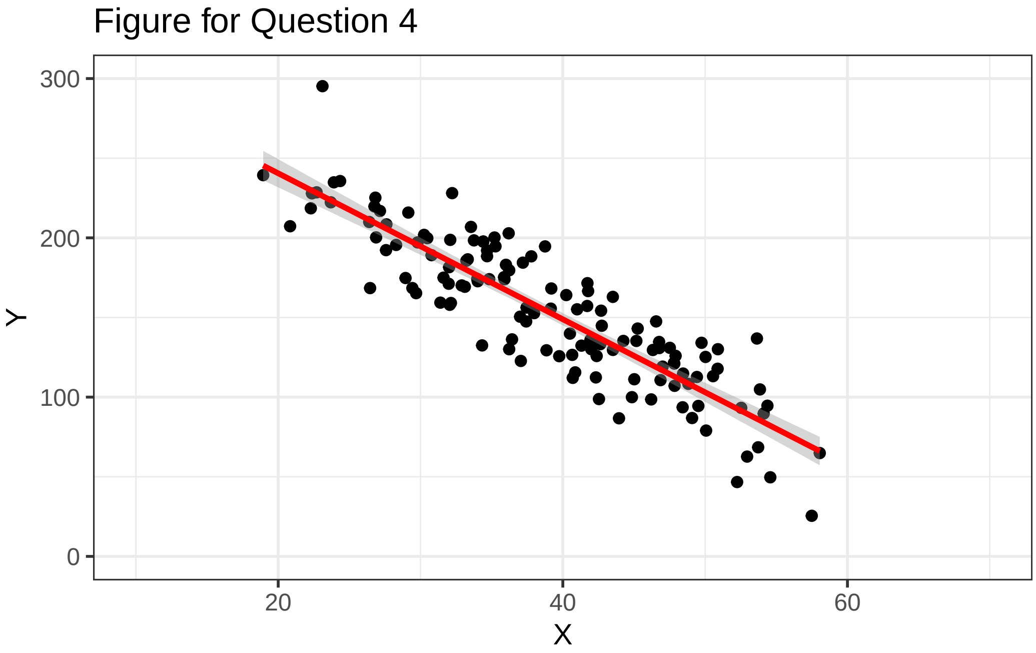
In the introduction and Chapters 1-4 of *The Art of Statistics*, David Spiegelhalter describes several common features of a strong visualization of data, including some identified by Alberto Cairo. Which of the following features are described as being characteristic of high-quality work in this regard? (CHECK ALL THAT APPLY.)

a. The graph's design is chosen so that relevant patterns become noticeable.
b. The graph contains reliable information.
c. The graph's appearance is presented in an attractive way.
d. The graph is honest, clear, and contains deep insights.
e. The graph should never connect data points gathered at different times.
f. The graph is accompanied by a meaningful title, and clear labels and captions.
g. The graph is of the raw data only.
h. The graph helps to raise more questions and encourage the reader to explore.

4 Question 4 (3 points)

Consider the following possible summaries of a linear model fit to predict Y from X, describing the scatterplot shown in the Figure for Question 4. As usual, the R-squared value is defined here as the square of the Pearson correlation coefficient between the outcome and the predictor. One of these summaries is correct. Which one?

a. Model: y = -4.6 + 332 x, with R-squared = 0.78
b. Model: y = -4.6 + 332 x, with R-squared = 0.28
c. Model: y = 4.6 + 332 x, with R-squared = -0.78
d. Model: y = 4.6 - 332 x, with R-squared = -0.28
e. Model: y = 4.6 + 332 x, with R-squared = 0.78
f. Model: y = 4.6 + 332 x, with R-squared = 0.28
g. Model: y = 332 + 4.6 x, with R-squared = 0.78
h. Model: y = 332 + 4.6 x, with R-squared = 0.28
i. Model: y = 332 - 4.6 x, with R-squared = -0.78
j. Model: y = 332 - 4.6 x, with R-squared = -0.28
k. Model: y = 332 - 4.6 x, with R-squared = 0.78
l. Model: y = 332 - 4.6 x, with R-squared = 0.28



5 Question 5 (3 points)

In this question, we consider data describing the age at onset (in years) for 23 women with a diagnosis of multiple sclerosis. The oldest age at onset was 46 years. The stem-and-leaf display shows the data for the first 23 subjects.

```
  The decimal point is 1 digit(s) to the right of the |

  1 | 46788889
  2 | 035679
  3 | 1123479
  4 | 26
```

If the next subject added to the data was 30 years of age at diagnosis, which of the following values will increase, as a result? (CHECK ALL VALUES THAT WILL INCREASE)

