431 Quiz 2

Thomas E. Love

due 2018-12-11 at 5 PM. Version 2018-12-04 12:31:43

General Instructions

The deadline for completing this quiz is 5 PM on Tuesday 2018-12-11, and this is a firm deadline.

If you wish to work on some of the quiz and then return to it later, you can do this by [1] scrolling down to the final question which asks you to type in your full name, affirming that you have done the work for the quiz alone, 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 without losing your progress.

There are 40 questions, labeled Q01 through Q40. The maximum score available is 110 points.

- Each question is worth between 2 and 6 points. Partial credit is available on some questions.
- Note that Q01 is probably the question which will take the longest amount of time for you to answer, so don't worry that the other 39 questions are as lengthy as that one.
- The order of the questions is arbitrary. Some are meant to be easy, some are not.
- You should attempt to answer every question. There is no advantage to failing to give a response, as an incorrect response is always at least as valuable as a missing one.
- Please select or type in your best response (or responses, as indicated) for each question.

Data sets are available for several of the questions. You will find those data sets at https://github.com/THOMASELOVE/431-2018/tree/master/quizzes/quiz02

- The `oscar_A` and `oscar_B` data sets apply to Q01 and Q02.
- The `swordfish` data set applies to Q03 and Q04.
- The `limestone` data set applies to Q12.
- The `wc_code.R` bit of R code applies to Q22.
- The `hospsim` data set applies to Q26 - Q32.
- The `wcgs` data set was used to develop Q19, but you will not actually use it in responding to the quiz.
- Our usual `Love-boost.R` script will also be used in developing the answer sketch.

For any question where we do not specify something different, you should assume the following:

- that you are looking to do two-sided statistical inference with a 95% confidence level.
- that you should round all numerical responses to two decimal places.
- that two-by-two tables should be built without Bayesian adjustments

Please use `set.seed(2018)` whenever you need to do work that requires random sampling.

The packages we loaded in R to complete the Answer Sketch for this Quiz were:

- `Hmisc`, `fivethirtyeight`, `car`, `broom`, `Epi`, `magrittr`, and the `tidyverse`.
- We also made use of functions from `knitr`, `mosaic` and `gridExtra` and those were installed, but not loaded with `library()`.
- You may also need to use a function from the `vcd` package, which should also be installed on your machine.
- As mentioned above, we also sourced in the `Love-boost.R` script.

You are welcome to consult the materials provided on the course website, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants at `431-help at case dot edu`. Please submit any questions you have about the Quiz to `431-help` through email.

Good Luck!

1 Q01 (6 points)

The `oscar_A.csv` and `oscar_B.csv` data files are available to support your work in Q01 and Q02.

Every year, the Academy Awards (also called the Oscars) are presented for artistic and technical merit in the American film industry. The `oscar_A.csv` and `oscar_B.csv` data files that are available to you each contain the names and ages of 51 winners of the Best Actor in a Leading Role and the Best Actress in a Leading Role awards since 1967. The two files simply arrange the same data in two different formats. Use whichever format works better for you.

The key question you will address in Q01 is "At the 5% significance level, which group (Best Actors or Best Actresses) tends to have older Oscar winners, and by how many years on average?"

In your response to Q01, we expect you to

a. specify the test or confidence interval procedure you used (this includes specifying whether the samples are paired or independent),
b. justify clearly why you chose that procedure,
c. state clearly what the calculated interval estimate you developed is (rounded to two decimal places), and
d. state clearly what your conclusion is regarding the comparison of Best Actor ages to Best Actress ages based on this sample, using complete English sentences, and using a maximum of 1500 characters.

The form will limit you to a maximum of 1,500 characters in your response. Dr. Love's response in the answer sketch is approximately 600 characters. Use complete English sentences, and address all four issues (a, b, c, and d) specified in the directions, and anything else you think is important in addressing the key question "At the 5% significance level, which group (Best Actors or Best Actresses) tends to have older Oscar winners, and by how many years on average?".

2 Q02 (3 points)

If you look closely at the data in the `oscar_A` and `oscar_B` files discussed in Q01, you'll see they are missing the 1990 data. In that year, Daniel Day-Lewis won his first Oscar (for "My Left Foot") at the age of 32, while Jessica Tandy also won her first Oscar (for "Driving Miss Daisy") at the age of 80.

Which of the following statements are true about what would happen, if the 1990 data were included along with the 51 observations we've already used in Q01? (Note that no calculations are required here.)

[*Set up as TRUE or FALSE in each case.*]

- a. Whether the data were best analyzed as paired or independent samples would change.
- b. The point estimate of the population age difference incorporating the new data would shift closer to 0.
- c. A t-based confidence interval estimate incorporating the new data would be wider.

3 Q03 (4 points)

The `swordfish.csv` data file is available to support your work in Q03 and Q04.

Swordfish absorb mercury in their bodies, and it is thought that a mercury concentration of more than 1.00 ppm (parts per million) is not good for human consumption.

