431 Quiz 3 and Answer Sketch

Thomas E. Love

due 2017-12-12, version 2017-12-14

0.1 Introduction

You must complete this quiz by **Noon on Tuesday December 12**. The documents you will need to take the Quiz include this PDF document, and the Google Form located at https://goo.gl/forms/yj5YuBusGF7hVvnx1 where you will submit your answers.

Please *select* or type in your best response for each of the 40 questions. The questions are not arranged in order of difficulty, and you should answer all of them. Some questions will require a few minutes of work, others you should be able to answer almost instantly. Some questions (include the last seven questions on the Quiz) require you to code in R to obtain answers, others provide all necessary output in this document. Each question is worth 3 points for a potential total score of 120.

In addressing questions involving R code, you may assume that Love-boost.R and the tidyverse packages have been pre-loaded.

Assume 5% significance in two-sided testing unless otherwise specified.

Data and code relevant to several questions are available to you online. Look for these three files at https://github.com/thomaselove/431data:

- `hospsim.csv`
- `surveyday1_2017.csv`
- `wc_code.R`

You will have the opportunity to edit your responses after completing the quiz, but this must be completed by the deadline. If you wish to complete part of the quiz and then return to it later, please scroll to the end of the quiz and complete the affirmation after the final Question. Then, you will be able to exit the quiz and save your progress.

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 or the Teaching Assistants. They may be reached at `431-help at case dot edu`. Any further announcements about this Quiz will be posted to your preferred email and https://github.com/THOMASELOVE/431slides/tree/master/wrapup.

1 Q01

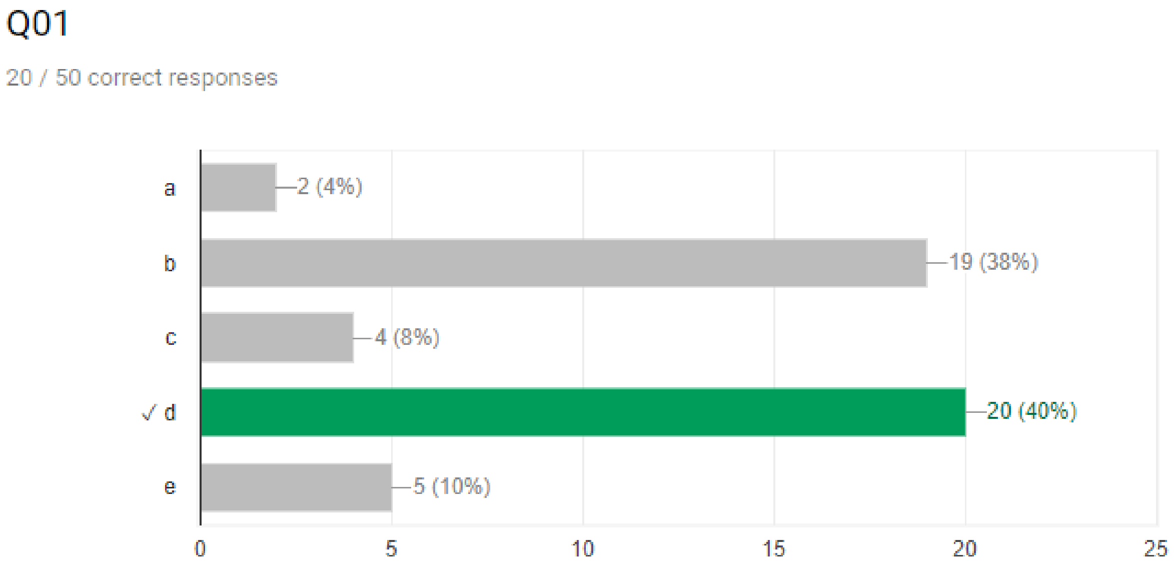
Suppose you have a data frame named `dat` containing a variable called `height`, which shows the participant's height in centimeters. Which of the following lines of code will create a new variable `tall` in the `dat` data frame which takes the value **TRUE** when a subject is more than 175 cm tall,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

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

1.1 A01 is `d`.

- Approach `a` will throw an error message.
- Approach `b` will create a YES/NO, rather than a TRUE/FALSE variable. This was the most common incorrect response.
- Approach `c` will pull the tall people into a data frame called tall.
- Approach `d` does what we're looking for.

1.2 Results


Q01

20 / 50 correct responses

2 Q02

I fit two linear regression models, called `m1` and `m2`, to predict the same outcome (y), where `m2` includes a proper subset of the five predictors included in `m1`. I then ran `anova(m1, m2)`.

```
anova(m1, m2)
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5
Model 2: y ~ x1 + x2 + x4

  Res.Df   RSS Df Sum of Sq     F  Pr(>F)
1     94 42667
2     96 45546 -2   -2879.5 3.172 0.04644
```

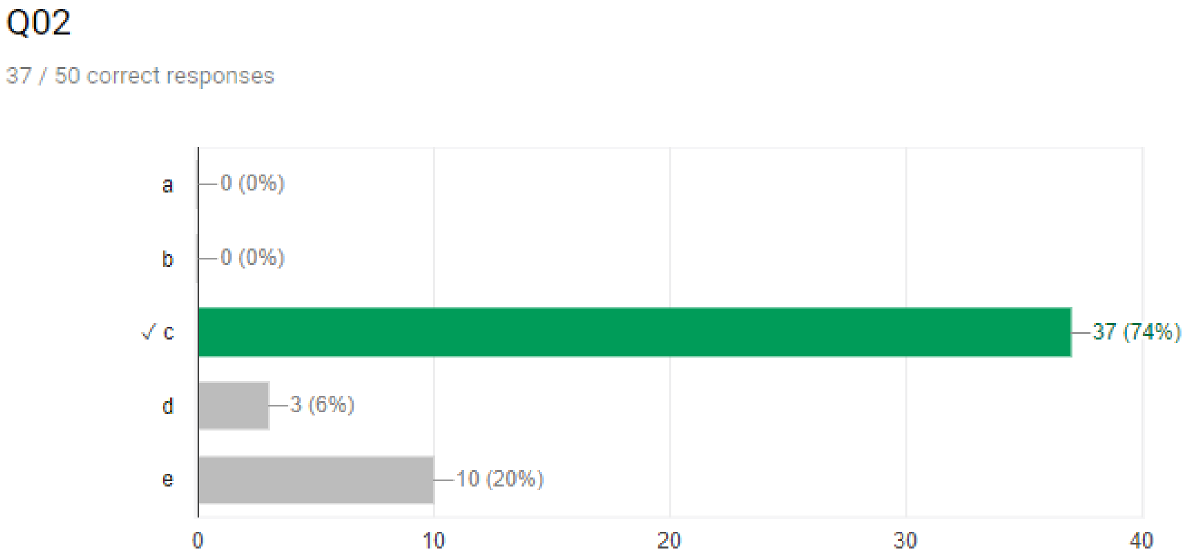
From the *p* value provided in this output, which of the following statements is the most appropriate conclusion at a 5% significance level?

