432 Quiz 1

Thomas E. Love

Deadline: 9 PM 2022-02-21. Version: 2022-02-21 09:58:41

Edits Since Initial Posting

- Corrected Plot for Question 3, which may affect your responses to Questions 3 and 4.
- Clarified that the coefficients in the table at the top of the Output for Questions 11 and 12 are exponentiated.

Links

All links for the Quiz will be made available at https://github.com/THOMASELOVE/432-2022/tree/main/quiz/quiz1 at 5 PM on 2022-02-17.

This will include links to:

- the Main Document (this document) containing the instructions and questions
- the Google Form Answer Sheet, and
- the data sets we are providing

Instructions

This PDF document is **20** pages long. There are **24** questions on this Quiz. 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 must be submitted through the Google Form Answer Sheet found on the links page by the deadline, without exception. The form will close at that time, and no extensions will be made available, so please do not wait until Monday evening 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. The Answer Sheet works like a Minute Paper, in that you must be logged into Google via CWRU to access it.

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 Answer Sheet as often as you like without losing your progress.

The Data Sets

I have provided **3** data sets (called **set1.csv**, **set13.Rds**, and **set22.csv**) that are mentioned in the Quiz. They may be helpful to you.

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

During the Quiz period (2022-02-17 through 2022-02-21) 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 2022-02-21 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 post to Piazza in the `quiz1` folder 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.

Scoring and Timing

All questions are worth between 2 and 6 points, adding to a total of 100 points. The questions are not in any particular order, and range in difficulty from "things Dr. Love expects everyone to get right" to "things that are deliberately tricky". Some questions will take more time than others to answer.

The Quiz is meant to take 4-5 hours to complete. I expect most students will take 3-6 hours, and some will take as little as 2 or as many as 8. Again, it is **not** a good idea to spend a long time on any one question.

Dr. Love will grade the Quiz, and results (including an answer sketch) will be available by class time on Tuesday 2022-03-01.

What does the Quiz cover?

Quiz A includes material from the first 12 classes in 432, including:

- Chapters 1-14 of the 432 course notes, excluding some of the `tidymodels` material
- Dr. Love's note on interpreting effect sizes from Class 09
- all of Jeff Leek's *How to be a Modern Scientist* and
- Chapters 1-5 of Nate Silver's *The Signal and the Noise*.

Writing Code into the Answer Sheet

Occasionally, we ask you to provide R code in your response. Do not include the `library` command at any time for any of your code. Instead, assume in all questions that all relevant packages have been loaded in R. A list of R packages that Dr. Love used in building the Quiz and its answer sketch is available in the next section.

Packages and Settings used by Dr. Love

This doesn't mean you need to use all of these packages, nor does it mean that you are prevented from using other packages we've discussed in class to complete the Quiz.

```
library(here)
library(knitr)
library(janitor)
library(magrittr)
library(naniar)
library(patchwork)
library(GGally)
library(ggrepel)
library(equatiomatic)
library(simputation)
library(rms)        # includes Hmisc
library(tidymodels) # includes broom, rsample, etc.
library(tidyverse)  # includes dplyr, ggplot, etc.

# Note that all data files were downloaded onto
# my machine into a subfolder called data below
# my main R Project directory for Quiz 1.

theme_set(theme_bw())
opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)
```

1 Question 1

The `set1.csv` data set provided to you contains information on eight variables for 1250 subjects, and we will use it for Questions 1-8. In this question, we refer to the eight columns containing data in the .csv file, from left to right, as columns A, B, C, D, E, F, G and H, as they are labeled in a spreadsheet.

Import the data from `set1.csv` into a tibble called `set1` in R, and in doing so, use the `type.convert(as.is = FALSE)` function to convert all `character` type variables to `factor` variables, and **then** apply the `clean_names()` function (using its default settings) from the `janitor` package.

Which of the variables (taken from the .csv file) have their names change as a result of this process? (Choose **all** of the correct responses.)

a. The name of the variable in column A (the leftmost column) of the .csv file.
b. The name of the variable in column B
c. Column C's name
d. Column D's name
e. Column E's name
f. Column F's name
g. Column G's name
h. The name of the variable in column H (the rightmost column) of the .csv file.
i. None of the variable names change.

2 Question 2

Question 2 has four parts, and the `set1` tibble it asks about is the one you created in Question 1. Please respond to each of these parts with the correct number.

a. What is the total number of missing data values in the `set1` tibble?

b. How many of the observations in the `set1` tibble have complete data on all eight variables contained in that tibble?