In a random sample of 28 swordfish, the concentrations listed below in the Output for Q03 and Q04 (and also in the `swordfish.csv` data file) were found.

```
0.07 0.24 0.39 0.54 0.61 0.72 0.81
0.82 0.84 0.90 0.95 0.98 1.02 1.08
```

```
1.14 1.20 1.20 1.26 1.29 1.30 1.37
1.40 1.44 1.58 1.62 1.68 1.85 2.10
```

The summary statistics from (`mosaic::favstats`) are:

```
  min     Q1 median     Q3 max     mean        sd  n missing
 0.07 0.8175   1.11 1.3775 2.1 1.085714 0.4757951 28       0
```

Suppose you want to know whether there is evidence at the 5% significance level that the mean concentration in the population of swordfish is different than 1.00 ppm, and you plan to use a two-sided bootstrap confidence interval with 1,000 bootstrap replications (which is the default choice) to make this decision.

In your response to Q03, (a) specify an appropriate confidence interval, and (b) clearly state your decision as to whether or not there is statistically significant evidence to support this claim. Do this in two or more complete sentences.

As is our general rule, round your interval to two decimal places. The form will limit you to a maximum of 500 characters in your response. Dr. Love's response in the Answer Sketch is approximately 160 characters.

4 Q04 (3 points)

Estimate a 95% confidence interval for the probability that a randomly selected swordfish will have a concentration of mercury above 1.00 ppm, based on the data in the `swordfish.csv` file, using the SAIFS approach. Express this probability as a proportion (between 0 and 1) rather than as a percentage (between 0 and 100). As is our general rule, round the endpoints of your confidence interval to two decimal places.

5 Q05 (3 points)

In *The Signal and The Noise*, Nate Silver writes repeatedly about a Bayesian way of thinking about uncertainty, for instance in Chapters 8 and 13. Which of the following statistical methods is **NOT** consistent with a Bayesian approach to thinking about variation and uncertainty?

- a. Gambling using a strategy derived from a probability model.
- b. Combining information from multiple sources to build a model.
- c. Establishing a researchable hypothesis prior to data collection.
- d. Significance testing of a null hypothesis, using, say, Fisher's exact test.
- e. Updating our forecasts as new information appears.

Background Information for Q06 - Q10

Suppose you fit four candidate regression models to predict the natural logarithm of a measure of predatory behavior in leopards.

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 seven predictors, Model B contains five of those seven predictors, and Model C contains three of the five Model B predictors, while Model D is a simple regression, using one of the predictors in Model C.

You obtain the results shown below in the Output for Q06 - Q10.

```
modelA <- lm(log(predatory.behavior) ~
               x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = leopard)
modelB <- lm(log(predatory.behavior) ~
               x1 + x2 + x3 + x4 + x5, data = leopard)
modelC <- lm(log(predatory.behavior) ~
               x1 + x2 + x3, data = leopard)
modelD <- lm(log(predatory.behavior) ~
               x1 , data = leopard)

BIC(modelA, modelB, modelC, modelD)
```

```
       df      BIC
modelA 10 4486.338
modelB  7 4470.056
modelC  5 5125.469
modelD  3 5304.291
```

6 Q06 (2 points)

Which of these models does the output above suggest will be the best choice to predict the natural logarithm of the predatory behavior measure?

- a. Model A
- b. Model B
- c. Model C
- d. Model D
- e. The output doesn't suggest a "best" choice.

7 Q07 (2 points)

The following output relates to `modelA` described in the previous question.

```
round(car::vif(modelA),3)
```

```
   x1    x2    x3    x4    x5    x6    x7    x8
1.018 1.014 1.074 1.007 1.014 2.665 2.554 1.013
```

Which of the following statements is the best conclusion from the output for Q07 shown above?

- a. Model A's residuals will show no problem with independence.
- b. Model A's residuals will show a serious problem with independence.

- c. Model A has no sign of meaningful collinearity.
- d. Model A has a serious problem with collinearity.
- e. Model A's residual variance will be larger than the residual variance of Model B, which is the model that includes predictors `x1, x2, x3, x4` and `x5`, only.

8 Q08 (2 points)

Which of the following R commands would provide fitted values of `log(predatory.behavior)` using the equation in Model A (from the previous two questions), for a new set of data contained in the `newleopard` tibble?

- a. `tidy(modelA, newdata = newleopard)`
- b. `glance(modelA, newdata = newleopard)`
- c. `augment(modelA, newdata = newleopard)`
- d. `split(modelA, newdata = newleopard)`
- e. None of these.

