431 Quiz 3

Thomas E. Love

Version: 2021-11-19 02:11:14

Instructions for Students

There are **30** questions on this Quiz, and this PDF is **21** pages long. Be sure you have all **21** 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, and a guess might be, so you should definitely answer all of the questions.

The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, located at https://bit.ly/431-2021-quiz3-answer-sheet. All of your answers (including the essay for Question 30) must be submitted through the Google Form by **Monday 2021-11-29 at 9 PM**, without exception. The form will close at that time, and no extensions will be made available, so do not wait until the last moment to submit. We will not accept any responses except through the Google Form.

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

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.

Scoring and Timing

Most questions are worth 3 points. Questions 15 (6 points), 29 (5 points) and 30 (8 points) are worth more, summing to 100 points. The questions are not in any particular order except that Question 30 is an essay, which will likely take a bit longer than most of the other questions. The questions range in difficulty from "things I expect everyone to get right" to "things that are deliberately tricky".

The Quiz is meant to take about 6 hours. I expect most students will take 3-9 hours, and some will take as little as 2 or as many as 10. 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 our final class session on Thursday 2021-12-02.

IMPORTANT NOTES: On Writing Code

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.

Getting Help

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

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

1. You can ask your question in a private post on Piazza to the instructors. (This is the only kind of post you will be able to make on Piazza during the Quiz.)
2. You can ask your question via email to **431-help at case dot edu**.

While the complete Quiz is available, we will not answer questions about the Quiz except through the two approaches listed above. We promise to respond to all questions received before 5 PM on Monday 2021-11-29 at that time, if not sooner.

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.

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.

The Data Sets

In addition to the Love-boost.R script and the wc_code.R script mentioned in Question 11, we are providing six data sets that are mentioned in the Quiz. You will find them in our Shared Google Drive in the Quiz 3 folder. They may be helpful.

- wti_exam.csv, first mentioned in Question 01
- leopard.csv, first mentioned in Question 06
- rhino.csv, first mentioned in Question 14
- score_dat.csv, first mentioned in Question 16
- hospsim.csv, first mentioned in Question 17
- survA.csv, first mentioned in Question 25

To load wti_exam.csv, for example, I used

```
wti_exam <- read_csv("data/wti_exam.csv",
                     show_col_types = FALSE) %>%
    clean_names()
```

and I **strongly** encourage the use of similar code for loading each of the other files.

Twenty R Packages I loaded when building this document

This doesn't mean you need to use all 20 packages, and in fact I used exactly 11 of them to build the Quiz and its sketch. It does mean that I didn't use any other packages, **other than** the mosaic package.

```
library(broom)
library(car)
library(Epi)
library(equatiomatic)
library(fivethirtyeight)
library(GGally)
library(ggrepel)
library(glue)
library(Hmisc)
library(janitor)
library(knitr)
library(magrittr)
library(modelsummary)
library(naniar)
library(patchwork)
library(psych)
library(pwr)
library(sessioninfo)
library(visdat)
library(tidyverse)

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

theme_set(theme_bw())
```

1 Question 01 (3 points)

The lab component of a core course in biology is taught at the Watchmaker's Technical Institute by a set of five teaching assistants, whose names, conveniently, are Amy, Beth, Carmen, Donna and Elena. On the second examination of the semester (each section takes the same set of exams) an administrator at WTI wants to compare the mean scores across lab sections. She produces the following output in R.

```
wti_exam <- read_csv("data/wti_exam.csv",
                     show_col_types = FALSE) %>%
    clean_names()

result01 <- anova(lm(exam2 ~ ta, data = wti_exam))

tidy(result01) %>% kable(digits = 3)
```

term	df	sumsq	meansq	statistic	p.value
ta	4	1172.20	293.050	3.265	0.013
Residuals	165	14808.41	89.748	NA	NA

Emboldened by this result, the administrator decides to compare mean `exam2` scores for each possible pair of TAs, using a Bonferroni correction. If she wants to maintain an overall α level of 0.05 for the resulting suite of pairwise comparisons, and plans to do each of them separately with a two-sample t test, then what is the largest significance level she can use for each of the individual two-sample t tests?

a. A significance level of 0.20 on each test.
b. 0.05 on each test.
c. 0.005 on each test.
d. 0.0125 on each test.
e. None of these answers are correct.