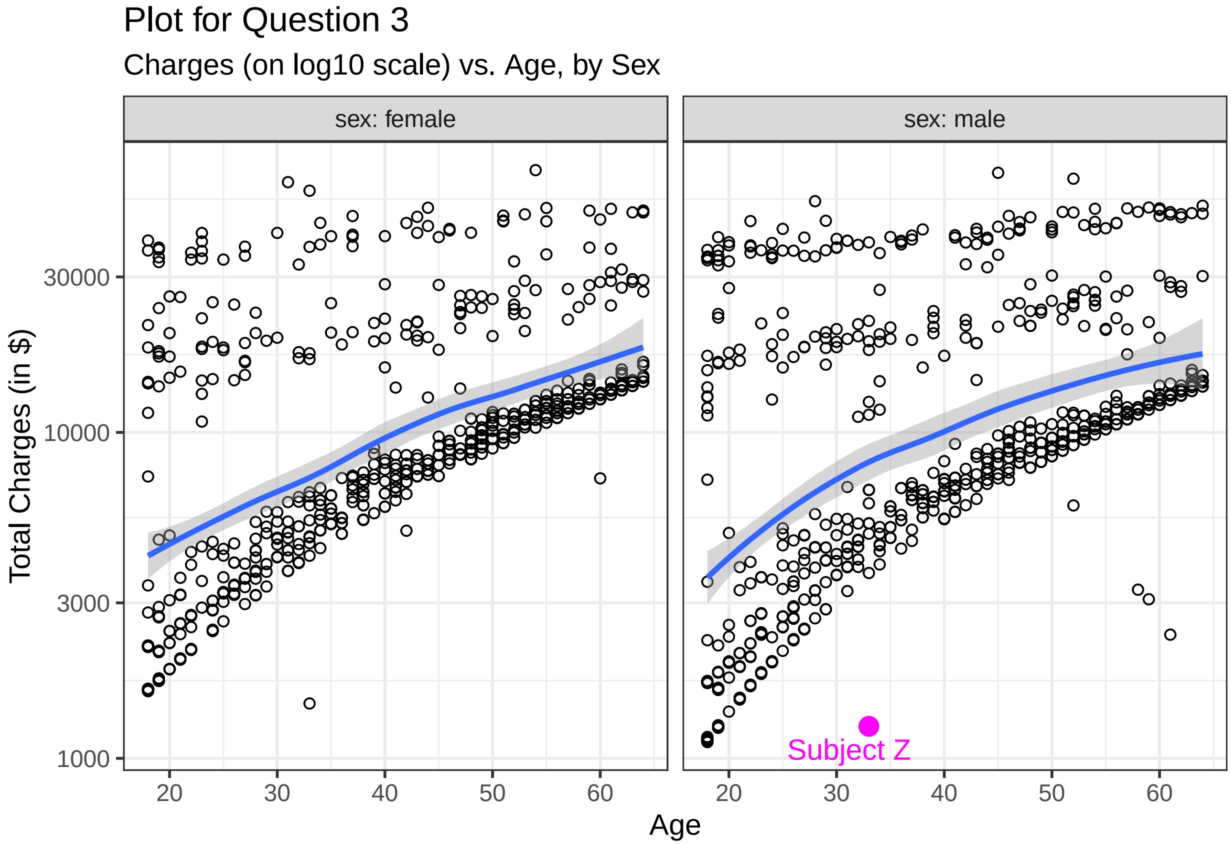
c. How many of the observations in the `set1` tibble are missing data on more than one of the eight variables in that tibble?

d. How many variables in the `set1` tibble contain complete data?

3 Question 3

Consider the Plot for Question 3, developed by Dr. Love using the `set1` tibble created in Question 1.

Note: This plot was corrected on **2022-02-18**.



Specify the subject number (`Subj_No` in the `set1.csv` file) for the subject who is identified in the Plot for Question 3, using a magenta color, and who is labeled there as Subject Z.

4 Question 4

In developing the Plot for Question 3, Dr. Love used several R commands, including each of the eight listed below except one. Which one did he **not** use?

a. `facet_wrap(~ sex, labeller = "label_both")`
b. `geom_point(shape = 1)`
c. `geom_smooth(method = "loess", formula = y ~ x)`
d. `labs(caption = "Charges (on log scale) vs. Age, by Sex")`
e. `labs(title = "Plot for Question 3")`
f. `labs(x = "Age", y = "Total Charges (in $)")`
g. `scale_y_log10()`

Output for Questions 5-6

The output for Questions 5-6 describes results from fitting a model m5 to predict the base-10 logarithm (log10 in R) of the total charges in the set1 tibble, using two predictors and an interaction term.

ANOVA table for model m5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
us_region	3	0.884	0.295	3.305	0.020
smoker	1	81.990	81.990	920.086	0.000
us_region:smoker	3	0.722	0.241	2.701	0.044
Residuals	1160	103.369	0.089	NA	NA

Tidied Coefficients for model m5

term	estimate	conf.low	conf.high
(Intercept)	3.87	3.83	3.91
us_regionnorthwest	-0.05	-0.10	0.01
us_regionsoutheast	-0.09	-0.15	-0.04
us_regionsouthwest	-0.08	-0.13	-0.02
smokerTRUE	0.57	0.49	0.66
us_regionnorthwest:smokerTRUE	0.07	-0.06	0.19
us_regionsoutheast:smokerTRUE	0.16	0.05	0.28
us_regionsouthwest:smokerTRUE	0.12	-0.01	0.24

Excerpts from glance() for model m5

r.squared	adj.r.squared	df	df.residual	AIC	BIC	nobs
0.447	0.444	7	1160	500.54	546.107	1168

5 Question 5

Again, the Output for Questions 5-6 displays results obtained by fitting the m5 model to predict the base-10 logarithm (log10 in R) of the total charges in the set1 tibble, using two predictors and an interaction term. Suppose we have two new subjects, for whom we intend to make predictions for their total charges (in $) using model m5.

- Amy is a smoker who lives in the Southeast.
- Ben is a non-smoker who lives in the Southwest.

Use the table of tidied coefficients provided in the Output for Questions 5-6 to obtain a predicted value of total charges (in $) from model m5 for Amy, and then for Ben. Then tell us…

a. Based on the tidied coefficients for model m5, which subject, Amy or Ben, is predicted to have a larger value of total charges?
b. Again using the tidied coefficients, how much larger are that subject's predicted total charges, rounded to the nearest dollar?