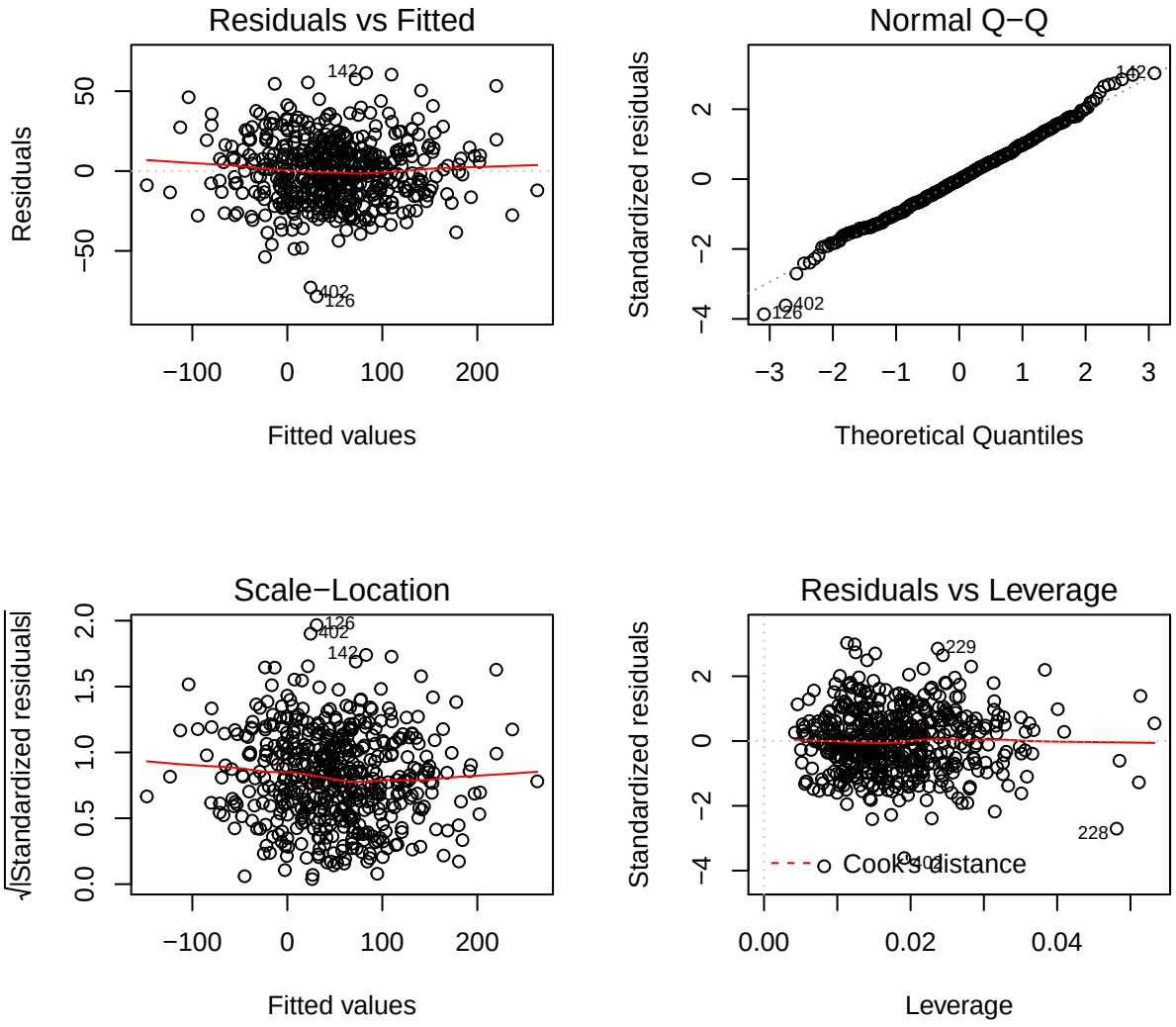
9 Q09 (2 points)

Suppose the first predicted subject in the `newleopard` tibble yields a prediction of `log(predatory.behavior)` of 3.5, with a 95% uncertainty interval of (3, 4). 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. 3 to 4
- b. `log(3)` to `log(4)`
- c. `10*3` to `10*4`
- d. `exp(3)` to `exp(4)`
- e. None of these would work.

10 Q10 (3 points)

Behold the residual plots for Model A that we have been discussing since Q06.



In the output for Q10 provided above, you see the residual plots for the Model A that we have been discussing since Q06. Which of the following conclusions is most appropriate?

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

11 Q11 (2 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 each of the changes listed below, indicate the impact on the width of the confidence interval.

- Rows
  - a. Increase the level of confidence.
  - b. Increase the sample size.
  - c. Increase the standard error of the estimate.
  - d. Use a bootstrap approach to estimate the CI.
- Columns
  - 1. CI will become wider
  - 2. CI will become narrower
  - 3. CI width will not change
  - 4. It is impossible to tell

12 Q12 (3 points)

The `limestone.csv` data file is available to support your work in Q12.

A geologist collects 130 hand-specimen sized pieces of limestone from a particular area. A qualitative assessment of both texture (either Fine, Medium or Coarse) and color (either Red, Orange or Brown) is made with the results shown in the Output for Q12.

—	Red	Orange	Brown
Fine	9	21	9
Medium	8	17	18
Coarse	19	22	7

Suppose you want to know if there is evidence of an association (at the 99% confidence level) between color and texture for these limestones. Which of the following conclusions is most appropriate?

- a. There is evidence of a significant color-texture association, since the p value is greater than 0.01
- b. There is evidence of a significant color-texture association, since the p value is less than 0.01
- c. There is no evidence of a significant color-texture association, since the p value is greater than 0.01
- d. There is no evidence of a significant color-texture association, since the p value is less than 0.01
- e. It is impossible to tell from the information provided.

13 Q13 (3 points)

Suppose you have a data frame named `dat` containing a variable called `height`, which shows the subject's height in centimeters. Which of the following lines of code will create a new variable `tall` in the `dat` data frame which takes the value **TRUE** when a subject is more than 175 cm tall, and **FALSE** when a subject's height is at most 175 cm?