- a. `x1` carries statistically significant predictive value for `y`, assuming that the other four predictors are in the model.
- b. The combination of `x1`, `x2` and `x3` contains statistically significant predictive value for `y`, assuming that none of the other predictors are in the model.
- c. The combination of `x3` and `x5` contains statistically significant predictive value for `y`, assuming that `x1`, `x2` and `x4` are already in the model.
- d. The combination of `x3` and `x5` contains statistically significant predictive value for `y`, assuming that `x1`, `x2` and `x4` are NOT already in the model.
- e. None of these statements are appropriate.

2.1 A02 is `c`.

- Model 1 contains x1, x2, x3, x4, and x5. Model 2 contains x1, x2 and x4.
- The ANOVA F test comparing Model 1 to Model 2 has a *p* value of 0.046.
  - This indicates that there is statistically significant predictive value for the terms included in Model 1 but not Model 2 (x3 and x5), assuming that the terms in Model 2 (x1, x2 and x4) are already included in the model.
  - That's what choice `c` says.
- Choice `a` is incorrect because the p value compares x1, x2, x4 to the model with all five predictors, so it doesn't tell us about x1 by itself.
- Similarly, choice `b` is incorrect because the p value in the output doesn't look at the model with x1, x2 and x3 alone.
- And choice `d` misinterprets what the ANOVA comparison of models does.

2.2 Results



Q02

37 / 50 correct responses

3 Q03

Using the `surveyday1_2017.csv` data file from https://github.com/thomaselove/431data, I developed the following table of information describing the association of these two items:

- `sex`: What is your sex? (f = Female, m = Male)
- `smoke`: Do you smoke? (1 = Non-Smoker, 2 = Former Smoker, 3 = Current Smoker)

–	`smoke` = 1	`smoke` = 2	`smoke` = 3	Total
Female	91	5	0	**96**
Male	94	9	3	**106**
Total	**185**	**14**	**3**	**202**

The resulting Pearson chi-squared test p-value is 0.157. Note that you could check this yourself using the data, but there is no need to do so.

Which of the following statements is appropriate?

- a. The assumptions required by the usual chi-squared test are reasonable here.
- b. The p value for a chi-square test is below our usual level to indicate significance.
- c. `sex` and `smoke` are independent of each other.
- d. The presented p value means that the `smoke` score is statistically significantly higher in females.
- e. None of these statements are appropriate.

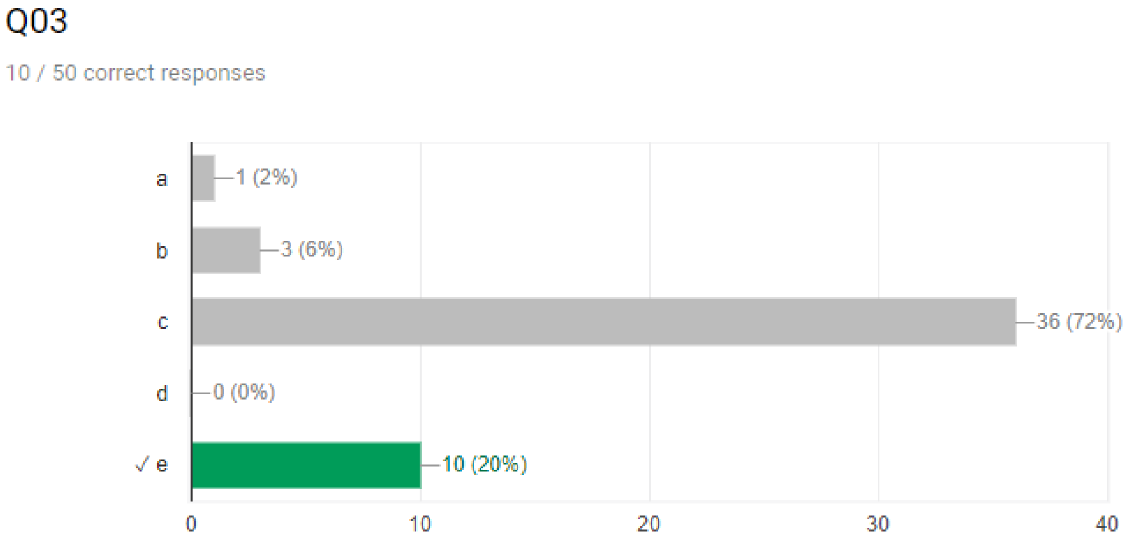
3.1 A03 is `e`.

This was the first "difficult" question. None of the provided statements are appropriate.

- The expected frequencies in the `smoke` = 3 cells are well less than five, so our Pearson chi-square test approximation is a problem, so `a` is incorrect. If you try running the chi-square test, you'll actually get warned about this.
- The p value for the chi-square test (0.157) is actually well above our usual signifance level of 0.05, so `b` is incorrect.
- There's clearly *some* association in the sample - the probabilities aren't identical for males and females in each category. The chi-square test allows us to retain the null hypothesis of independence. But retaining the null doesn't mean that they were truly independent of each other, just that the dependence was too weak for us to detect it in our sample, so `c` is incorrect. You cannot prove a null hypothesis.
- The mean `smoke` score is a fairly silly idea here, but for what it's worth, the p value listed doesn't tell us anything about means. Plus, the sample mean for females (101/96

= 1.05) is actually less than that of the males (121/106 = 1.14) so the direction is wrong, at best. So d is incorrect, too.

