432 Quiz 1 for Fall 2019

Thomas E. Love

due 2019-03-05 at 7 AM. Version 2019-02-27 09:21:21

Introduction

- There are 25 questions. Each question is worth between 2 and 5 points, with partial credit awarded on some questions (so don't leave anything blank.) The maximum score is 60 points.
- The Google Form to submit your responses will be found at http://bit.ly/432-2019-quiz1-form. All of your responses must be submitted using that Form.
- You are not permitted to discuss the Quiz with anyone other than Dr. Love and the teaching assistants.
- If you have questions, we will be happy to answer them, if you submit them via email to `431-help at case dot edu` before 7 PM on Monday 2019-03-04. We will not guarantee a response after that time.
- If you need to set a seed to respond to a question (and you will), use `set.seed(2019)` every time.
- Remember that the Google Form will close and the Quiz is due at **7 AM on Tuesday 2019-03-05**.

What Will You Need?

To do this quiz, you'll need to load the following packages in R. At any rate, I used them in creating the answer sketch. (I also used `GGally` and `gridExtra`, actually.)

```
car
leaps
modelr
rms
broom
janitor
tidyverse
```

Other packages may be useful, but are not required. You'll also need the following data sets and chunks of code, each of which is provided in the Quiz 1 section of our web site.

```
bootdif.R
childfev.csv
data24.Rds
dbptrial.csv
riff1.csv
```

You'll need to source in the `bootdif.R` function, and load in the four data sets.

Setup for Questions 1-8

The `dbptrial` data set is typical of diastolic blood pressure (DBP) data from small clinical trials in hypertension as completed 40 or 50 years ago. During this time, hypertension was even more severe than it is now, as the number of effective treatments was relatively small, and the definition (DBP $>$ 95 mm Hg) was not as stringent as it is now (DBP $>$ 80 mm Hg.)

In this randomized clinical trial, diastolic blood pressure (DBP) was measured (in mm Hg) in the supine position at baseline (`DBP0`) before randomization and monthly thereafter for four months (`DBP1`, `DBP2`, `DBP3`, and `DBP4`.) Each subject's `Age` (in years) and `Sex` were recorded at baseline. Our primary objective is to test whether Treatment A (a new drug) may be more effective in lowering DBP as compared to Treatment B (a placebo.)

1 Question 1

```
library(janitor); library(tidyverse)

dbptrial <- read_csv("data/dbptrial.csv") %>% clean_names()
dbptrial
```

```
# A tibble: 40 x 9
   subject_code treatment  dbp0  dbp1  dbp2  dbp3  dbp4   age sex
          <dbl> <chr>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
 1            1 A           114   115   113   109   105    43 F
 2            2 A           116   113   112   103   101    51 M
 3            3 A           119   115   113   104    98    48 F
 4            4 A           115   113   112   109   101    42 F
 5            5 A           116   112   107   104   105    49 M
 6            6 A           117   112   113   104   102    47 M
 7            7 A           118   111   100   109    99    50 F
 8            8 A           120   115   113   102   102    61 M
 9            9 A           114   112   113   109   103    43 M
10           10 A           115   113   108   106    97    51 M
# ... with 30 more rows
```

In loading the `dbptrial` data, I used `clean_names()`, which made some changes to the variable names in the data set as compared to what you see in the `dbptrial` data in `.csv` format. Which of the following variables **did not** change? (Note that more than one response may be selected.)

a. The variable describing each subject's diastolic blood pressure before randomization
b. The variable describing the age of the subject
c. The variable describing the subject's sex
d. The variable describing whether the subject received the new drug or the placebo
e. The subject identification code
f. None of the above

2 Question 2

Your job in Question 2 is to create a new variable, called `dbp_diff`, which represents the change in DBP between the baseline result and the result obtained at the final visit four months after baseline. Calculate your `dbp_diff` value so that `dbp_diff` $= 9$ indicates a subject whose DBP at their baseline visit was 9 mm Hg larger than their result at their visit after month 4.

Specify the R command you used to generate this `dbp_diff` variable inside the `dbptrial` data.

In providing your response, assume that all necessary packages have been loaded with `library` and that the `dbptrial <- read_csv("data/dbptrial.csv") %>% clean_names()` command has already been run.

3 Question 3

Build a boxplot (using `ggplot`) to compare the `dbp_diff` values for the new drug (treatment A) to the `dbp_diff` values for the placebo (treatment B). According to your boxplot, which of the following statements are true? (Note that more than one response may be selected.)

a. The median DBP change in the new treatment is larger than the median DBP change in the placebo.
b. The median DBP change in the new treatment is smaller than the median DBP change in the placebo.
c. The interquartile range of the DBP changes in the new treatment is substantially larger than the placebo.
d. There is substantial overlap in the distributions of DBP changes across the two treatment groups.
e. There is at least one apparent outlier candidate in the DBP changes in each treatment group.

4 Question 4

Suppose you want to compare the mean `dbp_diff` scores across the two treatment groups. Obtain an appropriate 95% bootstrap confidence interval for the true difference in mean `dbp_diff` values across the two treatments. Use `set.seed(2019)`.

In a single sentence, specify your point estimate and 95% confidence interval estimate in terms of mm Hg **and** specify what it tells you about the comparison of interest.