a. The mean
b. The standard deviation
c. The median

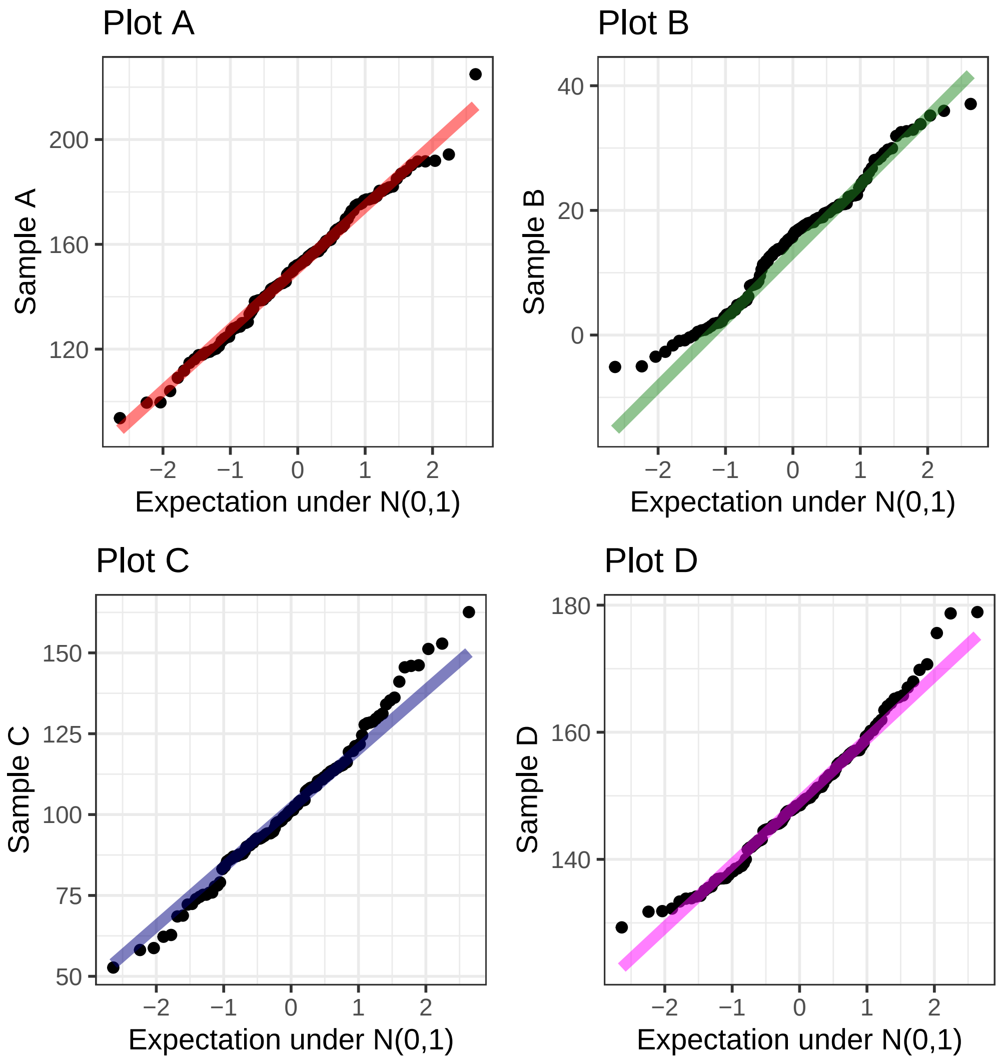
6 Question 6 (3 points)

There are four plots shown in the Figure for Question 6 on the next page. Each shows a Normal Q-Q plot describing a different set of 120 measurements. Which of the plots in the Figure for Question 6 shows data that could plausibly come from a Normal model with mean 150 and standard deviation 10?

(CHECK ALL APPROPRIATE RESPONSES)

a. Plot A
b. Plot B
c. Plot C
d. Plot D
e. None of the above.

Figure for Question 6



7 Question 7 (4 points)

Suppose that 90 of 175 applicants from students at private undergraduate institutions to a graduate school are accepted, while 135 of 325 from students at public undergraduate institutions are accepted. Estimate a two-sided 95% confidence interval for the relative risk of acceptance into graduate school for a "private undergrad" applicant as compared to a "public undergrad" applicant. Round your response to two decimal places, and provide both the point estimate and confidence interval.

8 Question 8 (4 points)

Each of the 500 applicants described in Question 7 applied to exactly one program at the graduate school: either Program A, B or C. Breaking down the applications, we find a few additional facts in addition to those specified in Question 7. In particular, we now also know that:

- Program A received 200 applications, and accepted 70.
- Program A accepted half of its 100 applicants who came from private schools.
- Program B also received 200 applications in total.
- Program B accepted 30 of its 50 applicants from private schools.
- Program C accepted 40 of its 75 applicants from public schools.
- Program C rejected 15 applicants from private schools.

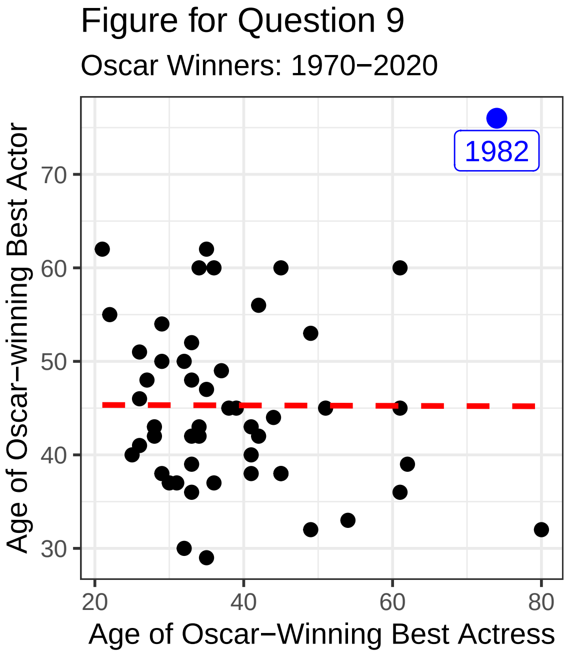
Which of the following statements are true? (CHECK ALL THAT APPLY.)

a. Students from private schools have lower odds of being accepted into Program A than do students from public schools.
b. Students from private schools have lower odds of being accepted into Program B than do students from public schools.
c. Students from private schools have lower odds of being accepted into Program C than do students from public schools.
d. None of these statements are true.
e. There is insufficient information to decide which of statements a-c are true.

9 Question 9 (4 points)

The data in the `oscar.csv` file I have provided to you describe the winners of the Academy Awards (also called the "Oscars") for Best Actor and Best Actress from 1970 to 2020.

The Figure for Question 9 is a scatterplot of 51 points, in each case displaying the age of the Best Actor (on the vertical, or y, axis) and the age of the Best Actress (on the horizontal, or x, axis) from the Academy Awards. Note that the Pearson correlation coefficient associated with these data is -0.003.



In 1982, Henry Fonda (age 76) and Katharine Hepburn (74) each won Oscars for the film *On Golden Pond*. This point is marked on the scatterplot by a blue dot, and labeled by its year. If the scatterplot were redrawn eliminating the 1982 awards, and including only the other 50 years, what would happen?

a. The slope of the linear model would DECREASE, and so would the model's R-squared.
b. The slope of the linear model would INCREASE, and so would the R-squared.
c. The slope of the linear model would DECREASE, and the R-squared would INCREASE.
d. The slope of the linear model would INCREASE, and the R-squared would DECREASE.
e. It is impossible to tell from the information provided.

10 Question 10 (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.

The entire report is available at https://www.nejm.org/doi/full/10.1056/NEJMoa1911614. 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.

Use a SAIFS procedure to 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. 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.