3.2 Results



Q03

10 / 50 correct responses

Clearly, most students fell into the trap set by choice c.

4 Q04

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?

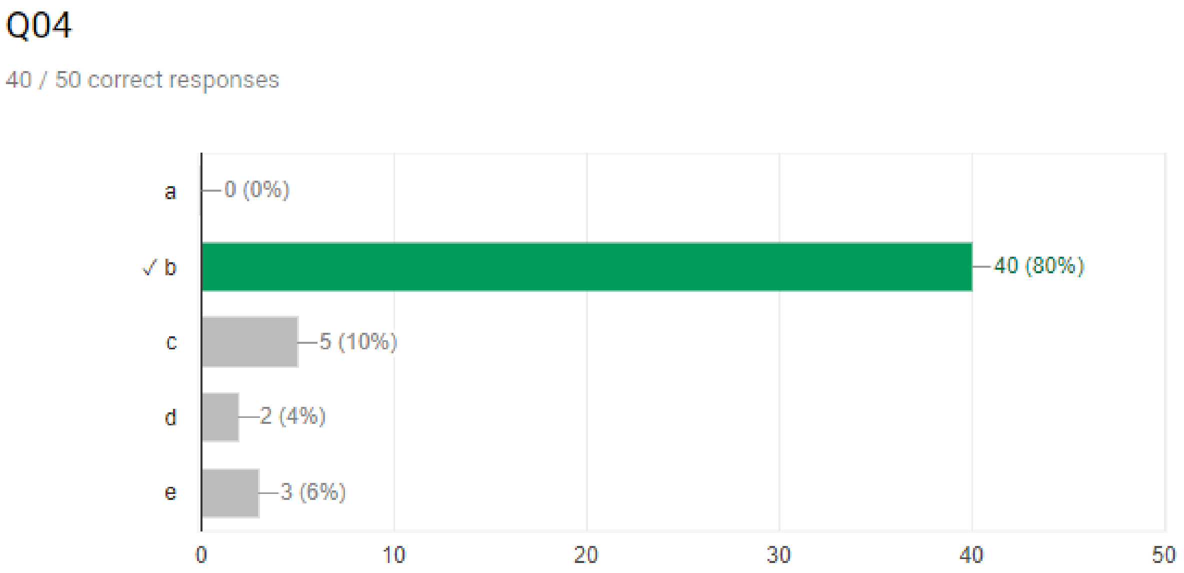
- a. She should use a significance level of 0.20 on each test.
- b. She should use 0.005 on each test.
- c. She should use 0.0125 on each test.
- d. She should use 0.05 on each test.
- e. None of these answers are correct.

4.1 A04 is `b`.

In total, there are **ten** pairwise comparisons to be made at Watchmaker's Technical Institute: A[my] vs. B[eth], A vs. C[armen], A vs. D[onna], A vs. E[lena], B vs. C, B vs. D, B vs. E, C vs. D, C vs. E and D vs. E.

- So if we want to retain a 5% significance level with a Bonferroni correction, we'd have to run the two-sample t tests at a significance level 1/10 that size, or 0.005.
- Only that rate (contained in answer `b`) will ensure that our overall error rate across all 10 comparisons will be no more than 0.05.
- So you'll need to make sure that's the approach taken the administrator at Tick-Tock-Tech.

4.2 Results



Q04

40 / 50 correct responses

5 Q05

If the administrator at the Watchmaker's Technical Institute that we mentioned in Q04 instead used a Tukey HSD approach to make her comparisons, she might have obtained the following output.

```
Tukey multiple comparisons of exam2 means, 95% family-wise confidence level

                diff     lwr   upr    ||                  diff     lwr   upr
               -----   ----- -----    ||                 -----  ------  ----
Beth-Amy       -3.09   -9.43  3.26    ||  Donna-Beth     -3.74  -10.08  2.61
Carmen-Amy     -7.03  -13.37 -0.69    ||  Elena-Beth     -2.53   -8.87  3.81
Donna-Amy      -6.82  -13.17 -0.48    ||  Donna-Carmen    0.21   -6.14  6.55
Elena-Amy      -5.62  -11.96  0.73    ||  Elena-Carmen    1.41   -4.93  7.76
Carmen-Beth    -3.94  -10.28  2.40    ||  Elena-Donna     1.21   -5.14  7.55
```

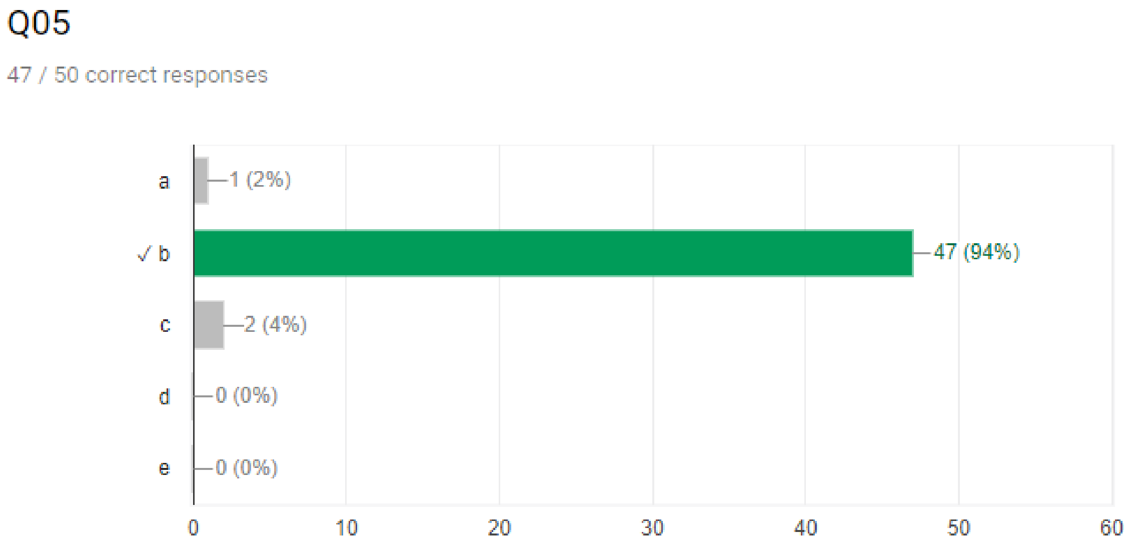
Note that when we refer in the responses below to Beth's scores, we mean the scores of students who were in Beth's lab section. Which conclusion of those presented below would be most appropriate?