- a. `dat %>% mutate(tall = height > 175)`
- b. `dat %>% tall <- height > 175`
- c. `dat$tall <- ifelse(dat$height > 175, "YES", "NO")`
- d. `tall <- dat %>% filter(height > 175)`
- e. None of these will do the job

14 Q14 (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.

```
Analysis of Variance Table

Response: exam2
           Df  Sum Sq Mean Sq F value  Pr(>F)
ta          4  1199.8 299.950  3.3355 0.01174
Residuals 165 14837.8  89.926
```

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 significance level should she use for each of the individual two-sample t tests?

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

15 Q15 (4 points)

Suppose you have completed a pilot study of average birth weight for full term infants whose gestational age is 40 weeks. In that sample, the infants whose mothers smoked during pregnancy had a mean birth weight that was 300 grams lighter than the mean birth weight of the infants whose mothers did not smoke during pregnancy, while the standard deviation of the birth weights was about 430 grams in each group.

Assume that you are planning to conduct a new study, with a balanced design where you will initially enroll in the study a total of 400 infants (again, who were at 40 weeks gestation) in a balanced design between the two exposure groups (by the mother's smoking status.) In this new study, you want to be able to detect a difference in means that is at least half as large as what you observed in the pilot study.

If 10% of your initially enrolled subjects (in each smoking group) cannot be used in the final analysis for some reason, then with what power will you be able to detect the desired effect? Please present your power estimate as a percentage, rather than as a proportion, and round it to the nearest integer.

16 Q16 (3 points)

Suppose that 80 of 100 male applicants to a graduate school are accepted, while 60 of 100 female applicants are accepted. Estimate a two-sided 95% confidence interval for the relative risk of acceptance for a male applicant as compared to a female one. Round all parts of your response to one decimal place.

17 Q17 (3 points)

Breaking down the applications described in Q16 into the school's two separate programs, we find that program A accepted 72 of its 80 male applicants, and program B accepted 41 of its 80 female applicants. Each of the 200 students described in Q16 applied to either program A or program B, and not to both. Which type of student (males or females) had lower odds of being accepted by the school ...

Rows:

- a. into Program A?
- b. into Program B?
- c. into the school overall?

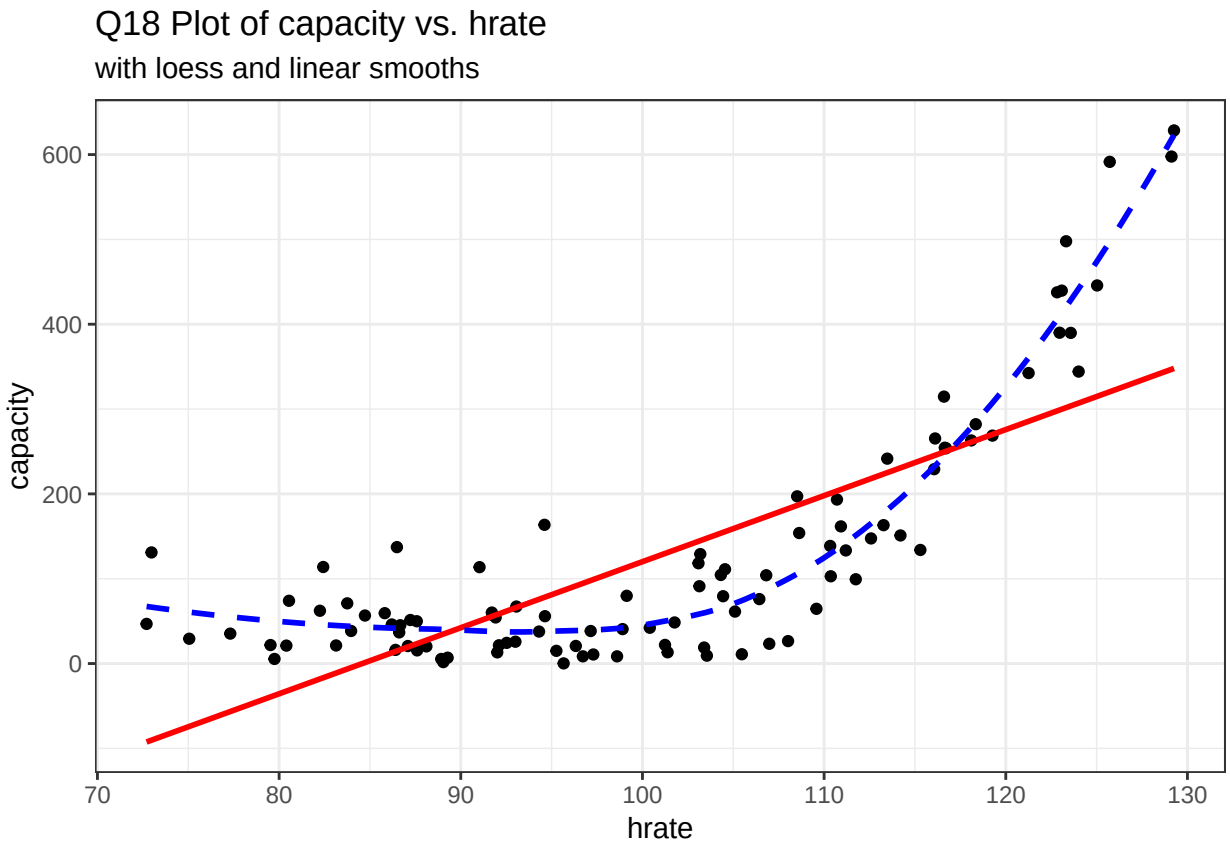
Columns:

- Females had lower odds.
- Males had lower odds.