11 Question 11 (4 points)

Suppose you have collected data as part of a cohort study to look at the impact of exposure to an industrial solvent (which is stored in a four-level character variable called `solvent` which can be either none, modest, moderate or profound) on the probability of a bladder cancer diagnosis (stored as a three-level character variable called `diagnosis` which can be either definite, possible, or no.)

You can assume that a tibble containing these variables called `q11` is available to you in R, and that the packages loaded by Dr. Love at the start of this Quiz are also already loaded for you, so you don't need to load them again and there should be no `library()` calls in your response.

Provide a single line of R code (using 0-2 pipes) to obtain an appropriate summary of the relationship between the `solvent` and `diagnosis` variables in the `q11` tibble.

12 Question 12 (4 points)

Suppose now that in continuing your work on the study from Question 11, you now have more granular information on the exposure level to the solvent. Specifically, you now have an `exposure` measure, expressed as the percentage of the Occupational Safety and Health Administration (OSHA) recommended exposure limit, so that 100 = the recommended exposure limit for this solvent, and values above 100 indicate exposures that exceed that limit, while values below 100 indicate exposures that are at least somewhat "safe".

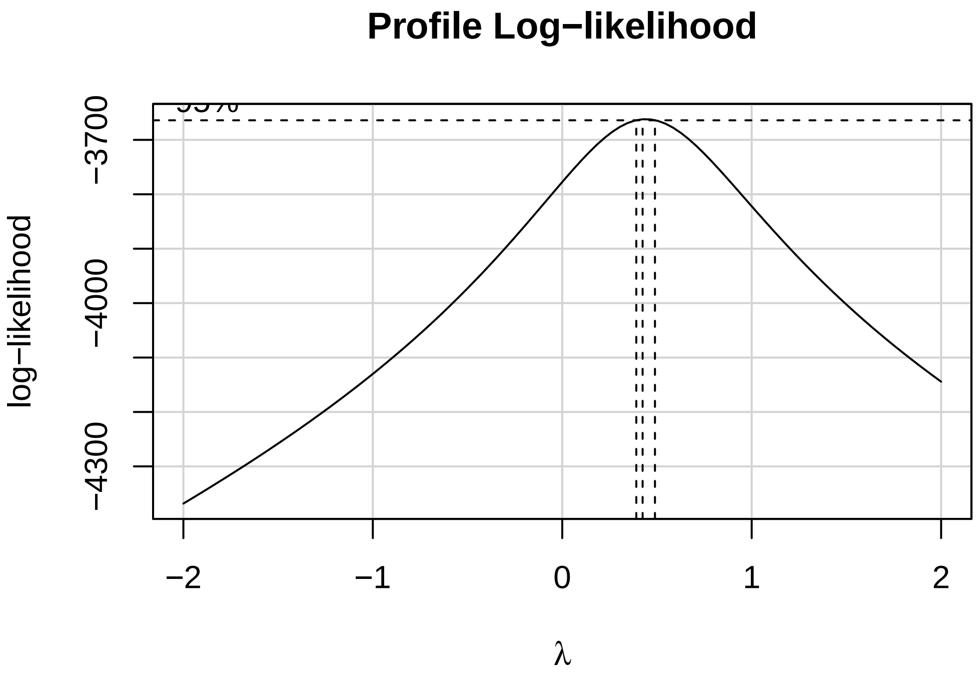
You can assume that a new tibble called `q12` is available to you in R containing this `exposure` measure as well as the bladder cancer `diagnosis` variable described in Question 11.

Again, you can also assume that the packages and scripts loaded by Dr. Love at the start of this Quiz are also already loaded for you, so you don't need to load them again.

Provide a single line of R code (again, you may use at most two pipes) to obtain an appropriate numerical summary description of the distribution of `exposure` within each `diagnosis` group in the `q12` tibble. This description should include, at least, the sample size, mean, median, standard deviation and MAD.

13 Question 13 (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.

14 Question 14 (3 points)

A new sample of 350 subjects ages 35-59 from the NHANES data generates the Table for Question 14, which summarizes the relationship between the subject's Self-Reported Overall Health (Excellent, Vgood = "Very Good", Good, Fair or Poor) and whether or not they have ever tried marijuana (Yes/No). In this sample, which group is more likely to report their Self-Reported Overall Health in one of the top three categories (Excellent, Very Good or Good)?

a. The "Yes" group, by more than three percentage points.
b. The "Yes" group, by 0.1 to 3 percentage points.
c. Neither group.
d. The "No" group, by 0.1 to 3 percentage points.
e. The "No" group, by more than three percentage points.
f. It is impossible to tell from the information provided.

Table for Question 14

	HealthGen					
Marijuana	Excellent	Vgood	Good	Fair	Poor	Total
No	13	43	54	22	5	137
Yes	20	96	67	25	5	213
Total	33	139	121	47	10	350

15 Question 15 (3 points)

The process of inductive inference, as described in *The Art of Statistics*, requires us to think hard about how we move from looking at the raw data to making general claims about the target population. Consider the following principles of effective measurement in this context.

I. We want to actually measure what we really want to measure without introducing systematic bias.

II. We want to sample at random whenever possible from the available subjects we are trying to make inferences about.

III. We want to use measures that give us a good chance of getting a similar result in a new study using the same measures.

Each of the principles listed above is associated primarily with a particular step in the process of building inductive inference. Identify the step (a, b or c) in the process associated with each of the statements (I, II or III) above.

a. Moving from the raw data to the sample
b. Moving from the sample to the study population
c. Moving from the study population to the target population

16 Question 16 (4 points)

Consider the `starwars` tibble that is part of the `dplyr` package in the tidyverse. Filter the data file to focus on individuals who are of the Human species, who also have complete data on both their height and mass. Then use a t-based approach to estimate an appropriate 90% confidence interval for the difference between the mean body-mass index of Human males minus the mean body-mass index of Human females, without assuming that the population variances of males and females are the same. The data provides `height` in centimeters and `mass` in kilograms. You'll need to calculate the body-mass index (BMI) values - the appropriate formula to obtain BMI in our usual units of $\frac{kg}{m^2}$ is:

$$\text{BMI} = \frac{10,000 * \text{mass in kg}}{(\text{height in cm})^2}$$

Specify your point estimate, and then the lower and upper bound, each rounded to a single decimal place, and be sure to specify the units of measurement.