- a. Beth's scores were significantly lower than Amy's.
- b. Amy's scores are significantly higher than Carmen or Donna.
- c. Amy's scores are significantly lower than Carmen or Donna.
- d. Elena's scores are significantly higher than Beth, Carmen or Donna.
- e. None of these answers are correct.

5.1 A05 is b.

There are ten comparisons being made, but only two are statistically significant, in that the Tukey HSD interval does not include a mean difference of zero. Those are the comparisons of Carmen to Amy and Donna to Amy. In each case, Amy has higher scores, since the C-A and D-A differences are negative.

5.2 Results

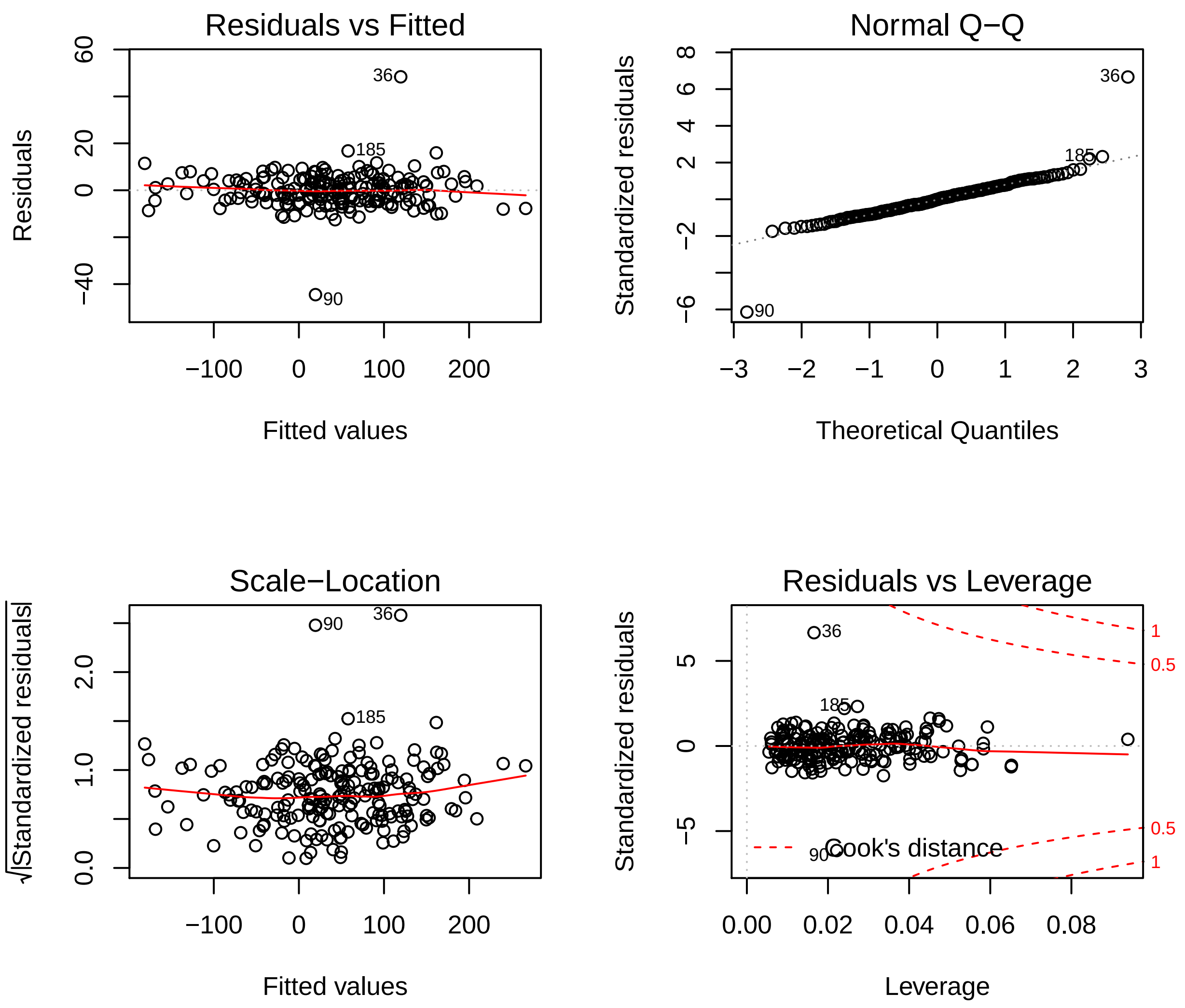


Q05

47 / 50 correct responses

6 Q06

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?