6 Question 6

While developing the model shown in the Output for Questions 5-6 using the `set1` data, Dr. Love made which of the following assumptions about the missing data mechanism?

a. Missing Completely at Random
b. Missing At Random
c. Missing Not At Random
d. We cannot tell from the information provided.

Output for Question 7

```
m7
```

```
Linear Regression Model

                  Model Likelihood    Discrimination
                        Ratio Test           Indexes
 Obs    1168    LR chi2    1798.76    R2       0.786
 sigma0.1863    d.f.            12    R2 adj   0.783
 d.f.   1155    Pr(> chi2)  0.0000    g        0.407

                             Coef    S.E.   t      Pr(>|t|)
 Intercept                    2.7001 0.1037  26.03 <0.0001
 us_region=northwest         -0.0301 0.0174  -1.73 0.0834
 us_region=southeast         -0.0588 0.0173  -3.39 0.0007
 us_region=southwest         -0.0683 0.0174  -3.92 <0.0001
 smoker                       1.1482 0.0470  24.43 <0.0001
 age_of_subject               0.0380 0.0049   7.74 <0.0001
 age_of_subject'             -0.3491 0.1166  -2.99 0.0028
 age_of_subject''             0.4504 0.1555   2.90 0.0038
 age_of_subject'''           -0.1410 0.0584  -2.41 0.0160
 age_of_subject * smoker     -0.0140 0.0010 -14.31 <0.0001
 us_region=northwest * smoker  0.0428 0.0401   1.07 0.2860
 us_region=southeast * smoker  0.1234 0.0368   3.35 0.0008
 us_region=southwest * smoker  0.1203 0.0404   2.98 0.0030
```

7 Question 7

Dr. Love then fit an alternative model to model the same outcome as in `m5`, but now including information about the age of the subject. Part of the output from that model is shown in the Output for Question 7. How was the subject's age included in the model `m7`? (Check all that apply.)

a. Using a 3rd degree (cubic) polynomial
b. Using a 4th degree (quartic) polynomial
c. Using a restricted cubic spline with 3 knots
d. Using a restricted cubic spline with 4 knots
e. Using a restricted cubic spline with 5 knots
f. Using an interaction term with `smoker`
g. None of the above

Output for Question 8

```
set.seed(432); validate(m7)
```

```
          index.orig training   test optimism index.corrected  n
R-square      0.7856   0.7894 0.7834   0.0059          0.7797 40
MSE           0.0343   0.0338 0.0347  -0.0009          0.0352 40
g             0.4070   0.4084 0.4064   0.0021          0.4050 40
Intercept     0.0000   0.0000 0.0195  -0.0195          0.0195 40
Slope         1.0000   1.0000 0.9951   0.0049          0.9951 40
```

8 Question 8

Consider again the model `m7` that Dr. Love fit in Question 7. In the Output for Question 8, we have provided some output associated with that model for predicting our outcome (the base-10 logarithm of total charges.) This question has two parts.

a. Based on this output, what is the mean squared error we obtain when predicting our outcome using `m7` within the data used to fit the model?
b. Based on this output, what is our best estimate of the proportion of variation in our outcome that will be explained using `m7` in a new batch of data?

9 Question 9

Suppose you are reviewing an academic paper and you have the four options listed below. In "How to be a Modern Scientist", Jeff Leek suggests that there is a #1 way to be a jerk reviewer. Which of the following recommendation decisions could be made by someone who was actively TRYING TO BE a jerk reviewer? (Select any that apply.)

a. Accept
b. Minor revisions
c. Major revisions
d. Reject

10 Question 10

In addition to the raw data, which of the following should be part of the "data package" that you share, according to Jeff Leek in *How to be a Modern Scientist*, when you are trying to maximize speed in the analysis of the data? (Check all of the correct responses.)

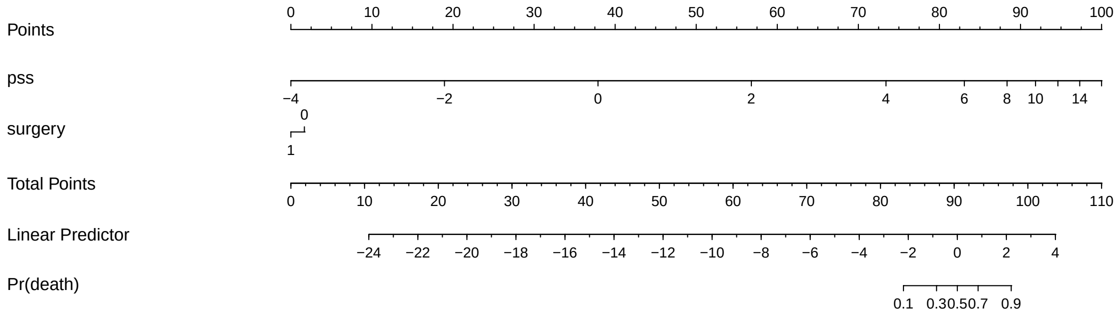
a. A tidy data set.
b. A research question.
c. A code book describing each variable and its values.
d. An explicit recipe describing how you went from the raw data to the tidy data set and code book.
e. An exploratory data analysis of the outcome of interest.
f. A substantial bribe.