2 Question 02 (3 points)

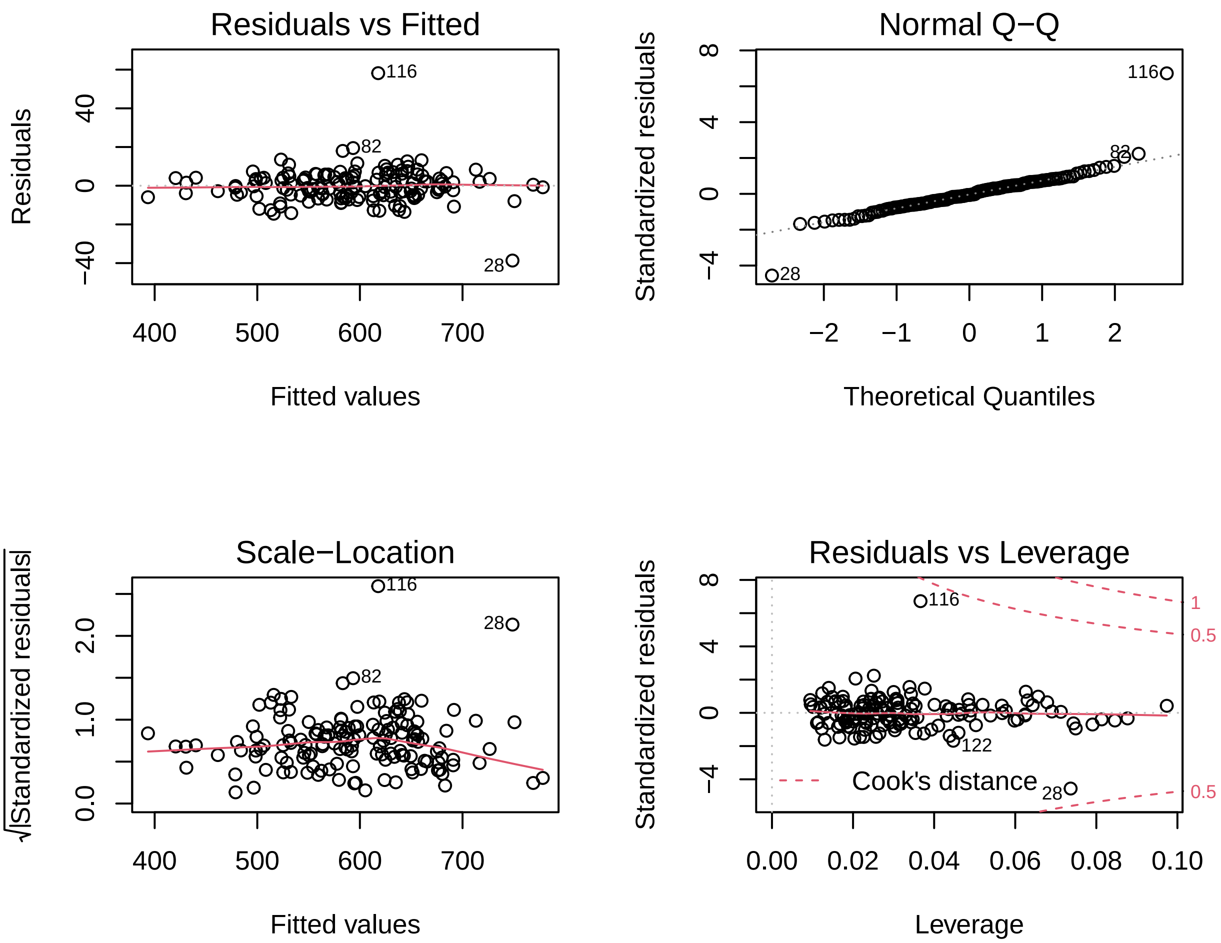
Suppose now that the administrator at the Watchmaker's Technical Institute that we mentioned in Question 01 instead used a Tukey HSD approach with a 95% family-wise confidence level. Note that when we refer in the responses below to Amy's scores, we mean the scores of students who were in Amy's lab section. Which conclusion of those presented below would be most appropriate?

a. Amy's scores were detectably higher than Donna's and Beth's.
b. Carmen's scores were detectably lower than everyone else's.
c. Donna's scores are detectably higher than Amy, Carmen or Elena.
d. Donna's scores are detectably lower than Amy, Carmen or Elena.
e. Donna's scores are detectably higher than Amy or Carmen.
f. Donna's scores are detectably lower than Amy or Carmen.
g. Elena's scores are detectably higher than Beth, Carmen or Donna.
h. None of these answers are correct.

3 Question 03 (3 points)

A regression model was developed to predict an outcome, `y`, based on a linear model using the four predictors `x1`, `x2`, `x3` and `x4`, in a sample of 150 subjects. We have **not** provided you with the data set for this Question, which I've called `dat03`.

```
m03 <- lm(y ~ x1 + x2 + x3 + x4, data = dat03)
par(mfrow=c(2,2)); plot(m03); par(mfrow=c(1,1))
```



```
vif(m03)
```

```
      x1       x2       x3       x4 
1.029302 1.005072 1.012276 1.031792 
```

Which of the following conclusions best describes the situation, based on the output provided?

a. Our main problem is with collinearity.
b. Our main problem is with the assumption of linearity.
c. Our main problem is with the assumption of constant variance.
d. Our main problem is with the assumption of normality.
e. We have no apparent problems with regression assumptions.

4 Question 04 (3 points)

The `Pottery` data are part of the `car` package in R, and we'll use those data in Questions 04 and 05. The data describe the chemical composition of ancient pottery found at four sites in Great Britain. For Question 04, we focus on the Na (Sodium) levels, and our goal is to compare the mean Na levels across the four sites. Run the following code.