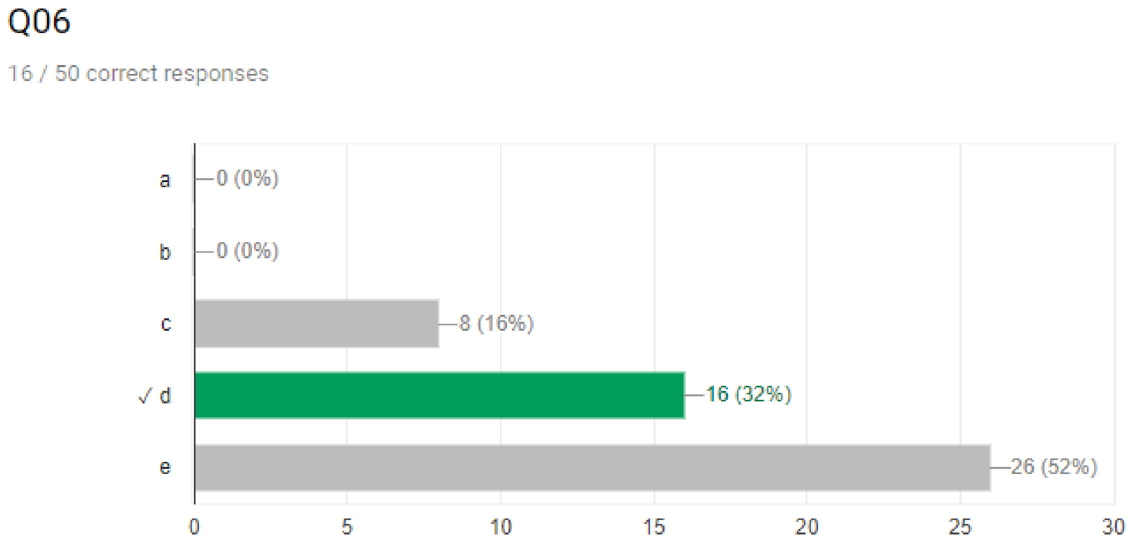
- a. Our main problem is with collinearity.
- b. Our main problem is with the assumption of linearity.
- c. Our main problem is with the assumption of constant variance.
- d. Our main problem is with the assumption of normality.
- e. We have no apparent problems with regression assumptions.

6.1 A06 is `d`.

We have two especially poorly fitted points in the data, and these appear to be rows 36 and 90. These two outliers don't have especially large leverage or influence, but they are very poorly fit, with standardized residuals around 6-7 (one positive, one negative). This is primarily a problem with the assumption of Normality. If those two points were removed, there's no suggestion in any of the other plots of a problem.

More than half of the people taking the quiz chose `e`, and I want to be clear - residuals of this size are definitely a violation of the Normality assumption. It is certainly true that in this case, the solution appears to be straightforward: removing these two outliers and refitting the model probably won't change our conclusions appreciably, based on the current leverage and influence profile.

6.2 Results

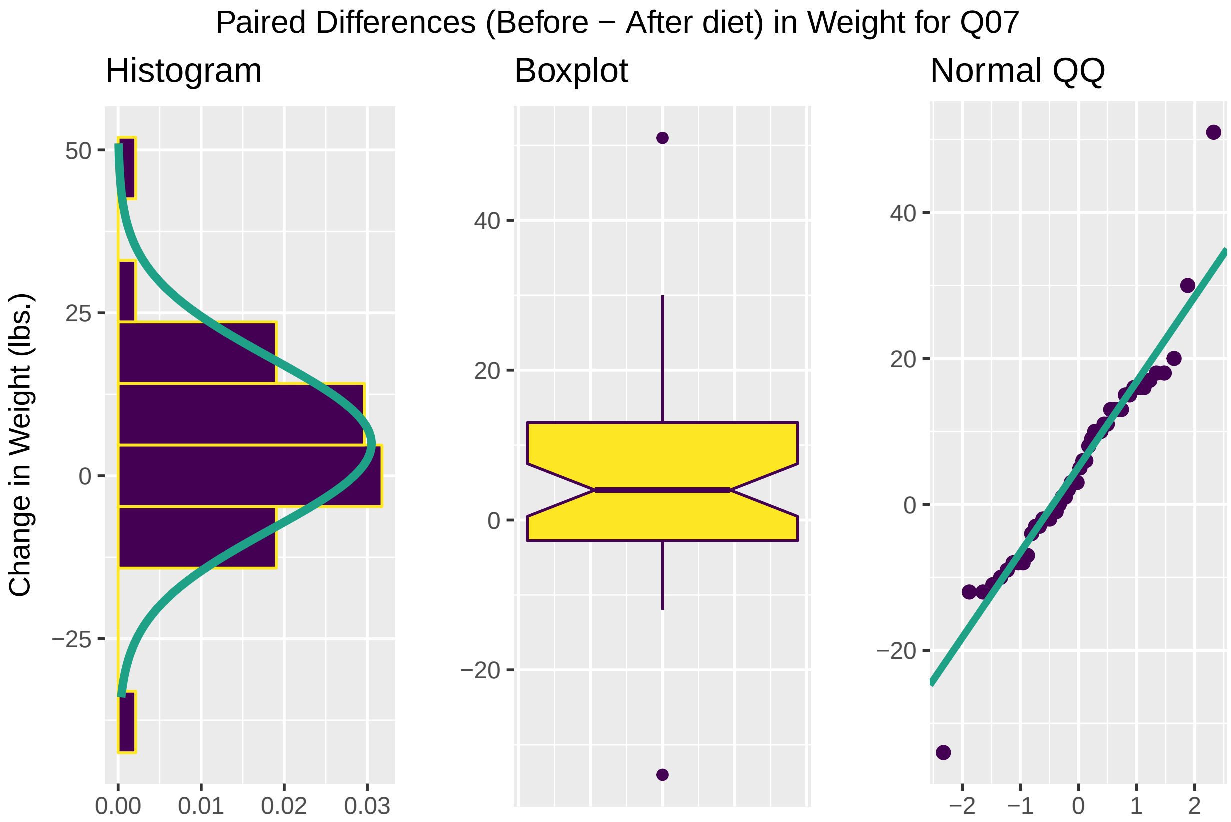


Q06

16 / 50 correct responses

7 Q07

Suppose we compare the change in weight (before - after, in pounds) for 50 overweight male adult subjects who enter into a rather strict nutritional regimen. Specifically, the subjects drink nothing other than water, and eat nothing but a variety of potatoes for two weeks, then spend four weeks eating only high-nutrition vegetables, and still drinking nothing but water. The team's statistical analyst prepares the following output.



```
One Sample t-test

data:  dat07$diff
t = 2.6613, df = 49, p-value = 0.0105
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval: -0.035  9.875
sample estimates: mean of x = 4.92


Result of applying Hmisc::smean.cl.boot
with a 99% confidence level

  Mean  Lower  Upper
4.9200 0.5191 9.7413
```

If the researchers are committed to the use of a 99% confidence level, then which of the following conclusions is most appropriate, given the output above?