Output for Questions 11-12

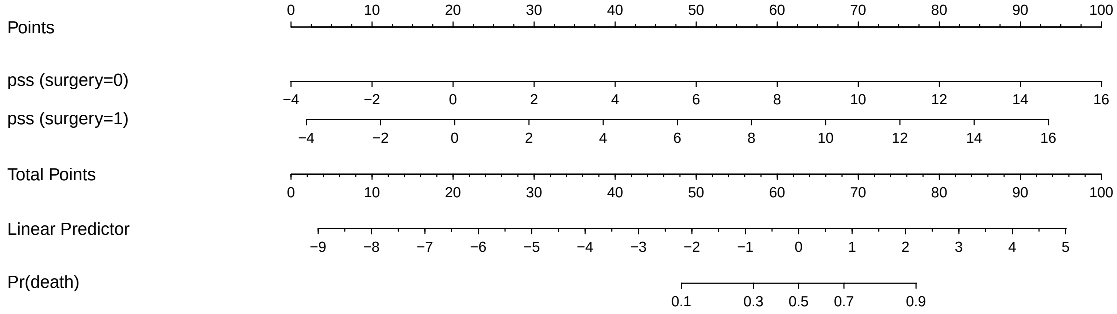
Model `m11` (Exponentiated) Coefficients with 95% confidence intervals

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.0015	1.4038	-4.6111	0.0000	0.0001	0.0157
pss	2.1356	0.1694	4.4780	0.0000	1.6094	3.1659
surgery	1.0323	1.7002	0.0187	0.9851	0.0435	41.1678
pss:surgery	0.9380	0.2085	-0.3069	0.7589	0.5994	1.3861

Nomogram A



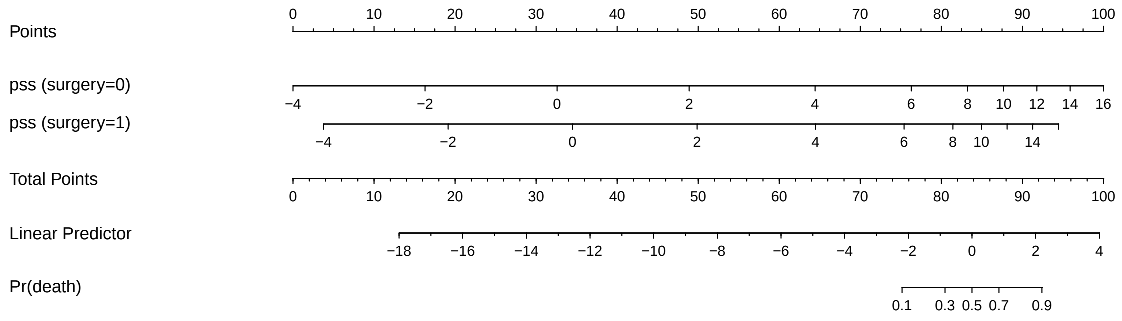
Nomogram B



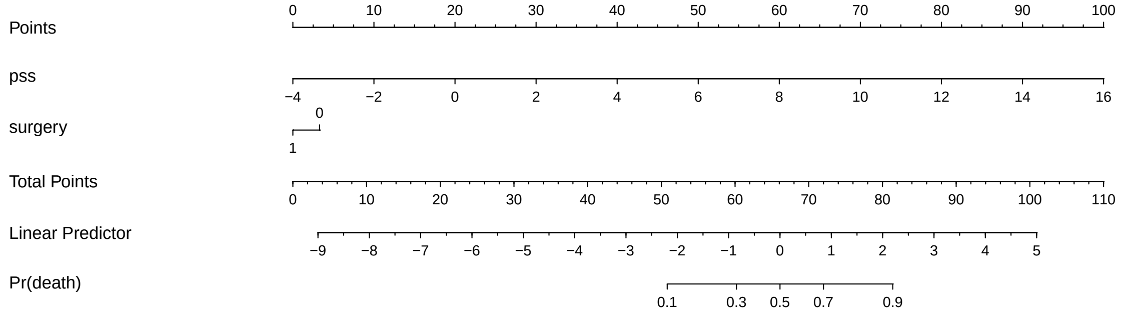
The Output for Questions 11-12 continues on the next page.

Output for Questions 11-12 (continued)

Nomogram C



Nomogram D



Note that Questions 11 and 12 appear on the next page.

11 Question 11

Questions 11 and 12 refer to a retrospective study of 288 patients with esophageal perforation, 57 of whom died during the study follow-up period.

The logistic regression model called model `m11` whose estimated (and exponentiated) coefficients are shown at the top of the Output for Questions 11-12 was fit to describe the probability that a patient with esophageal perforation would die (`death` = 1 if died, `death` = 0 if alive) based on their Pittsburgh severity score (`pss`) and whether or not they underwent surgery (`surgery` = 1 if the subject had surgery and 0 if they did not.)

An investigator produced four different nomograms and these are labeled in the Output for Questions 11-12 as Nomograms A, B, C and D. Exactly one of the nomograms shown in that output applies to model `m11`. Which one?

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

12 Question 12

According to the appropriate nomogram from the set presented in the Output for Questions 11-12, what is the predicted probability of death from model `m11` for a subject who had surgery and whose Pittsburgh severity score was 6?

a. Less than 0.2
b. Between 0.21 and 0.4
c. Between 0.41 and 0.6
d. Between 0.61 and 0.8
e. Greater than 0.8
f. It is impossible to tell from the information provided.

Setup for Questions 13-17

The data in the `set13.Rds` data frame we have provided will be used in Questions 13-17.