Notes:

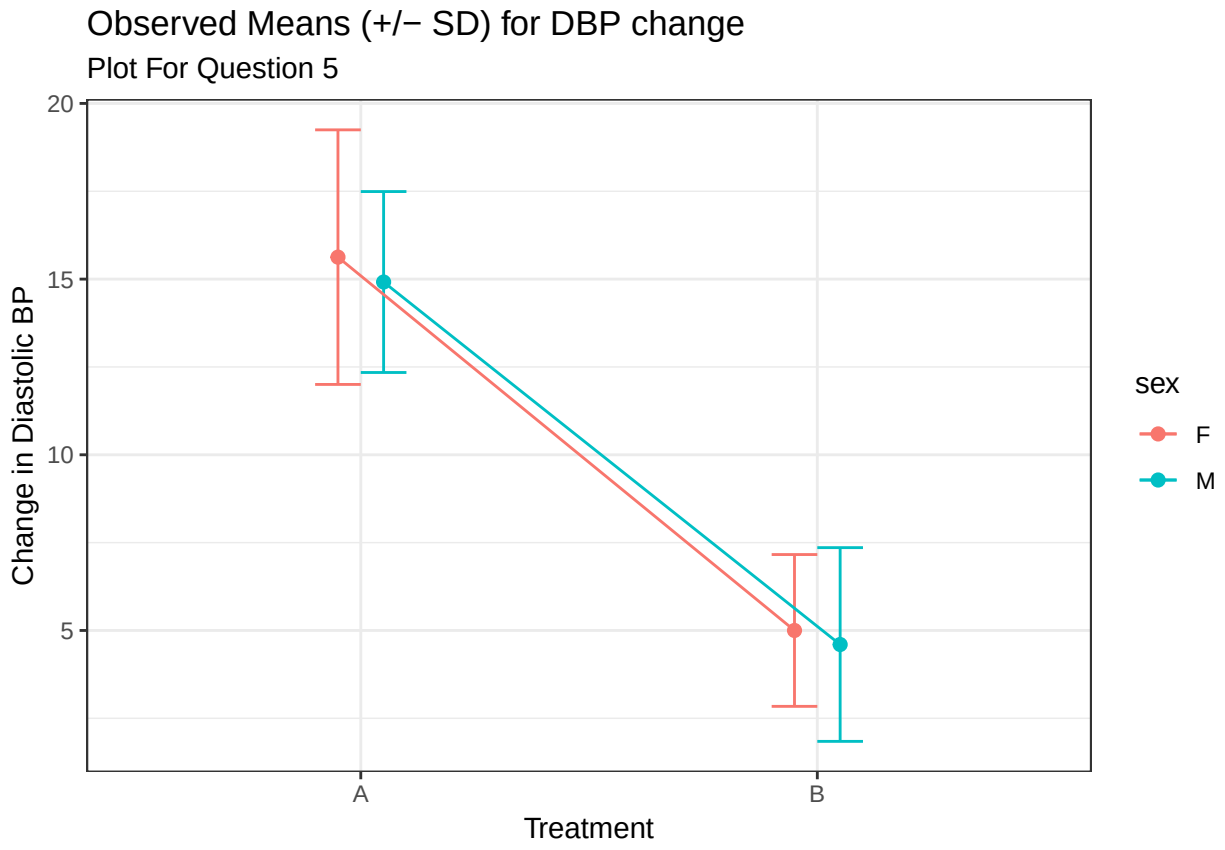
- Round all values in your response to two decimal places.
- Ensure that **positive** numbers indicate that the new treatment (A) is more effective than the placebo (B).

5 Question 5

Here's a plot I created and the code used to build it.

```
summaries_q5 <- dbptrial %>% group_by(treatment, sex) %>%
    summarise(n = n(), mean = mean(dbp_diff),
              stdev = sd(dbp_diff))

pd <- position_dodge(0.2)
ggplot(summaries_q5, aes(x = treatment, y = mean,
                         col = sex)) +
    geom_errorbar(aes(ymin = mean - stdev,
                      ymax = mean + stdev),
                  width = 0.2, position = pd) +
    geom_point(size = 2, position = pd) +
    geom_line(aes(group = sex), position = pd) +
    theme_bw() +
    labs(y = "Change in Diastolic BP",
         x = "Treatment",
         title = "Observed Means (+/- SD) for DBP change",
         subtitle = "Plot For Question 5")
```



See the top of the next page for the question.

Question 5. (continued) Which of the following conclusions is most appropriate, in light of the plot for Question 5?

a. There is no substantial interaction, and females appear to have smaller differences in DBP from baseline to four months later than did males, regardless of treatment.
b. There is no substantial interaction, and females appear to have larger differences in DBP than did males, regardless of treatment.
c. There is a substantial interaction, and in the new treatment, females appear to have smaller DBP differences than did males.
d. There is a substantial interaction, and in the new treatment, females appear to have larger DBP differences than did males.
e. None of these conclusions are appropriate.

6 Question 6

Fit an appropriate ANOVA model (taking into account what you learned in Question 5) to describe the impact of `sex` and `treatment` on the `dbp_diff` outcome.

Specify the η^2 value for the combined impact of all predictors in this ANOVA model.

Note:

- Your answer should be a **number**, expressed as percentage, and rounded to one decimal place.
- η^2 is read as "eta-squared".

7 Question 7

Use the `dbptrial` data set to assess the relationship between the subject's Age and the Treatment they received. Is the association of Age with treatment that we observe consistent with what we should expect from this randomized controlled clinical trial?

a. No, because there is a significant or substantial difference between the ages of those treated with the new drug and the ages of the placebo subjects.
b. Yes, because there is a significant or substantial difference between the ages of those treated with the new drug and the ages of the placebo subjects.
c. No, because there is no significant and substantial difference between the ages of those treated with the new drug and the ages of the placebo subjects.
d. Yes, because there is no significant and substantial difference between the ages of those treated with the new drug and the ages of the placebo subjects.
e. It is impossible to tell.

8 Question 8

Three models of potential interest are specified below. The `age_c` results are *centered* ages, where the mean age across all subjects has been subtracted from that subject's age, so an `age_c` value of 10 indicates a subject who is 10 years older than the average subject.

```
anova(model1, model2, model3)
```

```
Analysis of Variance Table

Model 1: dbp_diff ~ age_c * treatment + age_c * sex + treatment * sex
Model 2: dbp_diff ~ age_c * treatment + age_c * sex
Model 3: dbp_diff ~ age_c + treatment + sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     33 198.47
2     34 199.79 -1   -1.3256 0.2204 0.6418
3     36 226.28 -2  -26.4842 2.2018 0.1266
```

Suppose you want to use this output to get information about whether the following statement is reasonable.

- `STATEMENT`: **The association of treatment with differences in DBP from baseline to four-month follow-up is not substantially affected by the gender of the subject, even after adjusting for age.**

Which of the following responses is most correct?