Code for Question 04

```
data(Pottery)
anova(lm(Na ~ Site, data = Pottery))
```

Which of the following conclusions is most appropriate, based on the output generated by running the Code for Question 04?

a. The F test allows us to conclude that the population mean Na level in at least one of the four sites is different than the others, at a 1% significance level.
b. The F test allows us to conclude that the population mean Na level in each of the four sites is different than each of the others, at a 1% significance level.
c. The F test allows us to conclude that the population mean Na level is the same in all four sites, at a 1% significance level.
d. The F test allows us to conclude that the population mean Na level may not be the same in all sites, but is not statistically detectably different at the 1% level.
e. None of these conclusions are appropriate.

5 Question 05 (3 points)

On the next page, you'll find two visualizations Dr. Love generated to describe variables from the `Pottery` data set within the `car` package. Remember that we discussed these data in Question 04, as well.

Visualization 1 describes a variable Dr. Love has labeled `var1`, and Visualization 2 describes a variable labeled `var2`. Note that the white dots in the boxplots within each visualization are placed on the sample mean.

Based on this output, **and** whatever other work you deem appropriate, identify each of the variables I've visualized.

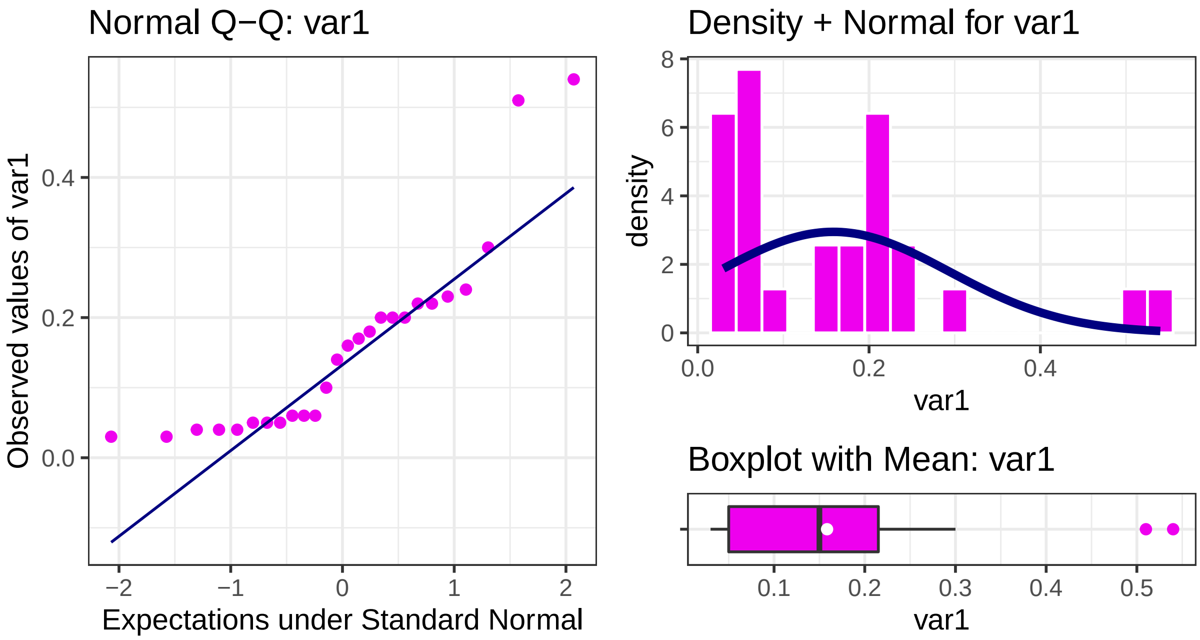
Rows:

1. Variable `var1`.
2. Variable `var2`.

Columns:

a. Calcium (`Ca`)
b. Iron (`Fe`)
c. Magnesium (`Mg`)
d. Sodium (`Na`)
e. Aluminum (`Al`).
f. We cannot determine.