18 Q18 (2 points)

Suppose we plotted the relationship between an outcome related to blood flow capacity, labeled `capacity`, and a predictor called `hrate`, which is a measure of peak comfortable heart rate. Each is measured for a cross-section of 100 subjects.

We then used the `geom_smooth` function in `ggplot2` to fit both a linear smooth and a loess smooth, producing the plot shown below in the output for Q18. One smooth is shown with blue dashes and the other is shown as a red solid line.



Which of the following statements is true?

- a. The linear fit is shown as a blue dashed line in this plot.
- b. The linear model describing `capacity` using `hrate` has a problem with independence.
- c. The Pearson correlation of `hrate` and `capacity` is negative.
- d. The linear model provides a better fit to the data than does the loess smooth.
- e. None of these statements are true.

19 Q19 (3 points)

Data describing a sample of subjects participating in the Western Collaborative Group Study (discussed in our Course Notes in several places) were used here to fit a model to predict the natural logarithm of systolic blood pressure (`sbp`) using the subject's age, height, smoking status (yes/no) and the natural logarithm of their weight. The `wcgs` file is on our website, but this question uses a sample from that data that is unknown to you, so you will not be able to duplicate the output that follows.

```
m19 <- lm(log(sbp) ~ age + log(weight) + height + smoke, data = wcgs19)
tidy(m19, conf.int = TRUE) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.528	0.414	10.941	0.000	3.710	5.346
age	0.008	0.002	3.977	0.000	0.004	0.011
log(weight)	0.163	0.084	1.934	0.055	-0.004	0.330
height	-0.012	0.004	-2.682	0.008	-0.021	-0.003
smokeYes	-0.023	0.021	-1.122	0.264	-0.064	0.018