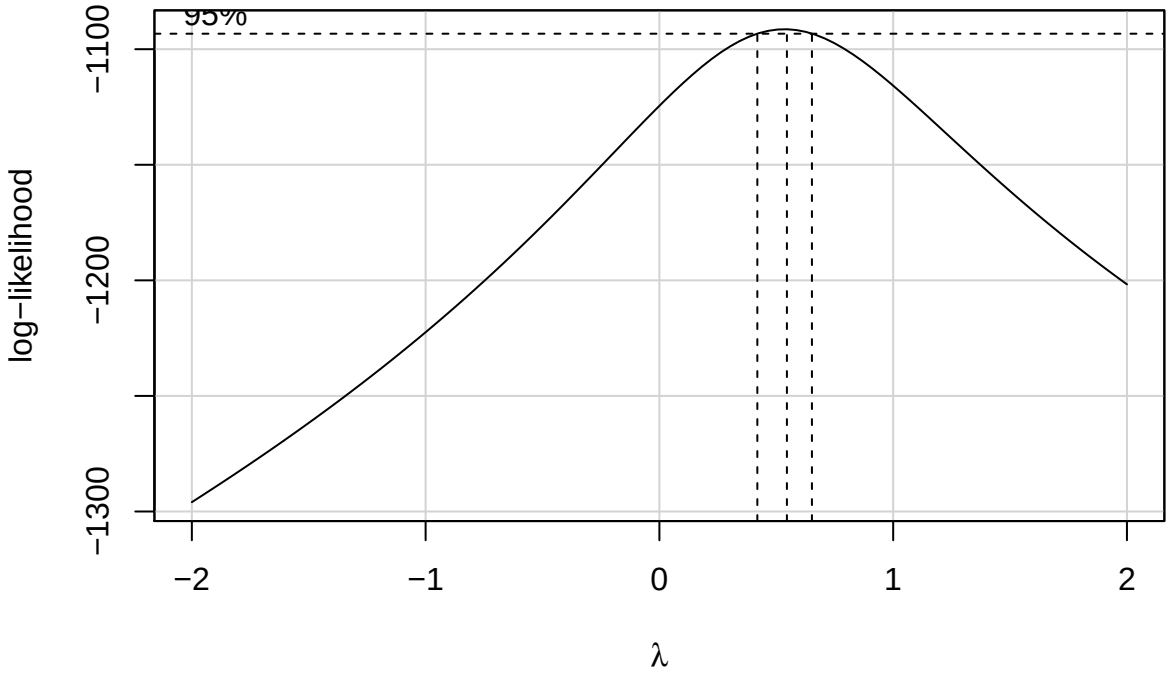
a. We should compare Model 1 to Model 2, the appropriate p value is 0.1266, so this supports the `STATEMENT` as being true.
b. We should compare Model 1 to Model 2, the appropriate p value is 0.1266, so this fails to support the `STATEMENT`.
c. We should compare Model 1 to Model 2, the appropriate p value is 0.6418, so this supports the `STATEMENT` as being true.
d. We should compare Model 1 to Model 2, the appropriate p value is 0.6418, so this fails to support the `STATEMENT`.
e. We should compare Model 1 to Model 3, the appropriate p value is 0.6418, so this supports the `STATEMENT` as being true.
f. We should compare Model 1 to Model 3, the appropriate p value is 0.6418, so this fails to support the `STATEMENT`.
g. We should compare Model 2 to Model 3, the appropriate p value is 0.1266, so this supports the `STATEMENT` as being true.
h. We should compare Model 2 to Model 3, the appropriate p value is 0.1266, so this fails to support the `STATEMENT`.
i. There is no way to tell from the information and output provided.

9 Question 9

Which of the following variables can be appropriately used as an outcome in a linear regression model? (Note that more than one response may be selected.)

a. The lifespan of a single-celled organism (for example, an amoeba), in minutes.
b. Whether the amoeba is predatory, or instead is a detrivore, that eats dead organic material.
c. The probability that an amoeba has infected a host organism pathogenically to cause amoebic dysentery.
d. The location in which our amoeba is found, (protozoa, fungus, algae or animal).
e. None of these.

10 Question 10



Which of the following bits of R code does this Box-Cox plot suggest will yield a good-fitting linear regression?

a. `lm(y ~ x)`
b. `lm(y ~ log(x))`
c. `lm(y ~ sqrt(x))`
d. `lm(y ~ 1/x)`
e. `lm(sqrt(y) ~ x)`
f. `lm(log(y) ~ x)`
g. `lm(1/y ~ x)`
h. `lm(log(y) ~ log(x))`
i. None of these models.

Setup for Questions 11-18

301 male patients were examined. Each exhibited one of several reasons to suspect problems with their prostate glands. These data are available in the `riff1.csv` data set. For each patient, the following data are provided:

- `ptnum` = patient identification code
- `age` = age (in years)
- `dre` = digital rectal examination result (0 = negative, 1 = positive)
- `tru` = transurethral ultrasound result (0 = negative, 1 = positive)
- `psa` = prostate-specific antigen level (in ng/ml)
- `vol` = volume of prostate (in ml)
- `psad` = prostate-specific antigen density level (this is just psa / vol)
- `biopsy` = biopsy result (0 = negative, 1 = positive)

```
riff1
```

```
# A tibble: 301 x 8
   ptnum   age   dre   tru   psa   vol  psad biopsy
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
 1     1    75     0     1   7.6  32.3  0.24      0
 2     2    68     1     0   4.1  27.0  0.15      0
 3     3    54     0     0   5.9  16.2  0.37      1
 4     4    62     1     1   9    33.0  0.27      1
 5     5    61     0     0   6.8  30.9  0.22      1
 6     6    61     0     1   8    73.7  0.11      0
 7     7    62     0     0   7.7  30.5  0.25      0
 8     8    61     1     1   4.4  30.5  0.14      0
 9     9    73     1     1   6.1  36.8  0.17      0
10    10    74     1     1   7.9  16.4  0.48      0
# ... with 291 more rows
```

The outcome which we are interested in predicting is the `biopsy` result, which we will assume indicates the "truth" in this case as to whether the patient actually has prostate cancer.