The `set13` data describes a 1970s study of potential insurance redlining (redlining = canceling insurance policies or refusing to renew them) in 47 zip codes in Chicago, Illinois. The measures we've included in the `set13` data are:

- `zip` = the zip code
- `fair` = High or Low (High means that there was at least 1 FAIR plan policy or renewal per 100 housing units in the zip code in the first half of 1978, Low means there were less than 1 per 100 housing units.)
- `minority` = percentage of the zip code's residents who describe themselves as being a member of a minority race
- `older` = percentage of housing units built before 1939 in the zip code
- `medinc` = median family income in thousands of dollars
- `side` = zip is located in the North Side (n) or South Side (s) of Chicago

Here's a quick numerical summary of the available data.

```
summary(set13 %>% select(-zip))
```

```
   fair       minority          older           medinc        side
 High:10   Min.   : 1.00   Min.   : 2.00   Min.   : 5.583   n:25
 Low :37   1st Qu.: 3.75   1st Qu.:48.60   1st Qu.: 8.447   s:22
           Median :24.50   Median :65.00   Median :10.694
           Mean   :34.99   Mean   :60.33   Mean   :10.696
           3rd Qu.:57.65   3rd Qu.:77.30   3rd Qu.:11.989
           Max.   :99.70   Max.   :90.10   Max.   :21.480
```

Note that FAIR plan policies are mostly given after normal insurance is denied, so areas with higher rates of such policies are potentially subject to more redlining.

13 Question 13

Our first hypothesis is that zip codes in this sample with high rates of non-white residents are more likely than other zip codes to be redlined in this way, even after accounting for the age of the housing stock in the zip code.

Fit a logistic regression model, which we'll call `m13`, to predict the probability of a zip code having a High `fair` value that will allow you to assess this first hypothesis effectively, ignoring for now the `medinc` and `side` information.

To demonstrate that you've done this, specify the estimated odds ratio for a High `fair` rate that is associated with a change in `minority` from 3.75% to 57.65%, while holding `older` constant, along with a 95% confidence interval for that odds ratio. Round your answers to two decimal places.

14 Question 14

Interpret the meaning of the point estimate you provided in Question 13 in a clear and complete English sentence (or two.)

15 Question 15

Now consider a new model where you will add two additional predictors (`side` and `medinc`) to the model you built in Question 13. Suppose that a main effects model of that type will use K degrees of freedom. Suppose you are willing to add a non-linear term to your model that uses no more than two additional degrees of freedom to your total of K from the main effects model. According to an appropriate Spearman ρ^2 plot in this scenario, which of the following additions to the model is the best option?

a. Add a restricted cubic spline with 4 knots in `minority`
b. Add a restricted cubic spline with 4 knots in `older`
c. Add a restricted cubic spline with 4 knots in `medinc`
d. Add an interaction of `side` with `medinc`
e. Add an interaction of `side` with `minority`
f. Add an interaction of `side` with `older`

16 Question 16

Fit the model suggested by your proposed addition in Question 15, and call the model `m16`, then obtain and specify bootstrap-validated estimates (using default choices and `set.seed(2022)`) from `m16` of the following summary statistics, rounded to three decimal places each:

a. the Nagelkerke R^2
b. the area under the ROC curve

17 Question 17

In this question, you will use your models fit in Question 13 (`m13`) and in Question 16 (`m16`) to make estimates for a fictional new zip code on the South side of Chicago, which has `minority` = 55%, `older` = 70%, and `medinc` = 10.

State the resulting probability (as a proportion rounded to two decimal places) that this fictional new zip code has a High `fair` rate…

a. according to the two-predictor model (`m13`) you fit in Question 13
b. according to the more complex model (`m16`) you fit in Question 16

18 Question 18

In *The Signal and the Noise*, Nate Silver encourages his readers to behave like foxes, rather than hedgehogs, when forecasting. Which of the following statements describe "foxes"? (Select all that apply.)

a. Someone willing to acknowledge mistakes in their predictions.
b. Someone who creates probabilistic forecasts.
c. Someone who expects the world to abide by relatively simple relationships once the signal has been separated from noise.
d. Someone who is willing to bet on their forecasts.
e. Someone who establishes a planned approach at the start, and uses new data just to refine the original plan.
f. None of these statements.

Output for Questions 19-21

Questions 19-21 involve the same study. In attempting to measure the complex relationships between four potential treatments and primary insurance on a summary measure of health obtained after treatment among 360 Northeast Ohio residents, two linear models were developed, called Model `m19a` and Model `m19b`. Each of the 360 subjects received exactly one of the four Treatments (although Treatments A and B were selected more often than C or D), and the sample was obtained to include equal numbers of Medicare, Medicaid and Commercially insured subjects. I have not provided the data for this example, but have given you some summary results below.

`describe` for the `set19` data

```
describe(set19)
```

```
set19

 4  Variables      360  Observations
--------------------------------------------------------------------------------
subject
       n  missing distinct     Info     Mean      Gmd      .05      .10
     360        0      360        1    180.5    120.3    18.95    36.90
     .25      .50      .75      .90      .95
   90.75   180.50   270.25   324.10   342.05

lowest :   1   2   3   4   5, highest: 356 357 358 359 360
--------------------------------------------------------------------------------
treatment
       n  missing distinct
     360        0        4

Value          A     B     C     D
Frequency    120   120    60    60
Proportion 0.333 0.333 0.167 0.167
--------------------------------------------------------------------------------
insurance
       n  missing distinct
     360        0        3

Value      Commercial   Medicaid   Medicare
Frequency         120        120        120
Proportion      0.333      0.333      0.333
--------------------------------------------------------------------------------
health
       n  missing distinct     Info     Mean      Gmd      .05      .10
     360        0      360        1    180.7    57.51     92.0    118.3
     .25      .50      .75      .90      .95
   149.5    180.1    213.2    242.1    262.4

lowest :  33.20560  36.94144  49.16756  49.36557  51.21994
highest: 304.51666 312.66653 328.61451 340.38944 362.45856
--------------------------------------------------------------------------------
```

Output for Questions 19-21 continues on the next page.