Visualization 1 for Question 05: var1

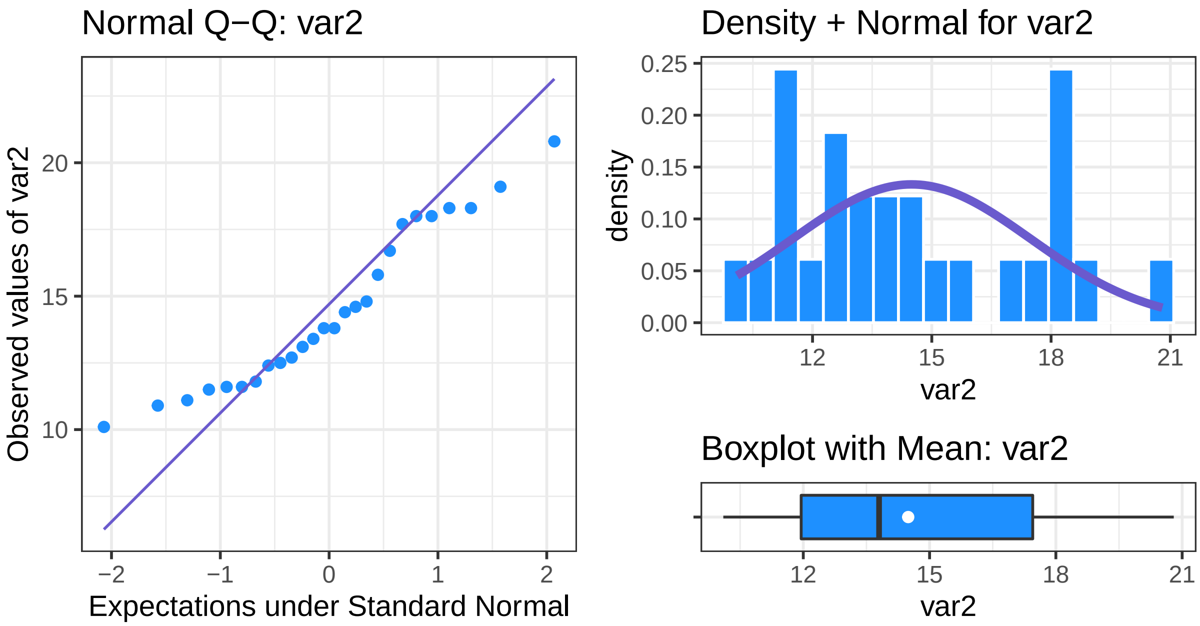


Normal Q−Q: var1

Density + Normal for var1

Boxplot with Mean: var1

Visualization 2 for Question 05: var2



Normal Q−Q: var2

Density + Normal for var2

Boxplot with Mean: var2

Background for Questions 06 - 10

In Questions 06-10, we will work with a data set (`leopard.csv`) I have provided which will let you predict a measure of predatory behavior in leopards using a square root transformation. The available data includes the outcome (`behavior`) which is measured on a scale from 0 to 125 with higher scores indicating greater levels of predatory behavior, as well as a set of six quantitative predictors, cleverly labeled `x1` through `x6`. The `animal` code I have provided uniquely identifies each of the 300 leopards included in the (fictional) data.

6 Question 06 (3 points)

Suppose you fit four candidate models to predict our measure of predatory behavior in leopards, while using a square root transformation for the outcome (`behavior`). The four models are nested, in that D is a proper subset of C, which is a proper subset of B, which is a proper subset of A. Specifically, Model A contains all six predictors, Model B contains five of those six predictors, and Model C contains four of the five Model B predictors, while Model D uses two of the predictors in Model C. Here are the predictors in each model:

Model	Predictors
A	x1, x2, x3, x4, x5 and x6
B	x1, x2, x3, x4 and x5
C	x1, x2, x3 and x4
D	x1 and x2

```
leopard <- read_csv("data/leopard.csv",
                    show_col_types = FALSE) %>%
    clean_names()

modelA <- lm(sqrt(behavior) ~
               x1 + x2 + x3 + x4 + x5 + x6, data = leopard)
modelB <- lm(sqrt(behavior) ~
               x1 + x2 + x3 + x4 + x5, data = leopard)
modelC <- lm(sqrt(behavior) ~
               x1 + x2 + x3 + x4, data = leopard)
modelD <- lm(sqrt(behavior) ~
               x1 + x2, data = leopard)
```

Of the four candidates listed in the columns, which model does each criterion (specified in the rows) suggest will be the best choice for predicting the square root of the `behavior` variable?

Columns:

1. Model A
2. Model B
3. Model C
4. Model D

Rows:

a. Multiple R^2
b. Akaike Information Criterion (AIC)
c. Adjusted R^2
d. Bayes Information Criterion (BIC)

7 Question 07 (3 points)

Code for Question 07

```
round(car::vif(modelA),3)
car::powerTransform(modelA)
anova(modelA, modelB)
```

Run the three lines of code provided in the Code for Question 07 above, where `modelA` refers to the model A you fit in Question 06, and `modelB` refers to the model B you fit in Question 06.

Which of the following statements are reasonable conclusions from the output you obtain by running this code? **CHOOSE ALL CORRECT RESPONSES.**

a. Model A displays no sign of meaningful collinearity.
b. Model A has a serious problem with collinearity.
c. Model A's residuals show a serious problem with independence.
d. Model A's residual variance is larger than the residual variance of Model B.

8 Question 08 (3 points)

Which of the following R commands would calculate fitted values of `sqrt(behavior)` using the equation in Model D for a new set of data contained in the `newleopard` tibble? **CHOOSE ALL CORRECT RESPONSES.**

a. `augment(modelD, newdata = newleopard)`
b. `broom(modelD, newdata = newleopard)`
c. `fit(modelD, newdata = newleopard)`
d. `glance(modelD, newdata = newleopard)`
e. `predict(modelD, newdata = newleopard)`
f. `split(modelD, newdata = newleopard)`
g. `tidy(modelD, newdata = newleopard)`
h. None of the above responses.

9 Question 09 (3 points)

Suppose the first predicted subject in the `newleopard` tibble yields a prediction of `sqrt(behavior)` of 7, with a 90% uncertainty interval of (5, 9). To convert that uncertainty interval back to the original scale on which the predatory behavior measurements were obtained, we would obtain which of the following results?

a. 5 to 9
b. `log(5)` to `log(9)`
c. `10*(5)` to `10*(9)`
d. `exp(5)` to `exp(9)`
e. `5^2` to `9^2`
f. None of these methods would work.