⚠ Warning

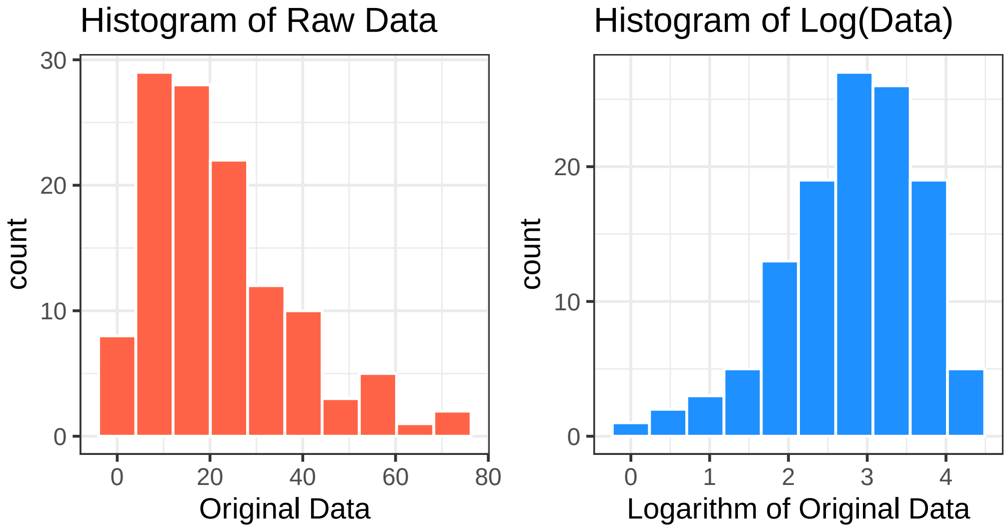
Be sure to read Question 16 very carefully to be sure you get **all** of the details right.

17 Question 17 (3 points)

Consider the two histograms shown in the Figure for Question 17. On the left, we show the original data set, in a red color. On the right, we show the natural logarithm of the data, in a blue color. Assuming you are unsatisfied with assuming a Normal distribution for each of these expressions of the data, what transformation would the ladder of power transformations recommend next, in an effort to re-express the data in a form that could be modeled effectively using a Normal distribution?

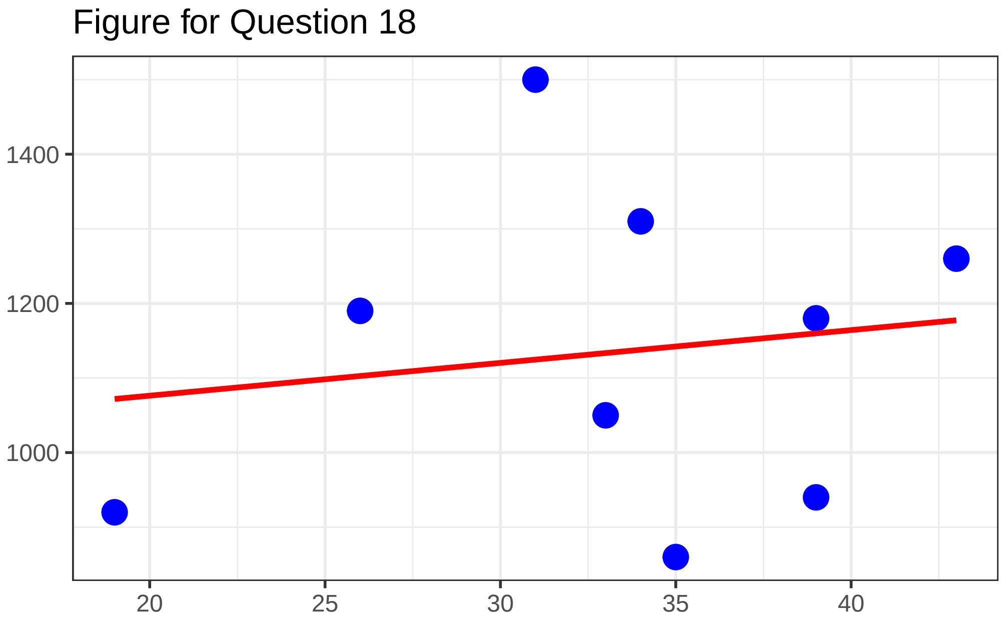
a. The square root of the data
b. The square of the data
c. The base 10 logarithm of the data
d. The inverse of the data
e. It is impossible to tell from the information provided

Figure for Question 17



18 Question 18 (4 points)

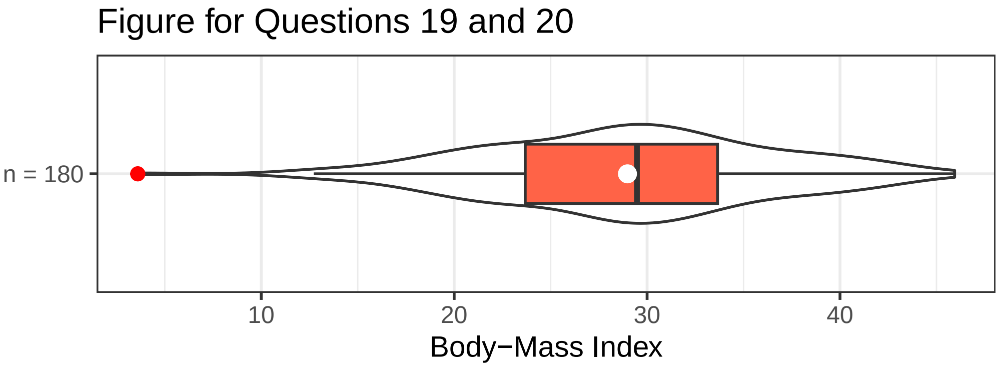
Fast food is often high in both fat and sodium. But are the two related? The scatter plot shown in the Figure for Question 18 describes the fat (in g) and sodium (in mg) contents of nine brands of hamburgers, and includes a linear model fit with geom_smooth, shown in red. I have provided the data in a file called `fastfood.csv`. In a sentence, what is the MOST IMPORTANT thing that should be done to improve the Figure?