Output for Questions 19-21 (continued)

anova() results after fitting our two models

```
anova(m19a)
```

```
Analysis of Variance Table

Response: health
           Df Sum Sq Mean Sq F value  Pr(>F)
treatment   3  24462  8154.1  3.1503 0.02509 *
insurance   2  22368 11184.1  4.3209 0.01400 *
Residuals 354 916276  2588.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m19a, m19b)
```

```
Analysis of Variance Table

Model 1: health ~ treatment + insurance
Model 2: health ~ treatment * insurance
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    354 916276
2    348 869941  6     46335 3.0892 0.005841 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(m19a, m19b)
```

```
     df      AIC
m19a  7 3858.744
m19b 13 3852.063
```

```
BIC(m19a, m19b)
```

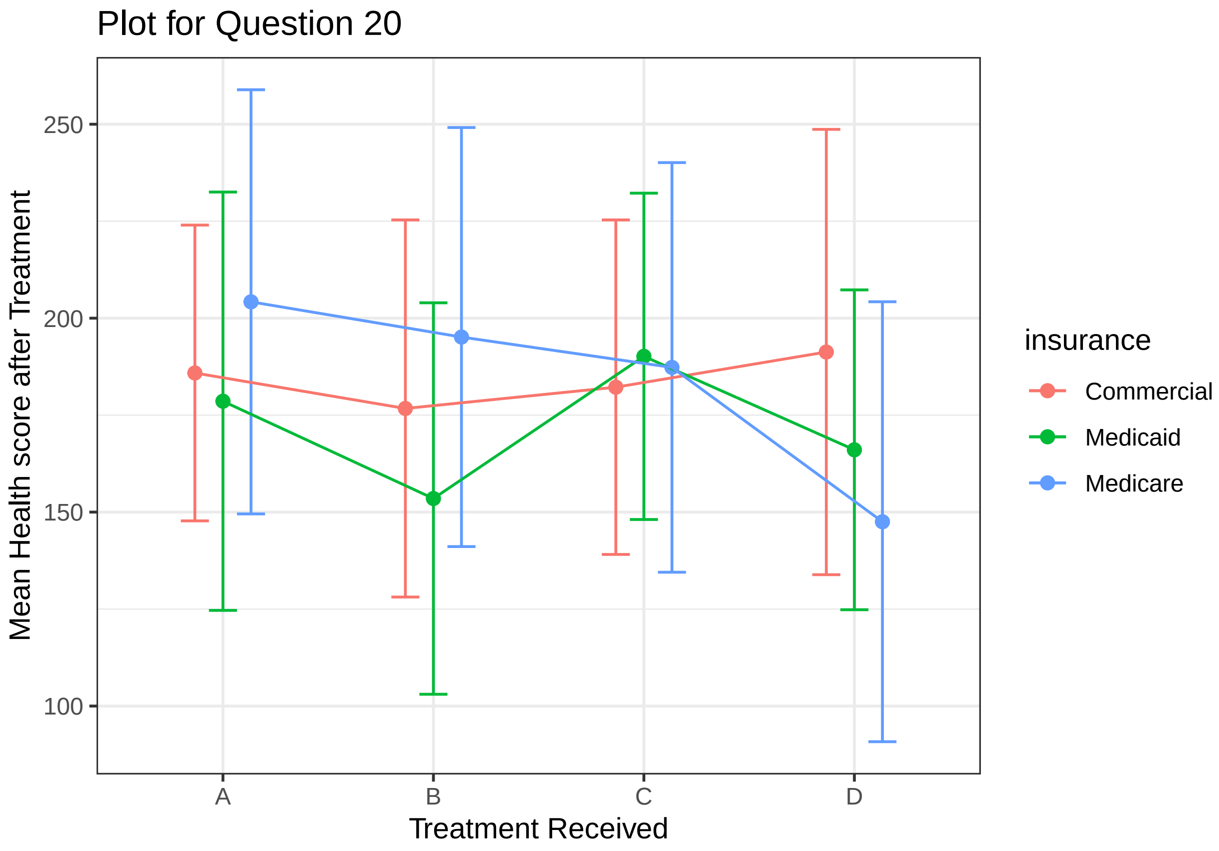
```
     df      BIC
m19a  7 3885.947
m19b 13 3902.583
```

19 Question 19

Question 19 has three parts. Please answer all three.

a. In a sentence, tell me what is included in model `m19b` but not in `m19a`?
b. Which model (`m19a` or `m19b`) looks better according to the Akaike Information Criterion?
c. Which model (`m19a` or `m19b`) looks better according to the Bayes Information Criterion?