11 Question 11

To begin, build a regression model to predict whether the patient actually has prostate cancer on the basis of their PSA level, prostate volume, transurethral ultrasound result, digital rectal examination result and age.

Which predictors show a statistically significant effect (at the 5% level) on the model, using Wald tests? (Note that more than one response may be selected.)

a. the subject's age
b. the result of the subject's digital rectal exam
c. the result of the subject's transurethral ultrasound
d. the subject's prostate volume
e. the subject's prostate-specific antigen level
f. None of the above

12 Question 12

Use a stepwise approach in an attempt to select a subset of predictors from the model you fit in Question 11. How many predictors remain in your new, smaller model?

a. 1
b. 2
c. 3
d. 4
e. 5

13 Question 13

Which of the following best describes the area under the receiver operating characteristic curve for the reduced model you fit in Question 12?

a. The area is less than 0.5
b. The area is between 0.50 and 0.699
c. The area is between 0.70 and 0.799
d. The area is between 0.80 and 0.899
e. The area is 0.9 or larger

14 Question 14

Which of the following statements best describes your interpretation of the area under the ROC curve, and the Nagelkerke R^2 results you obtained for the model you fit in Question 12?

a. The model displays essentially no predictive value, nor does it effectively discriminate cancer patients from those who are cancer-free.
b. The model displays only modest predictive value, and does not effectively discriminate cancer patients from those who are cancer-free.
c. The model displays only modest predictive value, and effectively discriminates cancer patients from those who are cancer-free.
d. The model displays strong predictive value, and effectively discriminates cancer patients from those who are cancer-free.
e. None of these statements are correct.

15 Question 15

A new patient, Mr. Smith, had a PSA level of 9.6 ng/ml, a prostate volume of 60 ml, positive results on both the digital rectal exam, and the transurethral ultrasound, and was 68 years old. What is the predicted probability of a positive biopsy for Mr. Smith, according to the reduced model you built in Question 12? Round your response to two decimal places, and present it as a proportion (between 0 and 1), not a percentage.

16 Question 16

Below, you'll see the results of running a validation for the five-predictor model fit back in Question 11.

```
set.seed(2019); validate(mymodel)
```

```
          index.orig training    test optimism index.corrected  n
Dxy           0.4765   0.5087  0.4595   0.0492          0.4273 40
R2            0.2413   0.2714  0.2250   0.0463          0.1950 40
Intercept     0.0000   0.0000 -0.1113   0.1113         -0.1113 40
Slope         1.0000   1.0000  0.8559   0.1441          0.8559 40
Emax          0.0000   0.0000  0.0545   0.0545          0.0545 40
D             0.1854   0.2118  0.1715   0.0402          0.1452 40
U            -0.0066  -0.0066  0.0047  -0.0113          0.0047 40
Q             0.1921   0.2184  0.1668   0.0516          0.1405 40
B             0.1768   0.1712  0.1805  -0.0094          0.1862 40
g             1.3804   1.5842  1.2844   0.2998          1.0806 40
gp            0.2026   0.2173  0.1945   0.0228          0.1798 40
```

Use this output to specify (to three decimal places) the most appropriate estimate of the area under the ROC curve that we would expect to see in new data.

17 Question 17

Consider the following two-predictor model for the same `riff1` data. In a sentence, state **and interpret** the odds ratio for `psa` in this model. Be sure to carefully describe the comparison made by the odds ratio.

```
model_17 <- glm(biopsy ~ tru + psa, family = "binomial", data = riff1)
summary(model_17)
```

```
Call:
glm(formula = biopsy ~ tru + psa, family = "binomial", data = riff1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9352  -0.8320  -0.6491   1.1155   1.9654

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.80729    0.24465  -7.387  1.5e-13 ***
tru          0.75402    0.26941   2.799 0.005131 **
psa          0.08113    0.02138   3.794 0.000148 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.36  on 300  degrees of freedom
Residual deviance: 333.27  on 298  degrees of freedom
AIC: 339.27

Number of Fisher Scoring iterations: 6
```

18 Question 18

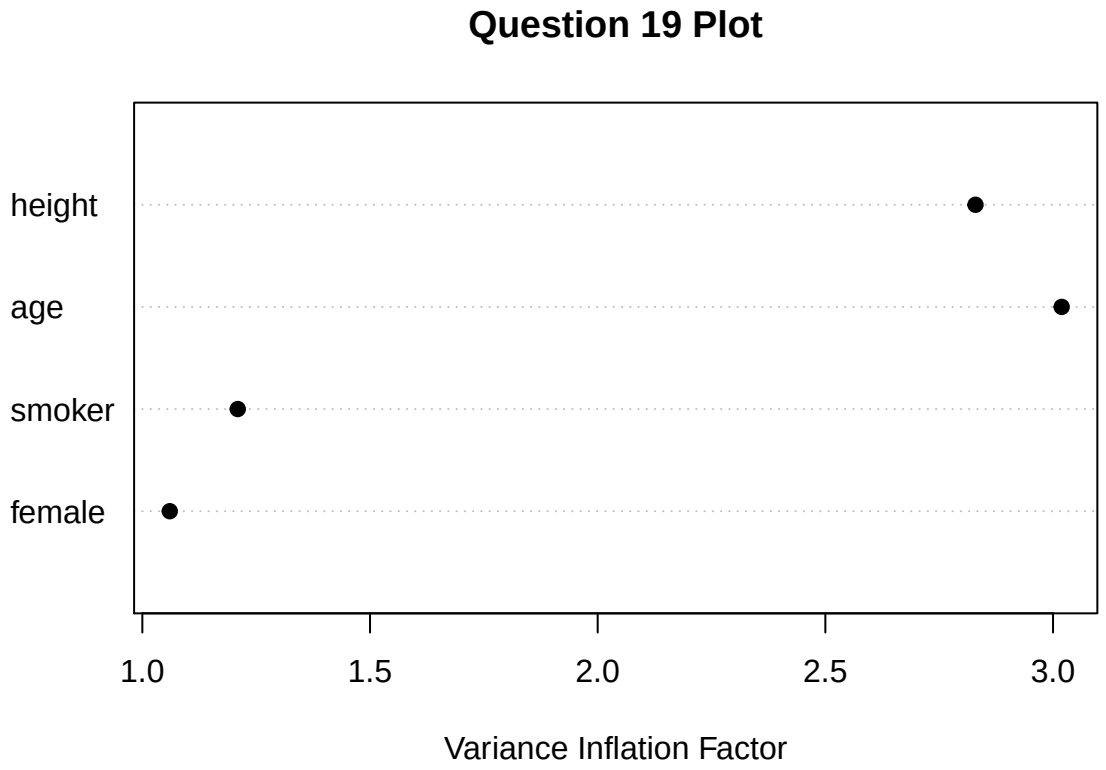
In a sentence, state **and interpret** an approximate 95% confidence interval for `tru` in the model fit in Question 17. You should calculate the result using the information provided. Specify the lower and upper bounds for the confidence interval after rounding to two decimal places.