- a. These should be treated as independent samples, and the t test is appropriate, so we find no significant difference with 99% confidence.
- b. These should be treated as independent samples, and the bootstrap is appropriate, so we find no significant difference with 99% confidence.
- c. These should be treated as paired samples, the boxplot suggests that we use the t-based interval and so we conclude that the mean weight loss was significantly more than zero.
- d. These should be treated as paired samples, the boxplot suggests that we use the t-based interval and so we conclude that the mean weight loss was not significantly more than zero.
- e. These should be treated as paired samples, the boxplot suggests that we use the bootstrap and so we conclude that the mean weight loss was significantly more than zero.

7.1 A07 is `e`.

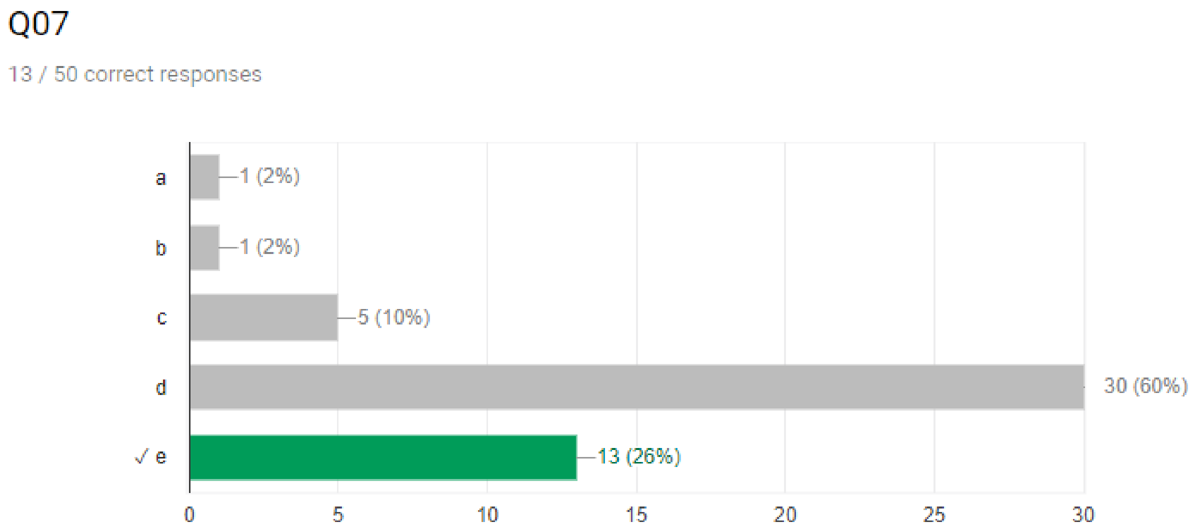
These are before-after differences, where each subject is measured twice. Clearly these described paired samples, so `a` and `b` are wrong.

The output shows two very substantial outliers, so a t test is less appropriate than a bootstrap would be (and thus `c` and `d` are wrong.) Note that a mean is still likely to be of interest, especially since there's no meaningful skew.

So the bootstrap estimate for the mean looks like the right choice, and with 99% confidence, we have an uncertainty interval of (0.52, 9.74) pounds of weight loss. So the mean weight loss was in fact significantly larger than 0, even at the $\alpha = 0.01$ level.

More than half of the people taking the quiz selected choice `d` here, which definitely surprised me. The two outliers are clearly a sign that a t test's assumption of Normality is problematic, and that a bootstrap is preferable here.