Plot for Question 20



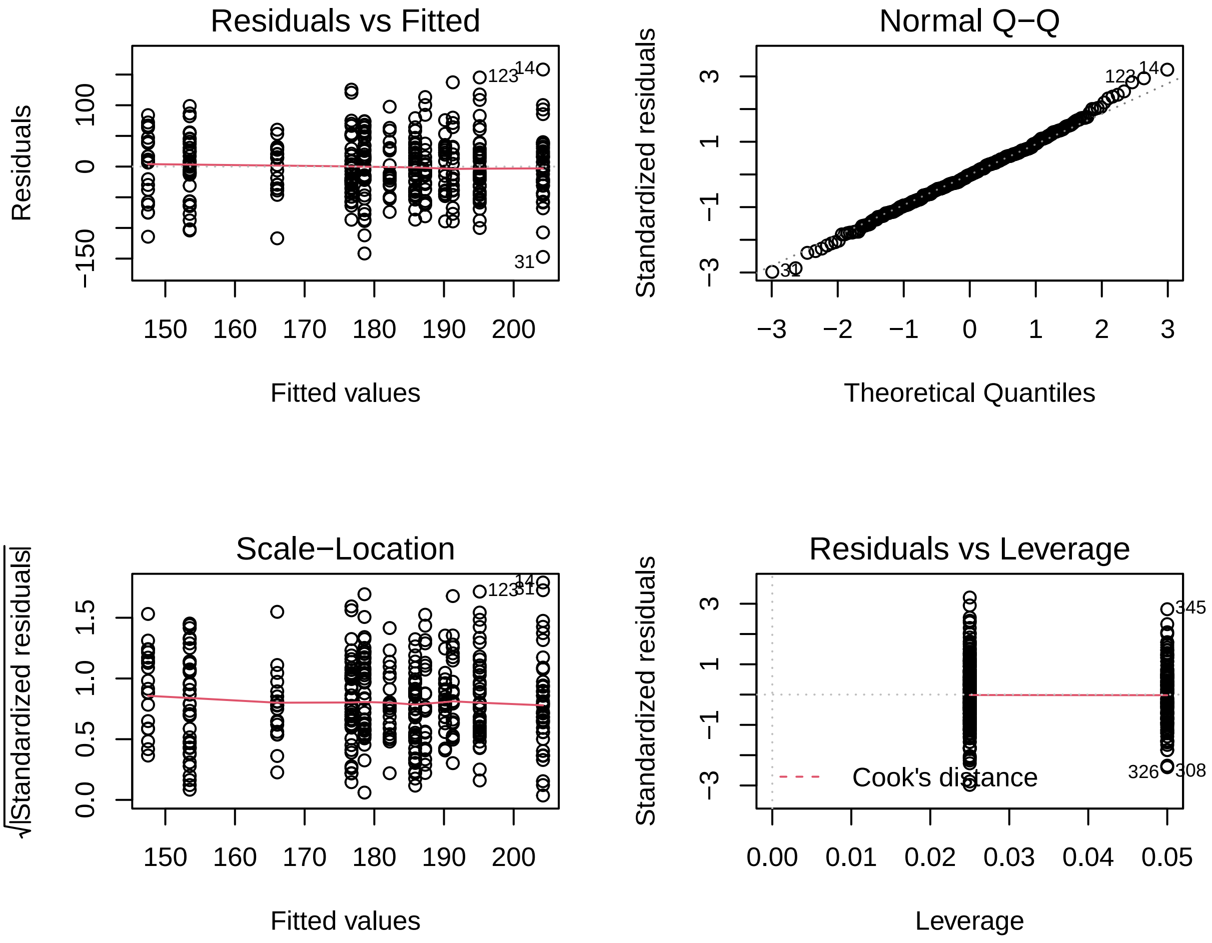
20 Question 20

What does the Plot for Question 20 of means with intervals indicating one standard deviation in either direction suggest about the best choice of model, comparing `m19a` to `m19b`?

a. `m19a` seems like the better choice.
b. `m19b` seems like the better choice.
c. This plot does not help us make the decision.
d. It is impossible to tell from the information provided.

Plots for Question 21

```
par(mfrow = c(2,2)); plot(m19b); par(mfrow = c(1,1))
```



21 Question 21

Consider the residual plots for model `m19b` shown as the Plots for Question 21. What conclusions can you draw about the validity of regression assumptions in this case? (Check all that apply.)

a. These residual plots display serious problems with collinearity.
b. These residual plots display serious problems with Normality.
c. These residual plots display serious problems with non-constant variance.
d. These residual plots display serious problems with highly leveraged points.
e. These residual plots display serious problems with influential points.
f. These residual plots display serious problems with linearity.
g. These plots display no serious problems with regression assumptions.

Code Attempt for Question 22

Here's the code that's OK

```
## THIS CODE IS OK

set22 <- read_csv(here("data/set22.csv"),
                  show_col_types = FALSE) %>%
    type.convert(as.is = FALSE) %>%
    clean_names()
```

Here's the code that's giving us trouble

```
set22 <- set22 %>%
    mutate(gov_ins = factor(insurance,
                         Medicare or Medicaid = Yes,
                         Commercial or Uninsured = No))
```