```
glance(m19) %>%
  select(r.squared, adj.r.squared, sigma, statistic, df,
         p.value, AIC, BIC) %>%
  knitr::kable(digits = 3)
```

	r.squared	adj.r.squared	sigma	statistic	df	p.value	AIC	BIC
value	0.171	0.148	0.122	7.491	5	0	-199.285	-181.221

What conclusions can you draw from this output, using a 5% significance level?

[*SET UP as SEPARATE TRUE-FALSE items*]

- a. Smokers have significantly lower systolic blood pressures than non-smokers, after we account for age and size (height and weight).
- b. Larger height is associated with significantly lower blood pressure, even after we've accounted for age, weight and smoking status.
- c. This model accounts for more than 15% of the variation in the log of systolic blood pressure.

20 Q20 (2 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.)

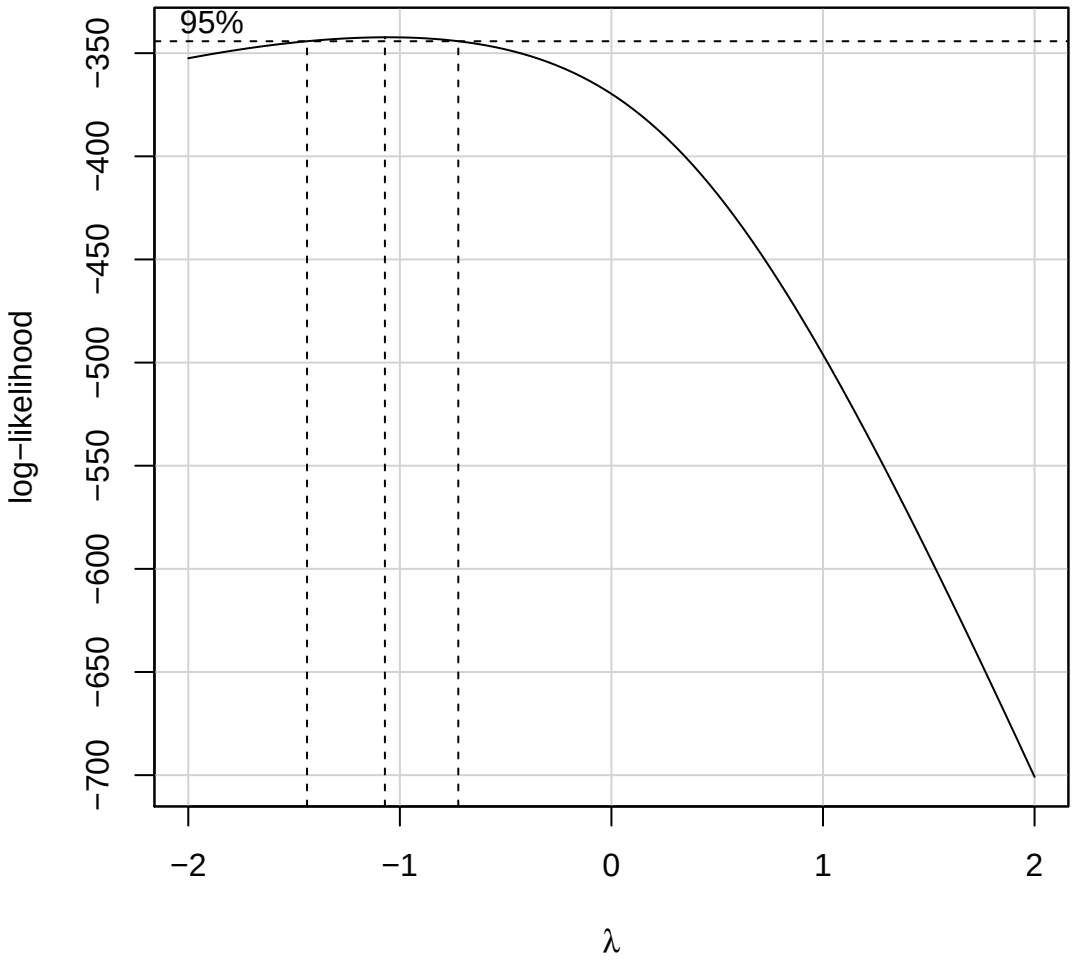
Which of the pairs of `tidyverse` functions identified below can be used to accomplish such a plot?

- 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`

21 Q21 (2 points)

Consider the Box-Cox plot below, which addresses a model built to predict an outcome called `score` using four predictors.



What transformation of our response does this plot suggest?

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

22 Q22 (2 points)

The code snippet `wc_code.R` is available to support your work in Q22 - Q24.

Consider the `weather_check` data frame within the `fivethirtyeight` package. We will use these data for Q22-Q24.

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`.

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

23 Q23 (3 points)

Perform an appropriate test to see if the odds ratio for a Yes (TRUE) response to "Do you typically check a daily weather report?" comparing Female to Male respondents is essentially consistent across age categories. What is the name of the test that you ran, and what is the conclusion? As usual, use a 5% significance level here.

- a. I ran a chi-square test on a 2x2 table using the `Epi` package's `twoby2` function, and the conclusion is that there is a significant association.
- b. I ran a chi-square test on a 2x2 table using the `Epi` package's `twoby2` function, and the conclusion is that there is not a significant association.
- c. I ran Woolf's test (`woolf_test`) to assess the homogeneity of odds ratios from the `vcd` package, and I conclude that the odds ratio is sufficiently consistent across age categories to allow me to collapse on age.
- d. I ran Woolf's test (`woolf_test`) to assess the homogeneity of odds ratios from the `vcd` package, and I conclude that the odds ratio is NOT sufficiently consistent across age categories to allow me to collapse on age.
- e. None of these statements describe an appropriate test.

24 Q24 (3 points)

Use the data we have been working with in the previous two questions, regardless of how you answered those questions. Suppose we want to use the Cochran-Mantel-Haenszel approach to estimate the common odds ratio across all age categories comparing Females to Males as to whether they check the weather daily. Which of the following statements is true?

- a. Females have higher odds of checking the weather, and a 95% confidence interval includes 1.
- b. Females have higher odds of checking the weather, and a 95% confidence interval does not include 1.
- c. Females have lower odds of checking the weather, and a 95% confidence interval includes 1.
- d. Females have lower odds of checking the weather, and a 95% confidence interval does not include 1.
- e. None of these statements are true.

25 Q25 (3 points)

Which of the following statements is **NOT** part of what Silver is trying to tell us in *The Signal and The Noise*? (You may wish to focus on Chapter 13, which summarizes the preceding arguments nicely.)

- a. Our bias is to think we are better at prediction than we really are.
- b. Make a lot of forecasts. It's the only way to get better.
- c. State, explicitly, how likely we believe an event is to occur before we begin to weigh the evidence.
- d. Revise and improve your estimates as you encounter new information.
- e. Nature's laws change quickly, and do so all the time.

Setup for Q26-32