7.2 Results

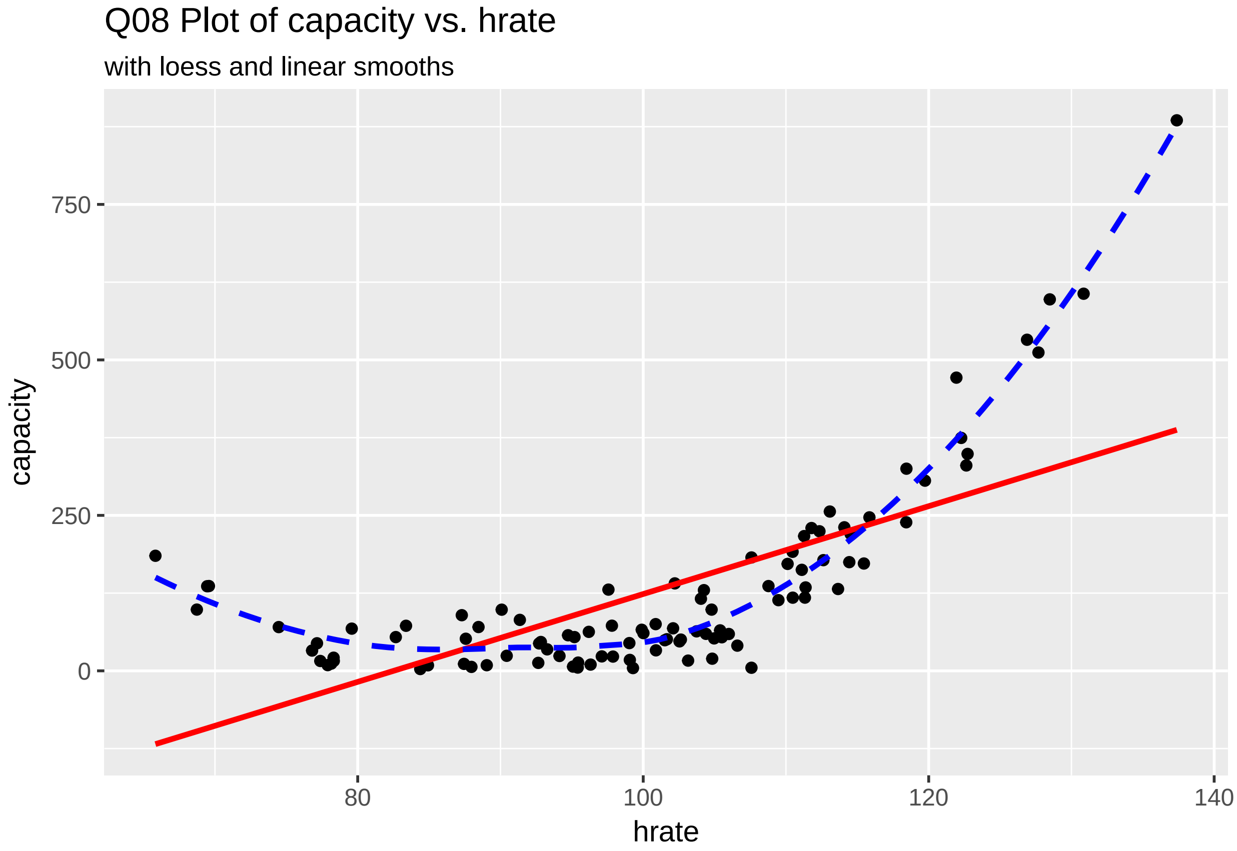


Q07

13 / 50 correct responses

8 Q08

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 below. If you have trouble distinguishing colors, I'll say that one smooth is shown with blue dashes and the other is in red without dashes, but just solid color.



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

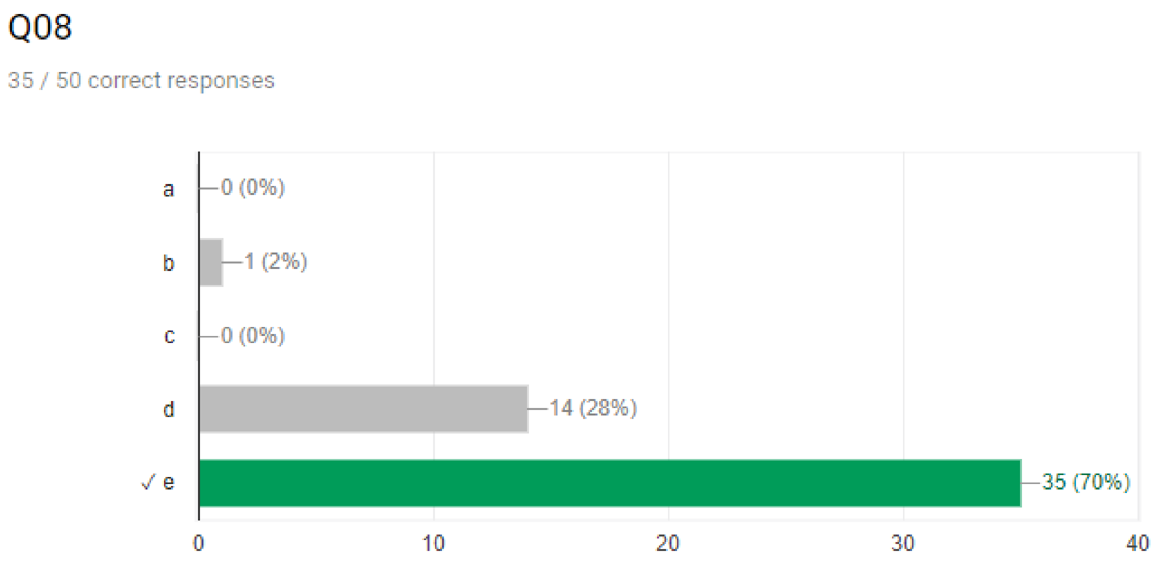
8.1 A08 is e.

e is correct, because the other four statements are false.

- a is false. The linear fit is clearly shown in red, not blue, and is a solid, not dashed line.
- b is false. The loess smooth is a much better fit to this curved association.
- c is false. The Pearson correlation will have the same sign as the slope of the regression line, which is positive.
- d is false. The data are taken from a cross-section of subjects. There's no time ordering here, so there's no way to have a problem with the independence assumption in a regression model.

14 people selected option d, which isn't right. I suspect they were mostly reluctant to select "None of the above" or they were unclear about the implications of the study being cross-sectional.

8.2 Results



Q08

35 / 50 correct responses

9 Q09

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.

```
m09 <- lm(log(sbp) ~ age + log(weight) + height + smoke, data = wcgs09)
arm::display(m09, digits = 3)
```

```
lm(formula = log(sbp) ~ age + log(weight) + height + smoke, data = wcgs09)
            coef.est coef.se
(Intercept)  4.528    0.414 
age          0.008    0.002 
log(weight)  0.163    0.084 
height      -0.012    0.004 
smokeYes    -0.023    0.021 
---
n = 150, k = 5
residual sd = 0.122, R-Squared = 0.17
```