19 Question 19 (4 points)

The Figure for Questions 19 and 20 displays the body-mass index (in $\frac{kg}{m^2}$) for 180 adults who suffer from rheumatoid arthritis, who are subjects in a new study. The mean BMI in the sample is 29 $\frac{kg}{m^2}$, the standard deviation is 7.3 $\frac{kg}{m^2}$ and there are no missing values. Which of the following statements are true? (CHECK ALL OF THE TRUE STATEMENTS.)

a. The distribution is substantially skewed and cannot be approximated well with a symmetric model.
b. The median is about 33 $\frac{kg}{m^2}$.
c. The IQR is about 10 $\frac{kg}{m^2}$.
d. The distribution cannot be approximated well with a Normal model because of outliers.
e. None of these statements are true.



20 Question 20 (3 points)

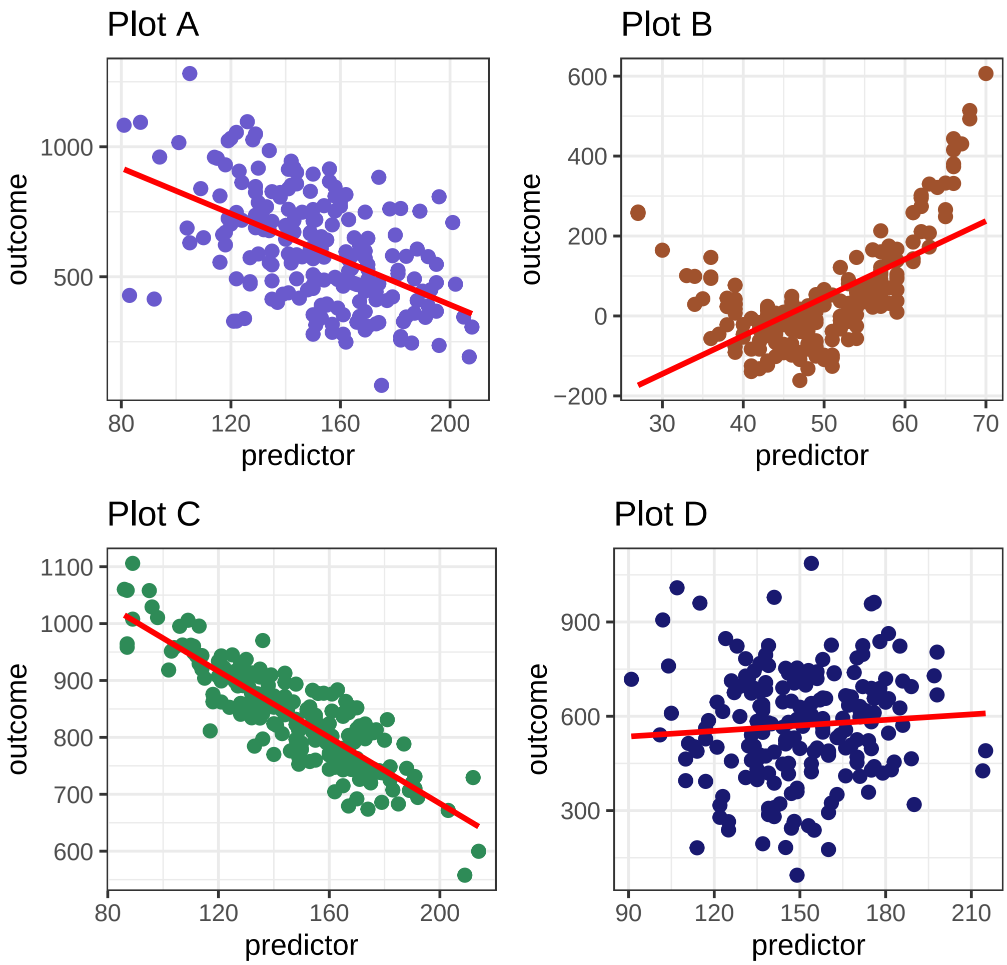
In the study of subjects with rheumatoid arthritis discussed in Question 19, adults with a BMI value of 25 or higher will be classified as overweight. Based on the Figure for Questions 19 and 20, how many of the 180 subjects would qualify as overweight using this standard?

a. Fewer than 45 subjects
b. Between 45 and 89 subjects
c. Exactly 90 subjects
d. Between 91 and 135 subjects
e. More than 135 subjects
f. There is insufficient information to answer the question.

21 Question 21 (4 points)

Consider the four scatterplots provided in the Figure for Question 21.

Figure for Question 21



Which of the four scatterplots in the Figure for Question 21 is associated with a linear model for `outcome` using `predictor` that has the largest R-square value?

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

Setup for Questions 22-28

The `hosp680.csv` data I have provided describe six characteristics of 680 simulated patients seen for out-patient care at a local hospital. Available variables in that data set are:

- `person` = Subject Identification Number (not a meaningful code)
- `sex` = the patient's sex (FEMALE or MALE)
- `statin` = does the patient have a prescription for a statin medication (YES or NO)
- `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)
- `sbp` = systolic blood pressure, in mm Hg, at the subject's most recent outpatient visit

These data are used in Questions 22-28 of the Quiz.

22 Question 22 (4 points)

Using the `hosp680` data, what is the 95% confidence interval for the odds ratio which compares the odds of receiving a statin if you are MALE divided by the odds of receiving a statin if you are FEMALE. Show the point and interval estimates, rounded to two decimal places. Do **NOT** use a Bayesian augmentation in responding to this question.