Setup for Questions 19 - 22

The `childfev` data set provided to you on our web site contains information on FEV1 (forced expiratory volume in one second), which measures the volume (in liters) of air expelled after one second of effort, on 654 children ages 6-22 who were seen in the Childhood Respiratory Disease Study, reported in 1980 in East Boston, MA. Potential predictors of FEV1 in the data include the child's age in years, their height in inches, their sex (indicated by a 1/0 variable called `female` which is 1 for females and 0 for males) and whether or not the child is a current smoker.

19 Question 19

We fit a main effects model to the `childfev` data to predict FEV1, and then obtained the following plot of variance inflation factors.



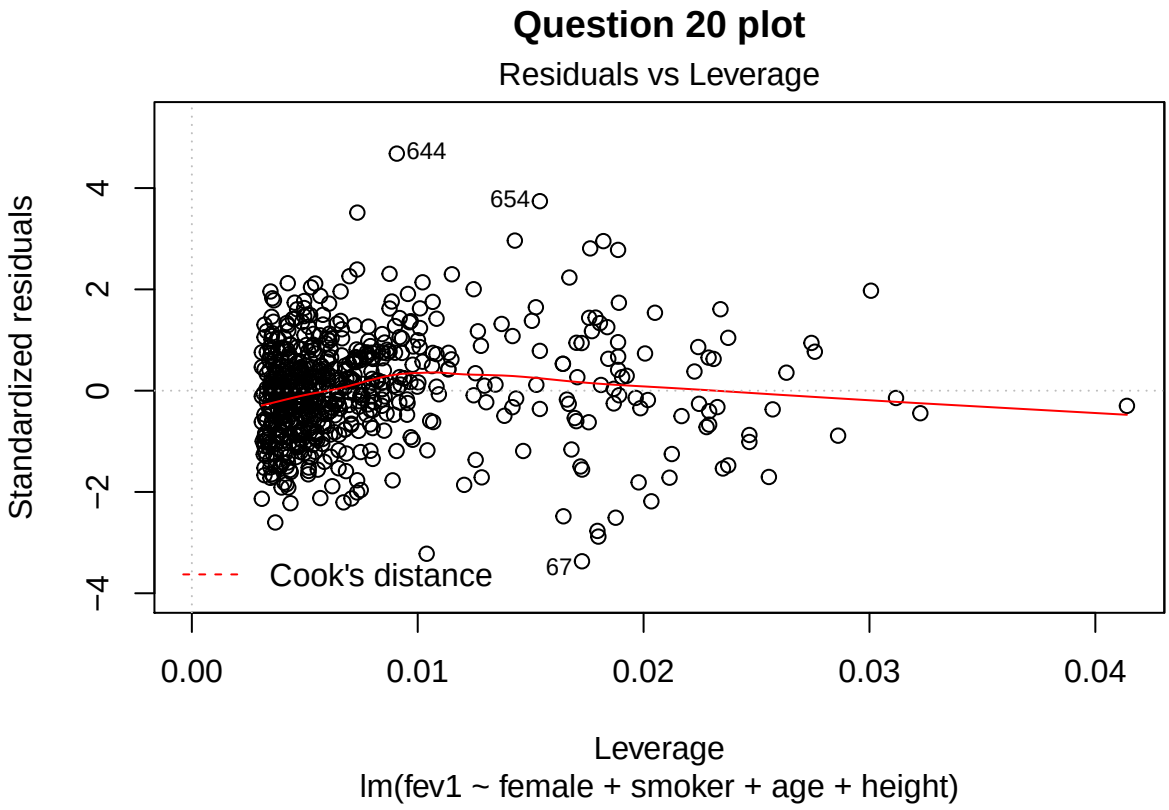
Which of the following is directly assessed by the Question 19 plot?

a. This is a check to see if multi-collinearity is an issue.
b. The assumption of linearity.

c. The assumption of homoscedasticity (constant variance).
d. This is a check of whether the model is parsimonious.
e. The assumption of independence.
f. This is a check of whether the model is well-calibrated.
g. The assumption of Normality.
h. This is a check on the influence of the observations on our model.

20 Question 20

The plot below is also based on the main effects model we discussed in Question 19.