10 Question 10 (3 points)

Create the main four residual plots for Model B from Question 06. Which of the following conclusions is most appropriate, based on an analysis of your residual plots for Model B?

a. There is a serious problem with the assumption of linearity.
b. There is a serious problem with the assumption of constant variance.
c. There is a serious problem with the assumption of Normality.
d. There are no serious problems evident in the residual plots.
e. None of these conclusions are appropriate.

11 Question 11 (3 points)

Consider the `weather_check` data frame within the `fivethirtyeight` package. We will use these data for Questions 11-13.

Suppose you want to build a table containing information from the `female`, `ck_weather` and `age` variables in that data frame. I suggest you use the following approach to place the data in the `wc` tibble, and adjust some of the coding.

Note I have provided this code snippet to you in a file called `wc_code.R` on our Shared Google Drive in the Quiz 3 folder.

```
wc <- fivethirtyeight::weather_check %>%
  select(female, ck_weather, age) %>%
  rename(sex = female) %>%
  mutate(sex = fct_recode(factor(sex),
                           "Female" = "TRUE",
                           "Male" = "FALSE"),
         ck_weather = fct_recode(factor(ck_weather),
                           "Check" = "TRUE",
                           "No Check" = "FALSE")) %>%
  mutate(sex = fct_relevel(sex, "Female"),
         ck_weather = fct_relevel(ck_weather, "Check"))
```

Build the specified table using your `wc` tibble. Which age group has exactly 105 female respondents who answered yes to the question "Do you typically check a daily weather report?"

a. Ages 18-29
b. Ages 30-44
c. Ages 45-59
d. Ages 60+
e. None of these.

12 Question 12 (3 points)

How many subjects in the `wc` tibble created in Question 11 have no missing data on any of the variables in that tibble?

13 Question 13 (3 points)

Again, use the `wc` tibble you developed in Questions 11 and 12. Report a point estimate and 90% confidence interval for the Female - Male difference in probability of "typically checking a daily weather report." Develop your estimate using only the subjects you identified in Question 12 as having no missing data on any of the variables in the `wc` tibble.

- Round your response to three decimal places, and report the probability difference in terms of a proportion, rather than as a percentage.
- Do not use a Bayesian augmentation in your response to Question 13.

THE QUIZ CONTINUES ON THE NEXT PAGE

Background for Questions 14-16

The southern white rhinoceros (*Ceratotherium simum simum*) has a wild population exceeding 20,000, making them the most abundant rhino species in the world, according to Wikipedia. Your outcome of interest is a measure of size. Suppose that you want to compare those living in an area of southern Africa subject to serious problems from poaching (you have data on 40 rhinos living near Watering Hole A) and those living in an area of southern Africa more than 1,000 km away with a less serious poaching problem (you have data on 40 rhinos living near Watering Hole B). Your interest is to understand how exposure to poaching is associated with average rhino size.

The `rhino.csv` data I have provided is, alas, simulated, and contains two columns of outcome information, one with the size data for each of the 40 rhinos living near watering Hole A, and the other providing the size for each of the 40 rhinos living near Hole B.

14 Question 14 (3 points)

Which of the following approaches to analyzing these data is appropriate?

a. A comparison of proportions from two independent samples.
b. A comparison of means from two paired samples.
c. A comparison of means from two independent samples.
d. A chi-square analysis of a contingency table.
e. A linear model with multiple predictors.

15 Question 15 (6 points)

Manage the `rhino` data I've provided to you within R (this may involve pivoting the data) and then make the comparison you identified in Question 14.

Write a **short** description[1] of your results providing:

- the statistical method you used to build your estimate (do not specify the code you used here but describe what you did),
- the point estimate and endpoints of an appropriate 90% confidence interval for an appropriate summary of the Hole A - Hole B difference in the size outcome, rounded to two decimal places.
- a clear statement about the direction of the effect you observed, and
- a clear statement about whether the effect you observed meets the standard for a statistically detectable effect with 90% confidence.

[1] Dr. Love's sketch for Question 15 uses 51 words and less than 300 characters (including spaces) in two sentences to address each of the elements listed here. So three or four sentences should suffice for your response (and the box on the Google Form will take up to 750 characters.) Just type your response into the box provided, using complete English sentences.

16 Question 16 (3 points)

Suppose we're using the data contained in the `score_dat.csv` file I've provided to build a model for an outcome called `score` using four predictors (`x1`, `x2`, `x3` and `x4`), gathered for a sample of 100 subjects. What transformation of our response does a Box-Cox analysis for the model for `score` containing the main effects of the four predictors suggest?

a. The inverse of our outcome, $1/score$.
b. The square root of our outcome, $\sqrt{score}$.
c. The logarithm of our outcome, $log(score)$.
d. The square of our outcome, $score^2$.
e. The original, untransformed outcome, $score$.

Background for Questions 17-22

For Questions 17-22, consider the data I have provided in the `hospsim.csv` file, which describe 750 patients at a metropolitan hospital. They are simulated. Available are:

- `subject.id` = Subject Identification Number (not a meaningful code)
- `age` = the patient's age, in years (all subjects are between 21 and 75)
- `sex` = the patient's sex (FEMALE or MALE)
- `a1c` = the patient's hemoglobin A1c level (in %)
- `ldl` = the patient's LDL cholesterol level (in mg/dl)
- `sbp` = the patient's systolic blood pressure (in mm Hg)
- `bmi` = the patient's body mass index (in kg/square meter)
- `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)
- `clinic.type` = whether the patient goes to a newly built clinic or an old clinic

17 Question 17 (3 points)

Using the `hospsim` data, what is the point estimate and 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. Do **NOT** use a Bayesian augmentation here, and round your answers (for the point estimate and each endpoint) to two decimal places.