23 Question 23 (3 points)

Use the `hosp680` data I have provided to assess whether the type of insurance is strongly associated with our measure of educational attainment. Which of the following ranges captures the proportion of variation in `hsgrads` captured by a linear model using `insurance` as its predictor (the η^2 value) in this analysis?

a. Less than 0.1
b. 0.1 up to 0.249
c. 0.25 up to 0.499
d. 0.5 or more
e. I don't have enough information to answer the question

24 Question 24 (4 points)

Using the hosp680 data provided to you, with the code shown below, use a set of Holm-Bonferroni confidence intervals, maintaining an overall confidence level of 90%, to compare each of the pairs of mean sbp values across the various insurance types.

```r
hosp680 <-
  read_csv("data/hosp680.csv", show_col_types = FALSE) |>
  mutate(insurance = factor(insurance))

fit24 <- lm(sbp ~ insurance, data = hosp680)

estimate_contrasts(fit24, contrast = "insurance",
                   ci = 0.9, p.adjust = "holm")
```

Which of these statements is true, based on these intervals? (CHECK ALL OF THE TRUE STATEMENTS)

a. The confidence interval for the mean difference in systolic blood pressure between subjects with Commercial insurance minus those with Medicare insurance includes only positive values.
b. The confidence interval for the mean difference in systolic blood pressure between subjects with Medicaid insurance minus those with Medicare insurance includes only positive values.
c. The confidence interval for the mean difference in systolic blood pressure between subjects with Medicaid insurance minus those with Commercial insurance includes only positive values.
d. The confidence interval for the mean difference in systolic blood pressure between subjects with Medicaid insurance minus those with Uninsured insurance includes only positive values.
e. None of the statements above are true.

Setup for Questions 25-28

Using the hosp680 data, I obtained the output shown on the next few pages. Use this output to help you build responses to Questions 25-28.

```
hosp680 <- read_csv("data/hosp680.csv", show_col_types = FALSE)
```

First Model: fit25A

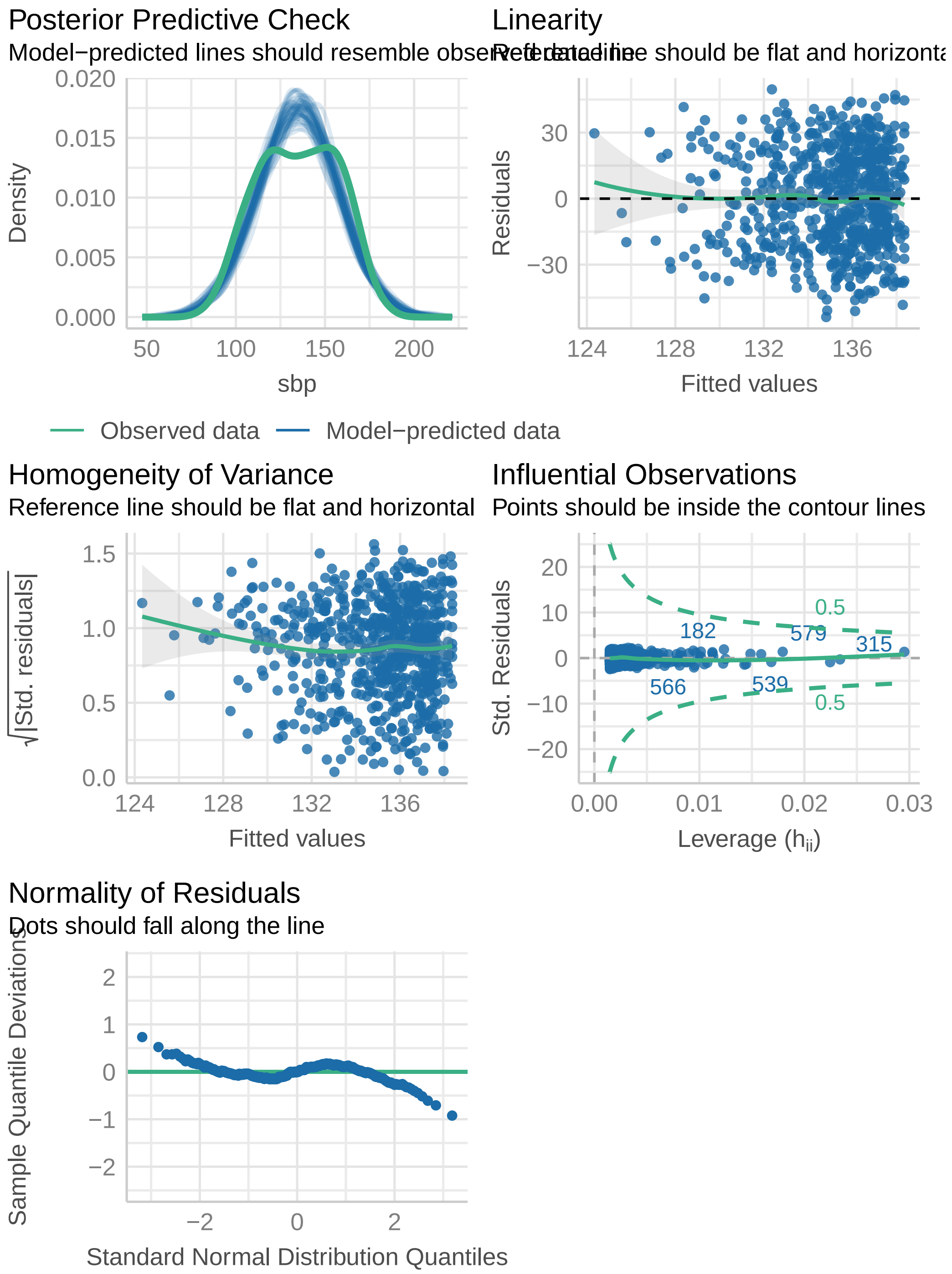
```
fit25a <- lm(sbp ~ hsgrads, data = hosp680)

model_parameters(fit25a, ci = 0.95)
```

```
Parameter   | Coefficient |   SE |           95% CI | t(678) |      p
--------------------------------------------------------------------
(Intercept) |      115.42 | 6.83 | [102.01, 128.83] |  16.90 | < .001
hsgrads     |        0.23 | 0.08 | [  0.07,   0.38] |   2.89 | 0.004

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

```
model_performance(fit25a, ci = 0.95)
```

```
# Indices of model performance

AIC      |     AICc |      BIC |    R2 | R2 (adj.) |   RMSE |  Sigma
---------------------------------------------------------------------
6141.315 | 6141.351 | 6154.882 | 0.012 |     0.011 | 22.028 | 22.061
```