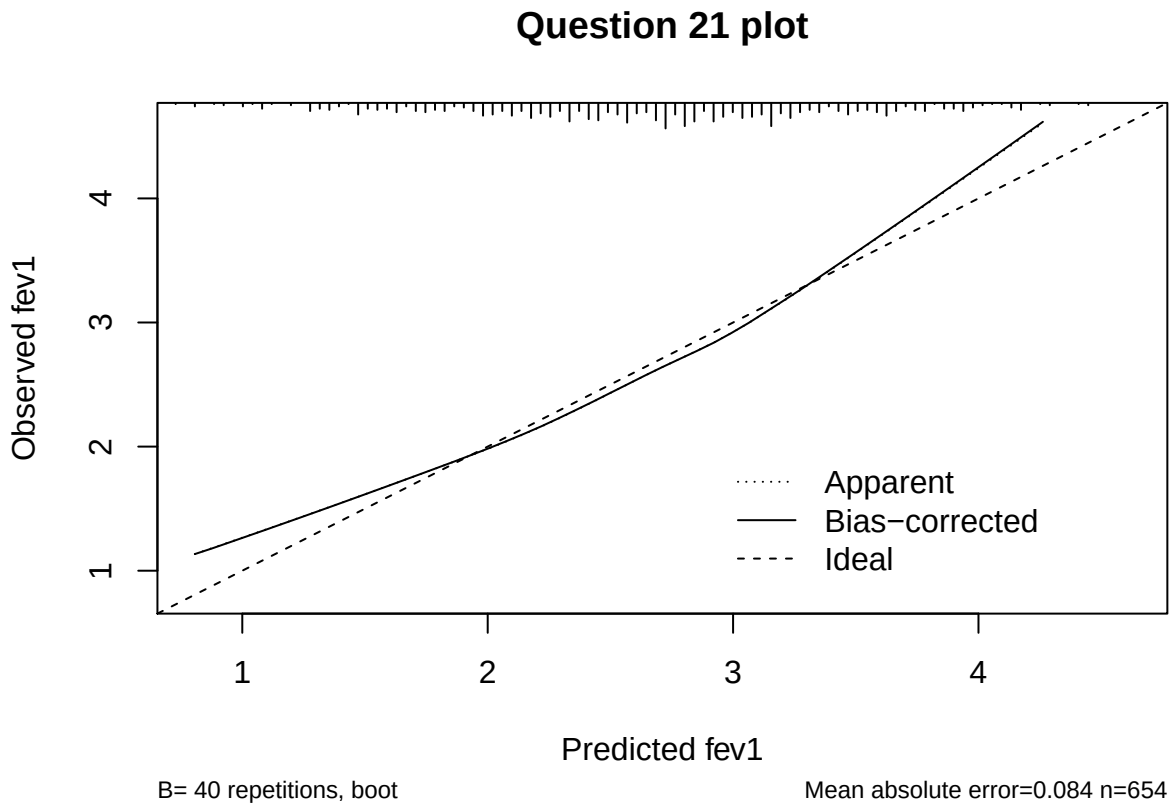


Which of the following is directly assessed by the Question 20 plot?

a. This is a check to see if multi-collinearity is an issue.
b. The assumption of linearity.
c. The assumption of homoscedasticity (constant variance).
d. This is a check of whether the model is parsimonious.
e. The assumption of independence.
f. This is a check of whether the model is well-calibrated.
g. The assumption of Normality.
h. This is a check on the influence of the observations on our model.

21 Question 21

The plot below is also based on the main effects model we discussed in Question 19.

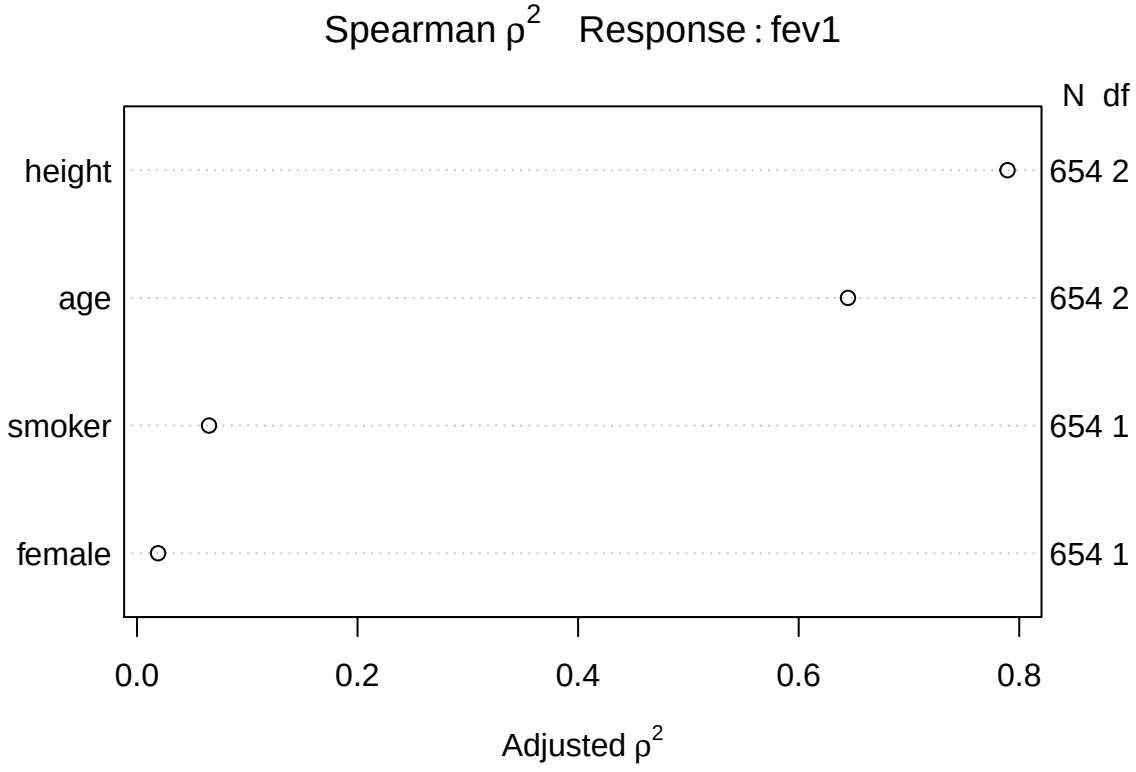



```
n=654   Mean absolute error=0.084   Mean squared error=0.01132
0.9 Quantile of absolute error=0.176
```

Which of the following is directly assessed by the Question 21 plot?

a. This is a check to see if multi-collinearity is an issue.
b. The assumption of linearity.
c. The assumption of homoscedasticity (constant variance).
d. This is a check of whether the model is parsimonious.
e. The assumption of independence.
f. This is a check of whether the model is well-calibrated.
g. The assumption of Normality.
h. This is a check on the influence of the observations on our model.

22 Question 22



In light of the Spearman plot for the `childfev` data shown above, you will answer two questions in this case.