18 Question 18 (3 points)

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 is significant at the 5% significance level, and a Tukey HSD comparison retaining a 5% family-wise significance level reveals that Medicare shows detectably higher education levels than Uninsured.
c. The ANOVA F test is significant at the 5% significance level, and a Tukey HSD comparison retaining a 5% family-wise significance level reveals that Medicaid's education level is detectably lower than either Medicare or Commercial.
d. The ANOVA F test is significant at the 5% significance level, and a Tukey HSD comparison retaining a 5% family-wise significance level reveals that Uninsured's education level is detectably lower than Commercial or Medicare.
e. None of these conclusions is appropriate.

19 Question 19 (3 points)

Build a model to predict LDL cholesterol using 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 5% significance 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.

20 Question 20 (3 points)

Run a backwards elimination stepwise procedure starting with the original nine regression inputs (`clinic.type`, `age`, `sex`, `insurance`, `hsgrads`, `a1c`, `bmi`, `sbp` and `statin`). Compare your initial "kitchen sink" model with all 9 inputs to the model generated by the stepwise approach using adjusted R^2, AIC and BIC. What conclusions can you draw?

The smaller model (stepwise result) is ...

a. better on adjusted R^2, worse on AIC and worse on BIC.
b. better on AIC, worse on BIC and adjusted R^2
c. better on AIC and BIC, and worse on adjusted R^2
d. worse on all three measures
e. better on all three measures

21 Question 21 (3 points)

Now build a model using sex and insurance type (and that's it) to predict hemoglobin A1c. In this new model, identify the subject with the largest residual. 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.

22 Question 22 (3 points)

You are comparing two regression models for the same outcome, which you built using a training sample of data. You then use each model to predict data in a test sample, that was not used to calculate the original regression equations. Which of the following summaries will be useful to you in assessing which model does a better job of out-of-sample prediction?

CHECK ALL RESPONSES THAT APPLY.

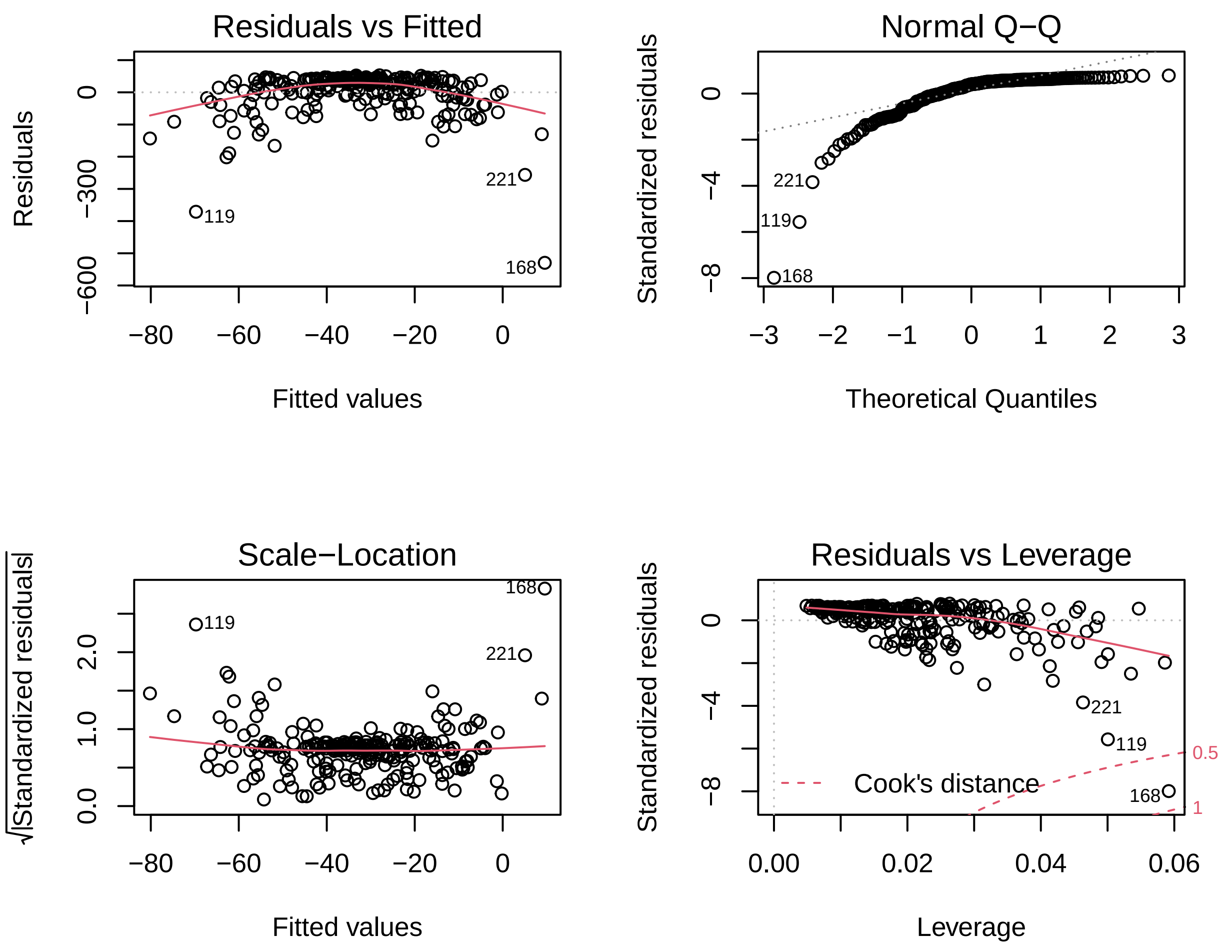
a. The square root of the mean of the squared prediction errors
b. The mean of the absolute values of the prediction errors
c. The maximum of the absolute values of the prediction errors
d. The adjusted R^2 statistic from the training sample
e. The squared correlation of the fitted and actual values in the test sample
f. None of these measures would be useful to you