```
check_model(fit25a)
```



Second Model: `fit25B`

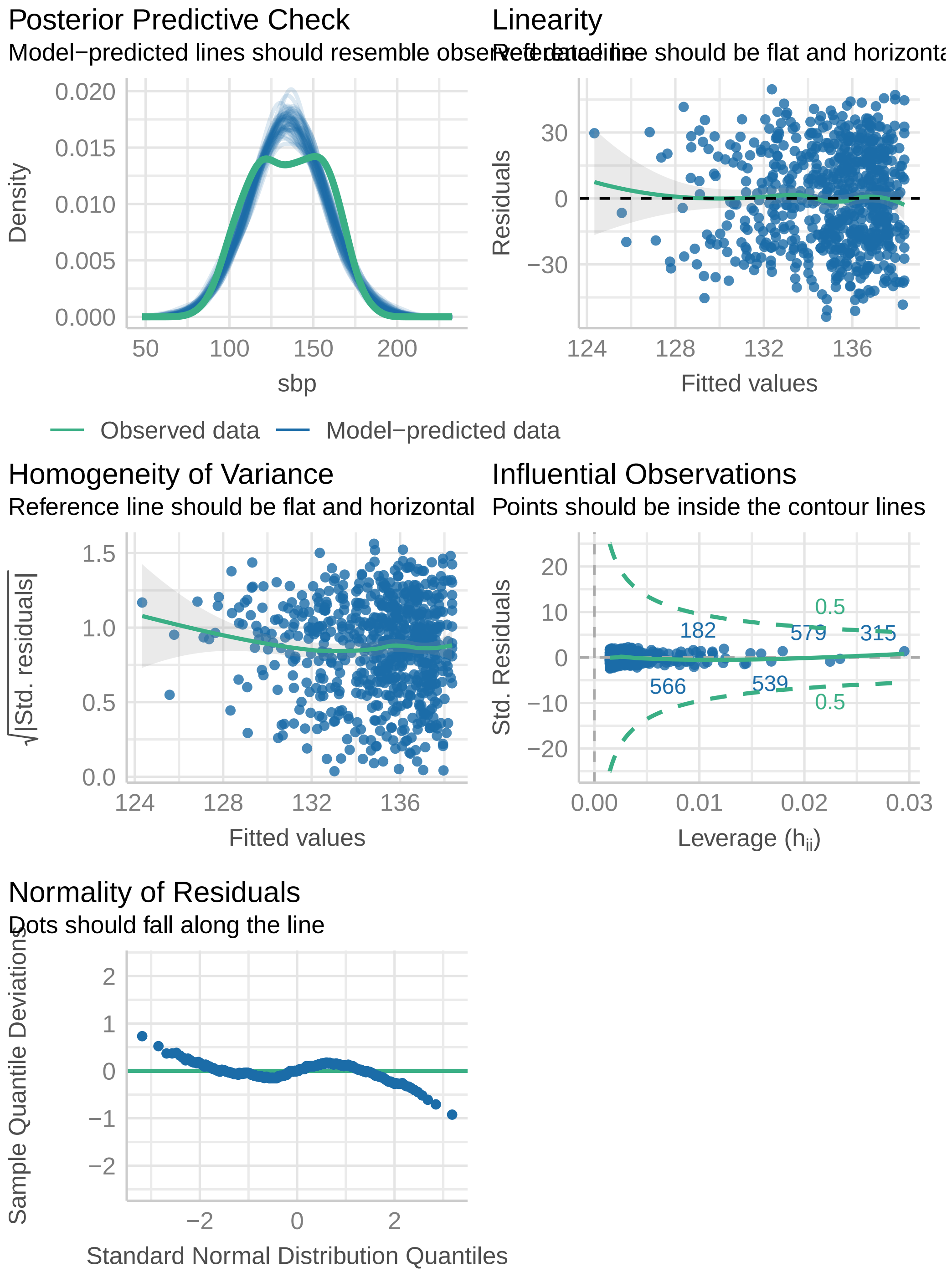
```
set.seed(25)
fit25b <- stan_glm(sbp ~ hsgrads, data = hosp680, refresh = 0)

model_parameters(fit25b, ci = 0.95)
```

```
Parameter   | Median |           95% CI |     pd |  Rhat |     ESS |                    Prio
--------------------------------------------------------------------------------------------
(Intercept) | 115.39 | [101.96, 128.83] |   100% | 1.000 | 4209.00 | Normal (135.03 +- 55.45
hsgrads     |   0.23 | [  0.08,   0.39] | 99.88% | 1.000 | 4241.00 |    Normal (0.00 +- 5.19
```

```
model_performance(fit25b, ci = 0.95)
```

```
# Indices of model performance

ELPD      | ELPD_SE |    LOOIC | LOOIC_SE |     WAIC |    R2 | R2 (adj.) |   RMSE |  Sigma
------------------------------------------------------------------------------------------
-3070.495 |  13.701 | 6140.991 |   27.402 | 6140.984 | 0.012 |     0.008 | 22.028 | 22.072
```

```
check_model(fit25b)
```



25 Question 25 (4 points)

In a single complete English sentence, state and interpret in context the point estimate of the slope of `hsgrads` in the `fit25a` model.

26 Question 26 (3 points)

According to the output I have provided, which of the models …

Rows

a. accounts for a larger proportion of variation in `sbp`?
b. indicates a better root mean squared error?
c. indicates a better residual standard deviation, as estimated by Sigma?

Columns

1. `fit25a`
2. `fit25b`
3. Neither model.

27 Question 27 (4 points)

On the basis of the output I have provided, which of the following statements are true? (CHECK ALL OF THE TRUE STATEMENTS)

a. The residual plots from the `fit25a` fit suggest a better match to the assumption of linearity than the plots from the `fit25b` fit.
b. In each model (`fit25a` and `fit25b`) there is a noticeable problem with highly influential outliers.
c. The Bayesian model shows a serious problem with the assumption of constant variance.
d. The `sbp` values appear to be highly skewed to the right.
e. None of statements `a` through `d` are true.

28 Question 28 (4 points)

In two or three sentences, describe your findings from the Posterior Predictive Check plots produced for each model (`fit25a` and `fit25b`.) Your response should specify a meaningful conclusion about the quality of fit for each model based on those plots, and describe what you see in the plots that motivates your conclusion. As part of your response, address whether we should prefer one of these fits over the other based solely on these plots, and if so, which one.

This is the end of the Quiz. Congratulations!

Session Information

```r
session_info()
```

```
R version 4.4.1 (2024-06-14 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