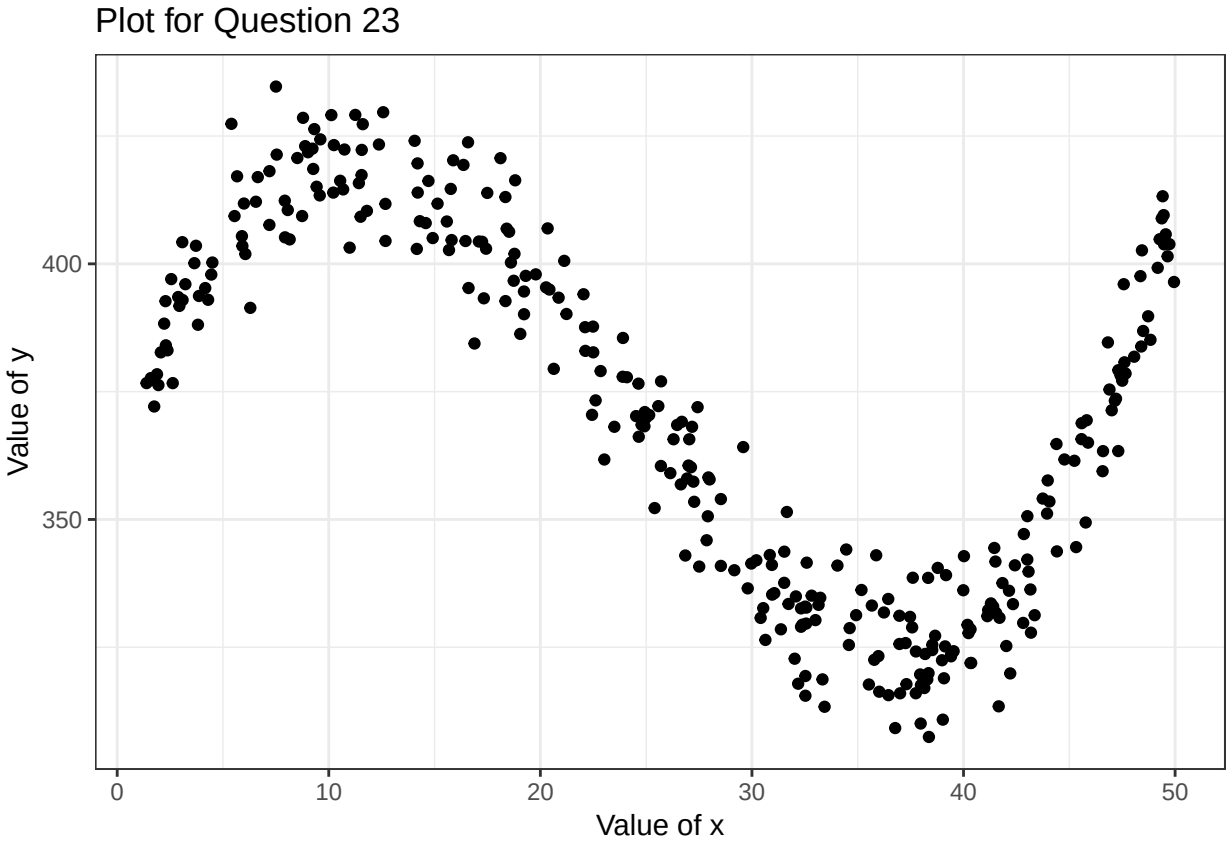
In part 1 you will specify the single most important of the following augmentations to our main-effects model for `fev1` to consider.

In part 2, you will specify the second most important of the following augmentations to our main-effects model for `fev1` to consider.

The augmentations under consideration are:

a. Add a restricted cubic spline in age
b. Add a restricted cubic spline in female
c. Add a restricted cubic spline in smoker
d. Add a restricted cubic spline in height
e. Add an interaction between smoker and age
f. Add an interaction between smoker and female
g. Add an interaction between smoker and height
h. Add an interaction between female and height
i. Add an interaction between female and age

23 Question 23



Consider the plot above. Which of the following approaches to fitting a restricted cubic spline would be most appropriate for modeling the association between y and x?

a. Using 2 knots
b. Using 3 knots
c. Using 4 knots
d. Using 5 knots
e. None of the above

Setup for Questions 24-25

For the last two questions, we use an R data set that has been made available to you called "data24.Rds" on our web site. It contains 509 observations on 7 potential predictors, labeled `a`, `b`, `c`, `d`, `e`, `f` and `g`, of a continuous `outcome` that has been rounded to 1 decimal place.

- Variable `a` can take any value between 0 and 1 (rounded to 3 decimal places)
- Variable `c` takes on integer values between 1 and 10
- Variables `d` and `f` are binary categorical variables
- Variables `b`, `e` and `g` are quantities (rounded to 2 decimal places, each)

```
data24
```