For Q26 - Q32, consider the data I have provided in the `hospsim.csv` file. The data describe 750 patients seen for care in the past year at a metropolitan hospital system. 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)
- `ehr_time` = Continuing or New, where Continuing means their electronic health record indicates they were also seen for care in this hospital system last year. New means they are "new" to the system this year.
- `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

26 Q26 (3 points)

Using the `hospsim` data, what is the 95% confidence interval for the odds ratio which compares the odds of receiving a statin if you were seen last year as well as this year divided by the odds of receiving a statin if

you were a new patient this year. Do **NOT** use a Bayesian augmentation.

- a. Odds Ratio is 0.51, CI is (0.36, 0.71)
- b. Odds Ratio is 0.72, CI is (0.60, 0.86)
- c. Odds Ratio is 1.18, CI is (1.08, 1.28)
- d. Odds Ratio is 1.98, CI is (1.41, 2.78)
- e. None of these answers are correct

27 Q27 (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. Which of the following conclusions is most appropriate based on your significance tests?

- a. The ANOVA F test is not significant, so it doesn't make sense to compare insurance types pairwise.
- b. The ANOVA F test is significant, and a Tukey HSD comparison reveals that Medicare shows significantly higher education levels than Uninsured.
- c. The ANOVA F test is significant, and a Tukey HSD comparison reveals that Medicaid's education level is significantly lower than either Medicare or Commercial.
- d. The ANOVA F test is significant, and a Tukey HSD comparison reveals that Uninsured's education level is significantly lower than Commercial or Medicare.
- e. None of these conclusions is appropriate.

28 Q28 (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 appears true? Do not transform your outcome.

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

29 Q29 (3 points)

Run a backwards elimination stepwise procedure. After doing so, how many of the original nine regression inputs (`clinic.type`, `age`, `ehr_time`, `insurance`, `hsgrads`, `a1c`, `bmi`, `sbp` and `statin`) remain in the model?

- a. 1, 2, or 3
- b. 4
- c. 5

- d. 6
- e. 7 or 8

30 Q30 (3 points)

Compare your initial "kitchen sink" model with all 9 inputs to the model generated by the stepwise approach in Q29 using adjusted R-squared, AIC and BIC. For each summary approach, which of the two models you are comparing gives BETTER results?

Rows:

- a. AIC
- b. BIC
- c. Adjusted R^2

Columns:

1. [SMALLER]. The smaller model (stepwise result from Q29)
2. [KITCHEN SINK]. The kitchen sink model with all 9 predictors

31 Q31 (3 points)

Now build a model using ehr_time and insurance type to predict a different outcome, hemoglobin A1c. Which of the following statements best describes the result?

- a. The model R^2 is below 10%, and both ehr_time and insurance type have a significant impact on hemoglobin A1c given the other predictor.
- b. The model R^2 is above 10%, and both ehr_time and insurance type have a significant impact on hemoglobin A1c given the other predictor.
- c. The model R^2 is below 10%, and neither ehr_time nor insurance type have a significant impact on hemoglobin A1c given the other predictor.
- d. The model R^2 is above 10%, although neither ehr_time nor insurance type have a significant impact on hemoglobin A1c given the other predictor.
- e. None of these statements are true.

32 Q32 (3 points)

In your model for Q31, identify the subject with the largest residual. Which of the following set of features best describe this subject?

- a. This is a continuing Medicare patient who is less than 65 years of age.
- b. This is a continuing Medicare patient who is 65 years old, or older.
- c. This is a new Medicare patient who is less than 65 years of age.
- d. This is a new Medicare patient who is 65 years old, or older.
- e. None of these accurately describe the subject in question.

33 Q33 (2 points)

For each listed research question, decide what statistical procedure (of those listed) would be **most** useful in answering the question posed. Assume all assumptions have been met for using the procedure.

Rows:

- a. Do college grade point averages differ for male athletes in major sports (e.g., football), minor sports (e.g., swimming), and in intramural sports?