THE QUIZ CONTINUES ON THE NEXT PAGE

23 Question 23 (3 points)

A regression model was developed to predict an outcome, `y`, based on a linear model using four predictors in a sample of 230 subjects. Here is the relevant R output.

```
m23 <- lm(y ~ x1 + x2 + x3 + x4, data = dat23)
par(mfrow=c(2,2)); plot(m23); par(mfrow=c(1,1))
```



```
vif(m23)
```

```
      x1       x2       x3       x4 
1.006073 1.006607 1.007781 1.004958 
```

Which of the following conclusions best describes this situation, based on the output?

a. The first thing we should try is dropping a highly influential point.
b. The first thing we should try is transforming the outcome to improve linearity.
c. The first thing we should worry about is the assumption of constant variance.
d. The first thing we should try is to focus on the problem of collinearity.
e. We have no apparent problems with the assumptions of linear regression.

24 Question 24 (3 points)

You have a tibble called `mydat` that contains 500 observations on 1 outcome and 5 predictors. Which of the following codes would most appropriately split the data into a test sample (called `mydat_test`) containing 20% of the observations, and a training sample (called `mydat_train`) containing the rest? **CHECK ALL CORRECT RESPONSES.**

a. `mydat_test <- slice_sample(mydat, prop = 0.80)` and `mydat_train <- anti_join(mydat, mydat_test)`
b. `mydat_test <- slice_head(mydat, 100)` and `mydat_train <- anti_join(mydat, mydat_test)`
c. `mydat_test <- partition(mydat, 400:100)` and `mydat_train <- anti_join(mydat, mydat_test)`
d. `mydat_test <- slice_sample(mydat, n = 100)` and `mydat_train <- anti_join(mydat, mydat_test)`
e. `mydat_train <- slice_sample(mydat, prop = 0.80)` and `mydat_test <- anti_join(mydat, mydat_train)`
f. None of these approaches would work.

Background for Questions 25-26

Consider the responses to the item “How important do you think statistics will be in your future career?” as gathered in the `statfuture` variable within the `survA.csv` data set I have provided to you. These data are drawn from the quick survey taken at the start of the semester by your class (and past classes) taking 431. The `year` variable (your `year` is 2021) indicates the year (Fall) in which the respondents took 431. Note that there were five years in the `survA` data where the class had no missing responses on this item.

25 Question 25 (3 points)

Rounding your responses to three decimal places, specify a point estimate and 90% confidence interval for the true value of the proportion of subjects who responded with the value 7 (meaning extremely important), using the approach due to Agresti and Coull. Use only the data from the five years where the class had no missing responses on the `statfuture` item to create your estimates.

26 Question 26 (3 points)

Is the sample proportion of “7” values from the 2021 survey (ignoring missing values completely) inside the interval you reported in Question 26?

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
d. It is impossible to tell from the information provided.

Background for Questions 27 and 28

The data in the table below are from a pilot study to 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

In questions 27 and 28, 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.

Consider the following output:

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

pilot <- tibble(subj, pre, post)

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

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

```r
mosaic::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
```

```r
mosaic::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
```

27 Question 27 (3 points)

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.

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.

What is the smallest number of subjects the new study will have to include?

28 Question 28 (3 points)

Provide your R code (a single line will suffice) to justify your response to Question 27.

29 Question 29 (5 points)

Livestock are given a special feed supplement to see if it will promote weight gain. The researchers report that the 77 cows studied gained an average of 56 pounds, and that a 95% confidence interval for the mean weight gain this supplement produced has a margin of error of +/- 11 pounds. Several students wrote conclusions about this situation, which are listed below. For each statement, please indicate whether the statement is a correct interpretation of this confidence interval.

Rows:

a. 95% of the cows studied gained between 45 and 67 pounds.
b. We're 95% confident that a cow fed this supplement will gain between 45 and 67 pounds.
c. We're 95% confident that the average weight gain among the cows included in this study was between 45 and 67 pounds.
d. The average weight gain of cows fed this supplement will be between 45 and 67 pounds 95% of the time.
e. If this supplement is tested on another sample of cows, there is a 95% chance that their average weight gain will be between 45 and 67 pounds.