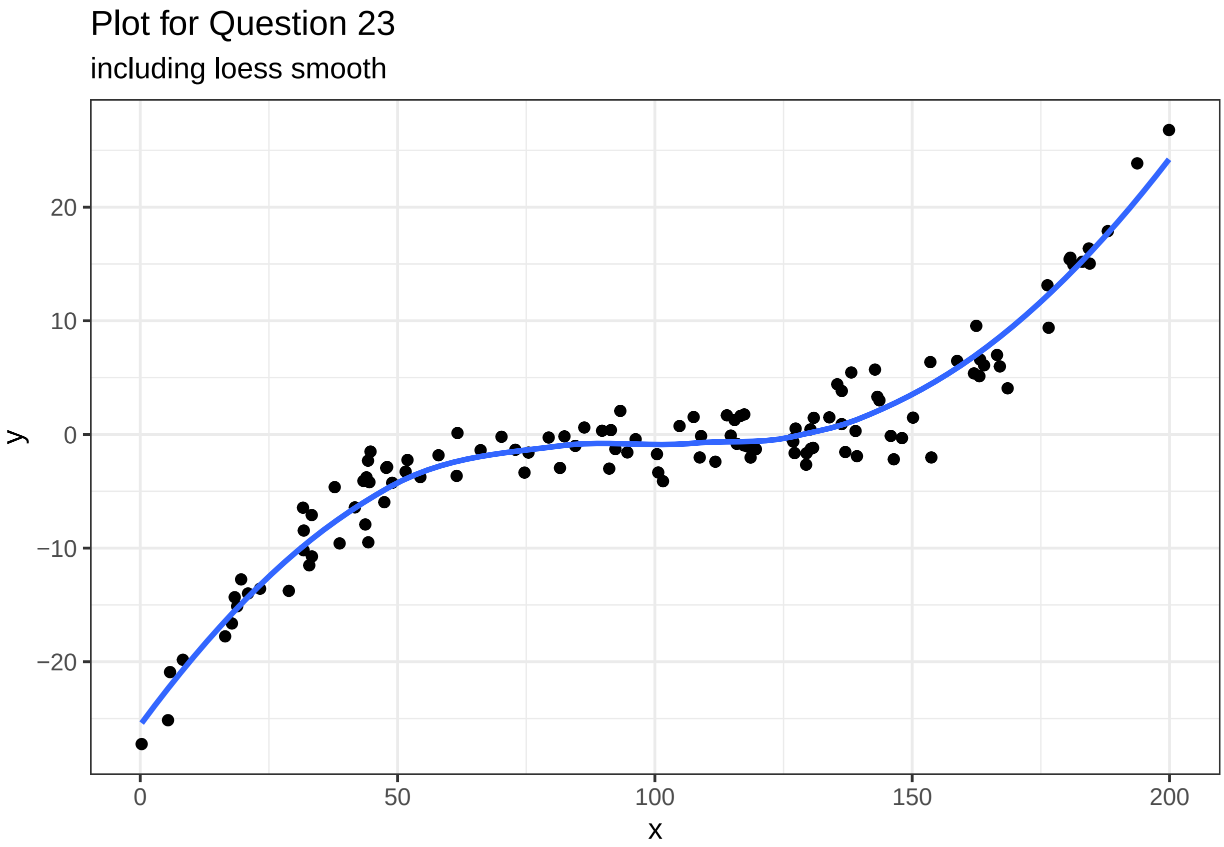
22 Question 22

We have provided a file called `set22.csv` as part of the available data for this Quiz. Among other things, it contains insurance data on hundreds of subjects, each of whom is classified as falling into one of four different insurance categories, specifically Medicare, Commercial, Medicaid, and Uninsured.

Suppose you have ingested these data into R, and now want to create a variable called `gov_ins` within a `set22` tibble that (a) is a factor, and (b) which takes the value Yes if the subject's insurance is provided by the government (Medicare or Medicaid) but No otherwise, while (c) retaining NA for missing values (if any.) Your first attempt to do this is shown in the Code Attempt for Question 22. Assume that all necessary packages have been loaded in R. Fix the call to the `mutate` function so that your resulting code will actually do what is required, when it follows the code chunk marked as OK.

Hint: Your response should begin with `mutate(`.

Plot for Question 23



23 Question 23

Suppose the relationship between a predictor, `x`, and an outcome, `y`, is described by the Plot for Question 23, which includes the fit from a loess smooth. What is the **minimum** number of knots that would be required in a restricted cubic spline on `x` to fit a model predicting `y` that approximates the general shape of the curve shown in the plot?

a. Less than three knots would be required.
b. At least Three knots would be required
c. At least Four knots would be required.
d. Five or more knots would be required.
e. It is impossible to tell from the information provided.

24 Question 24

In the first five chapters of *The Signal and the Noise*, Nate Silver reports on a program by the National Weather Service to evaluate how well its predictions regarding precipitation and temperature perform. In particular, the NWS compared the accuracy of its forecasts using computers alone to its forecasts made with a combination of computer models and humans.

Based on those comparisons as reported by Silver, which of the following statements are accurate? (Check all of the true statements.)

a. Temperature forecasts made by computers alone are 10% better than those where humans are involved.
b. Temperature forecasts where humans are involved are 10% better than those made by computers alone.
c. Precipitation forecasts made by computers alone are 25% better than those where humans are involved.
d. Precipitation forecasts where humans are involved are 25% better than those made by computers alone.
e. The impact on forecast quality associated with adding humans to the computer models has improved substantially in recent years.
f. The impact on forecast quality associated with adding humans to the computer models has remained very consistent in recent years.
g. The impact on forecast quality associated with adding humans to the computer models has declined substantially in recent years.
h. None of the statements listed above are correct.

This is the end of the Quiz.

Be sure to complete the Affirmation at the end of the Answer Sheet, and that you have submitted your Answer Sheet, and received your copy in your CWRU email by the deadline.