- b. Does intelligence as measured by IQ score differ between college students on academic probation and those not on probation?
- c. Does support for a school bond issue (For or Against) differ by neighborhood in the city?
- d. In twins of opposite sex, does the boy score higher or lower on a test of reading achievement?

Columns:

- 1. Independent Samples t test / CI.
- 2. Paired Samples t test / CI.
- 3. One-Way ANOVA comparing more than two means.
- 4. Chi-Square test of Association.

34 Q34 (2 points)

Suppose we have several potential models for a particular outcome, and we obtain the following output.

Model	Multiple R-squared	Adjusted R-squared
A	0.41	0.40
B	0.49	0.41
C	0.53	0.43
D	0.55	0.47

Which of these models is most likely to retain its nominal R-square value in predicting new data?

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

35 Q35 (3 points)

Anne is a researcher on the West Coast of the U.S. who wants to estimate the amount of a newly discovered antibody in human blood. Anne's research funds will only let her obtain blood samples from 41 people, so she decides to construct a two-sided 90% confidence interval thinking it will give her a more precise estimate of the mean antibody level in the population. Bill is a researcher on the East Coast of the United States who is researching the same antibody, but he has more research funding and can afford to obtain blood samples from 121 people. Bill decides to construct a two-sided 99% confidence interval. Suppose that the sample standard deviation will be about 10 in each of the samples, and that each researcher plans to use a t-based interval.

Which estimate, Anne's or Bill's, will produce the more precise estimate, if more precise is taken to mean the interval estimate with the smaller width?

a. Anne's estimate will be more precise.
b. Bill's estimate will be more precise.
c. The estimates will be equally precise.
d. It is impossible to tell from the information provided.

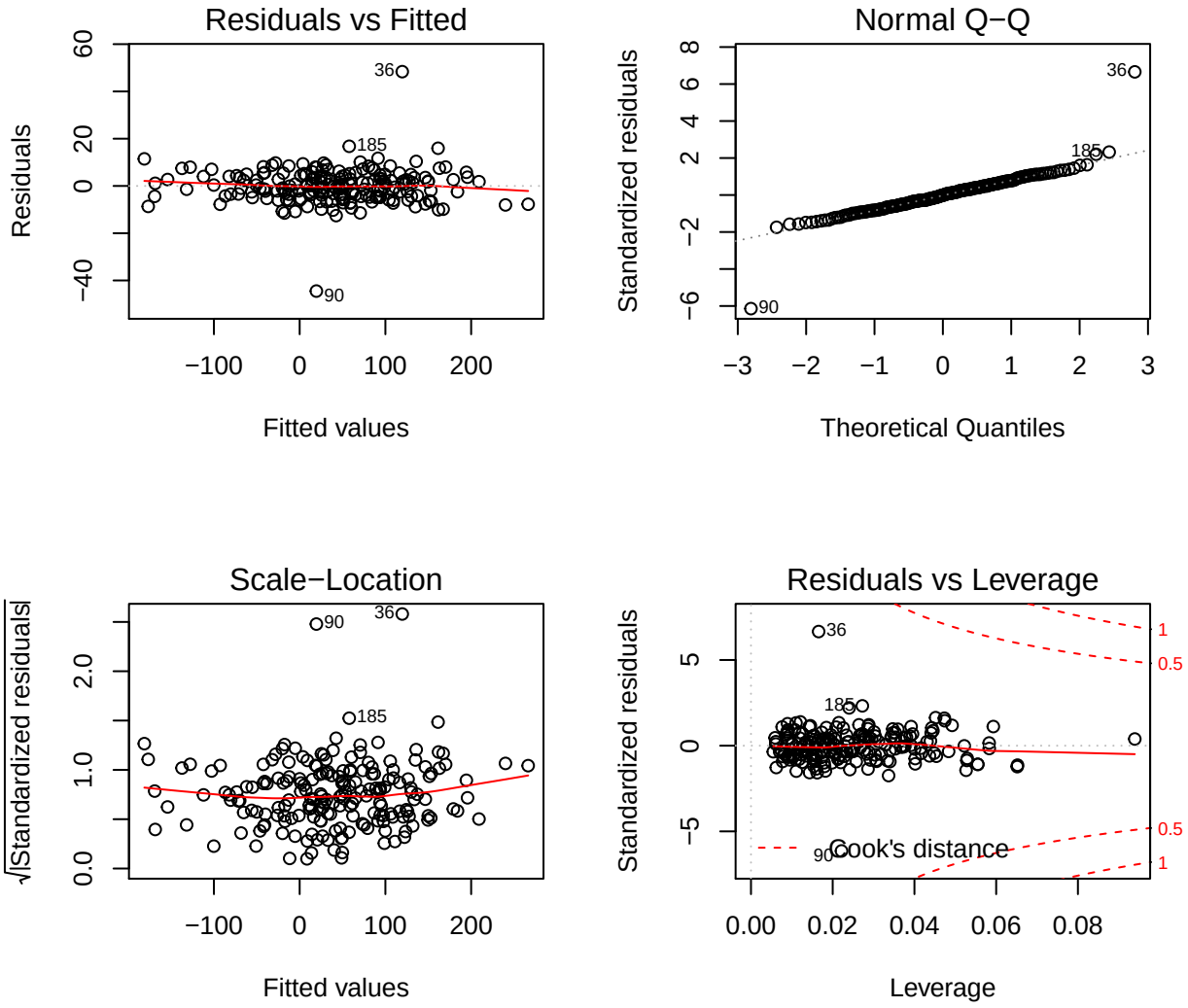
36 Q36 (2 points)

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 containing the rest?

- a. `mydat.test <- sample_n(mydat, 100)` and `mydat.train = anti_join(mydat, mydat.test)`
- b. `mydat.test <- partition(mydat, 400:100)` and `mydat.train = anti_join(mydat, mydat.test)`
- c. `mydat.test <- slice(mydat, 100)` and `mydat.train = anti_join(mydat, mydat.test)`
- d. `mydat.test <- sample_frac(mydat, 0.80)` and `mydat.train = anti_join(mydat, mydat.test)`
- e. None of these approaches would work.

37 Q37 (2 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 200 subjects. The following residual plots emerged from R.



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

- 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.

38 Q38 (3 points)

Suppose now that we want to build a study of the efficacy of a new drug formulation, as compared to the old formulation. We have decided that about 45% of people respond to the old formulation, and we want to declare a statistically significant effect of the new drug if we complete a two-sided test using a 5% significance level, if at least 55% of those receiving the new drug respond. If we want at least 90% power, and plan a balanced design, how many subjects will we need to enroll, in total, across the two formulation groups? Assume that no enrolled subjects will drop out of the study.

39 Q39 (2 points)

According to Jeff Leek in *The Elements of Data Analytic Style*, which of the following is **NOT** a good reason to create graphs for data exploration?

- a. To understand properties of the data.
- b. To inspect qualitative features of the data rather than a huge table of raw data.
- c. To discover new patterns or associations.
- d. To consider whether transformations may be of use.
- e. To look for statistical significance without first exploring the data.

40 Q40 (2 points)

A special method using regression strategies uses a sample of data to estimate a parameter as 2.35, with a standard error of 0.5. Which of the following statements best describes a 95% uncertainty interval (confidence interval) for that parameter, based on this sample?

- a. (2.35 - 0.5, 2.35 + 0.5)
- b. (2.35 - 0.1, 2.35 + 0.1)
- c. (2 - 2.35, 2 + 2.35)
- d. (2.35 - 0.35, 2.35 + 0.35)
- e. None of these.