Columns:

1. Appropriate interpretation of the CI
2. Not appropriate.

30 Question 30 (8 points)

Write a clear and well-composed essay of 200 to 300 words describing an important idea from David Spiegelhalter's *The Art of Statistics* about doing statistical science well that Dr. Love didn't cover in Classes 1-24. Your essay should state the idea in your own words, and should indicate why you feel it is important.

If you quote Spiegelhalter (and we prefer that you do), specify the Chapter containing your quote. If Dr. Love discussed your idea in class, you'll lose 1 of 8 available points. If your essay is unclear, or if you miss Spiegelhalter's point, that will have a bigger impact on your score. Each Chapter in Spiegelhalter includes a summary of key points. Feel free to use these summaries to help spark ideas, but do not quote the summaries.

Use any software to produce your Word document, Google Doc, or PDF submission.

These instructions are 150 words long.

Session Information (Dr. Love's Setup)

```
sessionInfo()
```

```
R version 4.1.2 (2021-11-01)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19043)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] ggridges_0.5.3       mosaicData_0.20.2     ggformula_0.10.1
 [4] ggstance_0.3.5       Matrix_1.3-4          forcats_0.5.1
 [7] stringr_1.4.0        dplyr_1.0.7           purrr_0.3.4
[10] readr_2.0.2          tidyr_1.1.4           tibble_3.1.5
[13] tidyverse_1.3.1      visdat_0.5.3          sessioninfo_1.2.1
[16] pwr_1.3-0            psych_2.1.9           patchwork_1.1.1
[19] naniar_0.6.1         modelsummary_0.9.4    magrittr_2.0.1
[22] knitr_1.36           janitor_2.1.0         Hmisc_4.6-0
[25] Formula_1.2-4        survival_3.2-13       lattice_0.20-45
[28] glue_1.4.2           ggrepel_0.9.1         GGally_2.1.2
[31] ggplot2_3.3.5        fivethirtyeight_0.6.2 equatiomatic_0.3.0
[34] Epi_2.44             car_3.0-12            carData_3.0-4
[37] broom_0.7.10

loaded via a namespace (and not attached):
 [1] colorspace_2.0-2    ellipsis_0.3.2      leaflet_2.0.4.1
 [4] snakecase_0.11.0    htmlTable_2.3.0     ggdendro_0.1.22
 [7] base64enc_0.1-3     fs_1.5.0            rstudioapi_0.13
[10] farver_2.1.0        bit64_4.0.5         fansi_0.5.0
[13] lubridate_1.8.0     xml2_1.3.2          mosaic_1.8.3
[16] splines_4.1.2       mnormt_2.0.2        polyclip_1.10-0
[19] jsonlite_1.7.2      cluster_2.1.2       dbplyr_2.1.1
[22] png_0.1-7           ggforce_0.3.3       shiny_1.7.1
[25] compiler_4.1.2      httr_1.4.2          backports_1.3.0
[28] assertthat_0.2.1    fastmap_1.1.0       cli_3.1.0
[31] tweenr_1.0.2        later_1.3.0         htmltools_0.5.2
[34] tools_4.1.2         gtable_0.3.0        tables_0.9.6
[37] Rcpp_1.0.7          cellranger_1.1.0    vctrs_0.3.8
[40] nlme_3.1-153        crosstalk_1.2.0     xfun_0.27
[43] rvest_1.0.2         mosaicCore_0.9.0    mime_0.12
[46] lifecycle_1.0.1     MASS_7.3-54         zoo_1.8-9
[49] scales_1.1.1        vroom_1.5.5         hms_1.1.1
```

```
[52] promises_1.2.0.1    parallel_4.1.2      RColorBrewer_1.1-2
[55] yaml_2.2.1          gridExtra_2.3       labelled_2.9.0
[58] rpart_4.1-15        reshape_0.8.8       latticeExtra_0.6-29
[61] stringi_1.7.5       highr_0.9           checkmate_2.0.0
[64] rlang_0.4.12        pkgconfig_2.0.3     evaluate_0.14
[67] labeling_0.4.2      htmlwidgets_1.5.4   cmprsk_2.2-10
[70] bit_4.0.4           tidyselect_1.1.1    plyr_1.8.6
[73] R6_2.5.1            generics_0.1.1      DBI_1.1.1
[76] pillar_1.6.4        haven_2.4.3         foreign_0.8-81
[79] withr_2.4.2         mgcv_1.8-38         abind_1.4-5
[82] nnet_7.3-16         etm_1.1.1           modelr_0.1.8
[85] crayon_1.4.2        utf8_1.2.2          tmvnsim_1.0-2
[88] tzdb_0.2.0          rmarkdown_2.11      jpeg_0.1-9
[91] grid_4.1.2          readxl_1.3.1        data.table_1.14.2
[94] reprex_2.0.1        digest_0.6.28       xtable_1.8-4
[97] httpuv_1.6.3        numDeriv_2016.8-1.1 munsell_0.5.0
```

THIS IS THE END OF THE QUIZ.