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

Package version:
  abind_1.4-8           arrangements_1.1.9    askpass_1.2.1
  backports_1.5.0       base64enc_0.1-3       bayesplot_1.11.1
  bayestestR_0.14.0     BH_1.84.0.0           bit_4.5.0
  bit64_4.5.2           blob_1.2.4            boot_1.3-31
  broom_1.0.7           bslib_0.8.0           cachem_1.1.0
  callr_3.7.6           car_3.1-3             carData_3.0-5
  cellranger_1.1.0      checkmate_2.3.2       cli_3.6.3
  clipr_0.8.0           cmprsk_2.2-12         coda_0.19-4.1
  codetools_0.2-20      colorspace_2.1-1      colourpicker_1.3.0
  commonmark_1.9.2      compiler_4.4.1        conflicted_1.2.0
  correlation_0.8.5     cowplot_1.1.3         cpp11_0.5.0
  crayon_1.5.3          crosstalk_1.2.1       curl_5.2.3
```

```
data.table_1.16.0      datasets_4.4.1         datawizard_0.13.0
DBI_1.2.3              dbplyr_2.5.0           Deriv_4.1.6
desc_1.4.3             digest_0.6.37          distributional_0.5.0
doBy_4.6.24            dplyr_1.1.4            DT_0.33
dtplyr_1.3.1           dygraphs_1.1.1.6       easystats_0.7.3
effectsize_0.8.9       emmeans_1.10.4         Epi_2.56
estimability_1.5.1     etm_1.1.1              evaluate_1.0.0
exactRankTests_0.8-35  fansi_1.0.6            farver_2.1.2
fastmap_1.2.0          fontawesome_0.5.2      forcats_1.0.0
foreach_1.5.2          Formula_1.2-5          fs_1.6.4
gargle_1.5.2           generics_0.1.3         ggdist_3.3.2
ggplot2_3.5.1          ggrepel_0.9.6          ggridges_0.5.6
glmnet_4.1-8           glue_1.8.0             gmp_0.7-5
googledrive_2.1.1      googlesheets4_1.1.1    graphics_4.4.1
grDevices_4.4.1        grid_4.4.1             gridExtra_2.3
gtable_0.3.5           gtools_3.9.5           haven_2.5.4
highr_0.11             hms_1.1.3              htmltools_0.5.8.1
htmlwidgets_1.6.4      httpuv_1.6.15          httr_1.4.7
ids_1.0.1              igraph_2.0.3           infer_1.0.7
inline_0.3.19          insight_0.20.5         isoband_0.2.7
iterators_1.0.14       janitor_2.2.0          jomo_2.7-6
jquerylib_0.1.4        jsonlite_1.8.9         knitr_1.48
labeling_0.4.3         later_1.3.2            lattice_0.22-6
lazyeval_0.2.2         lifecycle_1.0.4        lme4_1.1-35.5
loo_2.8.0              lubridate_1.9.3        magrittr_2.0.3
markdown_1.13          MASS_7.3-61            Matrix_1.7-0
MatrixModels_0.5.3     matrixStats_1.4.1      memoise_2.0.1
methods_4.4.1          mgcv_1.9-1             mice_3.16.0
miceadds_3.17-44       microbenchmark_1.5.0   mime_0.12
miniUI_0.1.1.1         minqa_1.2.8            mitml_0.4-5
mitools_2.4            MKdescr_0.8            MKinfer_1.2
modelbased_0.8.8       modelr_0.1.11          multcomp_1.4-26
munsell_0.5.1          mvtnorm_1.3-1          naniar_1.1.0
NHANES_2.1.0           nlme_3.1-164           nloptr_2.1.1
nnet_7.3-19            norm_1.0.11.1          numDeriv_2016.8-1.1
openssl_2.2.2          ordinal_2023.12.4.1    pan_1.9
parallel_4.4.1         parameters_0.22.2      patchwork_1.3.0
pbkrtest_0.5.3         performance_0.12.3     pillar_1.9.0
pkgbuild_1.4.4         pkgconfig_2.0.3        plyr_1.8.9
posterior_1.6.0        prettyunits_1.2.0      processx_3.8.4
progress_1.2.3         promises_1.3.0         ps_1.8.0
purrr_1.0.2            quadprog_1.5.8         quantreg_5.98
QuickJSR_1.4.0         R6_2.5.1               ragg_1.3.3
```

```
rappdirs_0.3.3         RColorBrewer_1.1.3     Rcpp_1.0.13
RcppArmadillo_14.0.2.1 RcppEigen_0.3.4.0.2    RcppParallel_5.1.9
readr_2.1.5            readxl_1.4.3           rematch_2.0.0
rematch2_2.1.2         report_0.5.9           reprex_2.1.1
reshape2_1.4.4         rlang_1.1.4            rmarkdown_2.28
rpart_4.1.23           rstan_2.32.6           rstanarm_2.32.1
rstantools_2.4.0       rstudioapi_0.16.0      rvest_1.0.4
sandwich_3.1-1         sass_0.4.9             scales_1.3.0
see_0.9.0              selectr_0.4.2          shape_1.4.6.1
shiny_1.9.1            shinyjs_2.1.0          shinystan_2.6.0
shinythemes_1.2.0      snakecase_0.11.1       sourcetools_0.1.7.1
SparseM_1.84.2         splines_4.4.1          StanHeaders_2.32.10
stats_4.4.1            stats4_4.4.1           stringi_1.8.4
stringr_1.5.1          survival_3.7-0         sys_3.4.3
systemfonts_1.1.0      tensorA_0.36.2.1       textshaping_0.4.0
TH.data_1.1-2          threejs_0.3.3          tibble_3.2.1
tidyr_1.3.1            tidyselect_1.2.1       tidyverse_2.0.0
timechange_0.3.0       tinytex_0.53           tools_4.4.1
tzdb_0.4.0             ucminf_1.2.2           UpSetR_1.4.0
utf8_1.2.4             utils_4.4.1            uuid_1.2.1
V8_5.0.1               vctrs_0.6.5            viridis_0.6.5
viridisLite_0.4.2      visdat_0.6.0           vroom_1.6.5
withr_3.0.1            xfun_0.48              xml2_1.3.6
xtable_1.8-4           xts_0.14.0             yaml_2.3.10
zoo_1.8-12
```