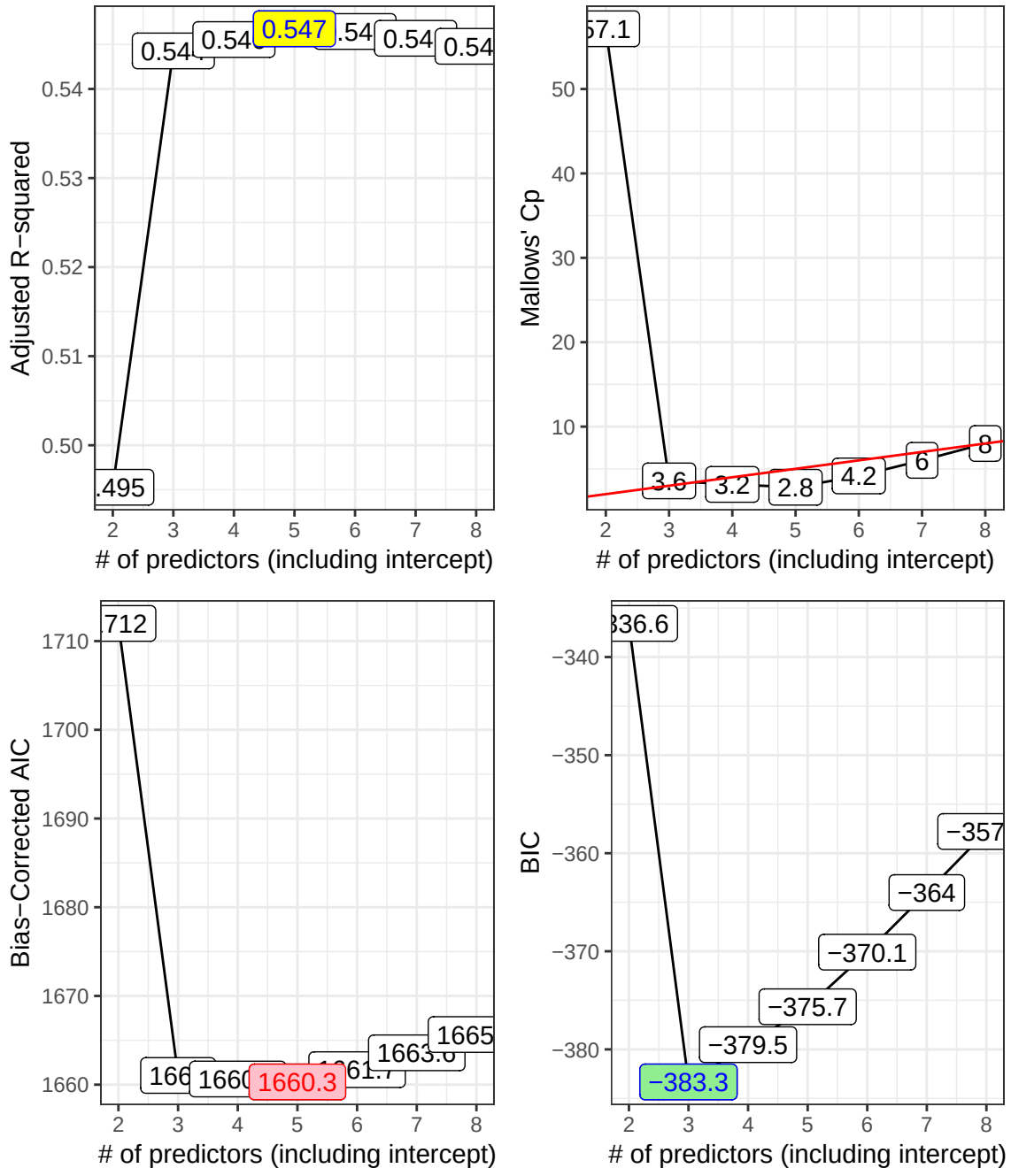
```
# A tibble: 509 x 8
       a     b     c     d     e     f     g outcome
   <dbl> <dbl> <int> <int> <dbl> <int> <dbl>   <dbl>
 1 0.69  122.      7     0 10.4      1 103.     56.2
 2 0.893 148.      2     0  8.71     1  98.0    54.3
 3 0.586 124.      3     0 11.0      0 108.     48.4
 4 0.697 113.      8     0 10.5      1 104.     56.4
 5 0.523  99.7     1     0 10.1      0  99.6    45.5
 6 0.584 105.      1     0 10.9      0 100.     59.2
 7 0.674 118.      3     0  9.54     0  96.8    54.7
 8 0.459  97.5     9     1  9.7      1 105.     69.1
 9 0.456  95.6     7     1 10.6      0 106.     66.2
10 0.455 101.      4     0  9.47     0 104.     46.6
# ... with 499 more rows
```

24 Question 24

Below is some output I obtained using `regsubsets` from the `leaps` package.

```
q24_win %>% print.data.frame
```

```
  (Intercept)     a     b     c    d     e     f     g k        r2
1        TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE 2 0.4962751
2        TRUE FALSE FALSE  TRUE TRUE FALSE FALSE FALSE 3 0.5460701
3        TRUE FALSE FALSE  TRUE TRUE FALSE  TRUE FALSE 4 0.5482366
4        TRUE FALSE FALSE  TRUE TRUE  TRUE  TRUE FALSE 5 0.5503305
5        TRUE  TRUE FALSE  TRUE TRUE  TRUE  TRUE FALSE 6 0.5509195
6        TRUE  TRUE FALSE  TRUE TRUE  TRUE  TRUE  TRUE 7 0.5510572
7        TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE 8 0.5510621
      adjr2        cp    aic.c       bic
1 0.4952816 57.140432 1711.994 -336.5691
2 0.5442759  3.570908 1661.037 -383.3172
3 0.5455528  3.153178 1660.634 -379.5199
4 0.5467617  2.816427 1660.309 -375.6521
5 0.5464555  4.159093 1661.690 -370.0869
6 0.5456914  6.005399 1663.590 -364.0105
7 0.5447895  8.000000 1665.649 -357.7836
```



Based on this output, there are somewhere between two and four models that are recommended for further exploration. Which models are recommended as reasonable candidates here?

For each row in the grid for this question on the Google Form, you will specify the model recommended by that statistic by listing its predictors.

The rows will be: [1] By adjusted R^2 [2] By Mallows' Cp [3] By corrected AIC and [4] By BIC

25 Question 25

Use 10-fold cross-validation to compare the models that you identified as reasonable candidates in Question 24, then specify the predictors in the model that has the best cross-validated results by root mean squared predictive error. Use `set.seed(2019)` in specifying your cross-validation approach.

As a hint, here's how I would code up the process of doing 5-fold cross-validation to assess a model containing predictor `a` only. Of course, that's not quite what you need to do, but we're getting close.

```
set.seed(2019)

fit_model_a <- data24 %>%
    crossv_kfold(k = 5) %>%
    mutate(model = map(train, ~ lm(outcome ~ a,
                                   data = .)))

fit_model_a_pred <- fit_model_a %>%
    unnest(map2(model, test, ~ augment(.x, newdata = .y)))

fit_model_a_results <- fit_model_a_pred %>%
    summarise(Model = "a only",
              RMSE = sqrt(mean((outcome - .fitted) ^2)),
              MAE = mean(abs(outcome - .fitted)))
```

Congratulations! You have reached the end of the Quiz.