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

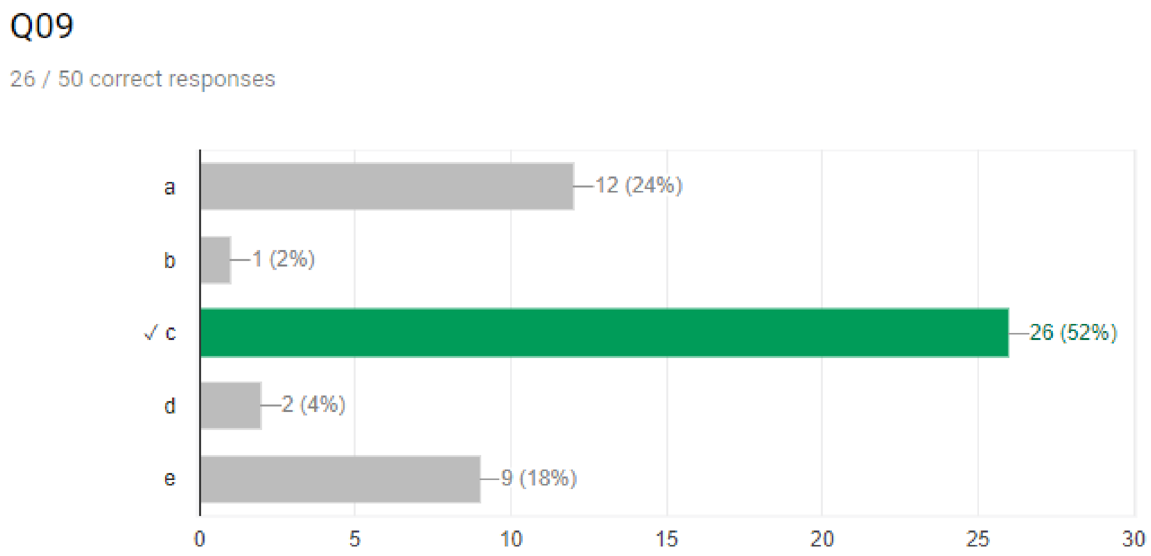
- a. Smokers have significantly lower systolic blood pressures than non-smokers, after we account for age and size (height and weight).
- b. Older people have lower blood pressures on average than do younger people.
- c. Larger height is associated with significantly lower blood pressure, even after we've accounted for age, weight and smoking status.
- d. This model accounts for about 12% of the variation in the log of systolic blood pressure.
- e. None of these statements are true.

9.1 A09 is `c`.

- `a` is false, because the estimated coefficient for smokeYes is -0.023, barely larger than its associated standard error (0.021), so a 95% uncertainty interval (estimate ± 2 se) will certainly include 0.
- `b` is false because the model suggests the opposite, if we are talking about significance and adjusting for these other predictors, and if we're not making adjustments, then we have no useful information from this output.

- `c` is true. `height` is negatively associated with log(`sbp`) and the approximate 95% confidence interval for the slope of height is (-0.020, -0.004), so additional height is associated with lower BP after accounting for the other variables in the model.
- `d` is false, because the R^2 here is 0.17, not 0.12.

9.2 Results

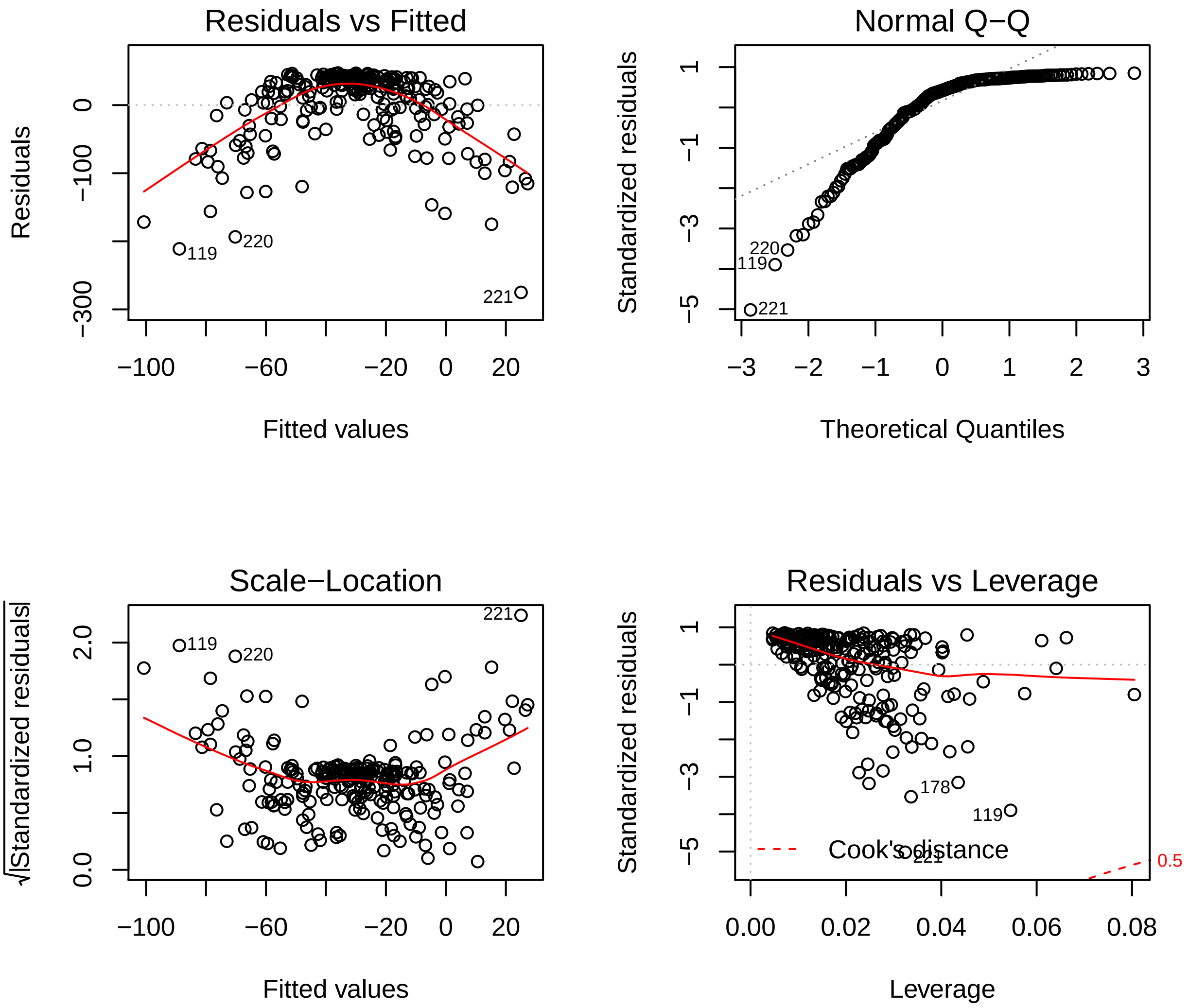


Q09

26 / 50 correct responses

10 Q10

A regression model was developed to predict an outcome, y, based on a linear model using four predictors in a sample of 240 subjects. The following residual plots emerged.



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

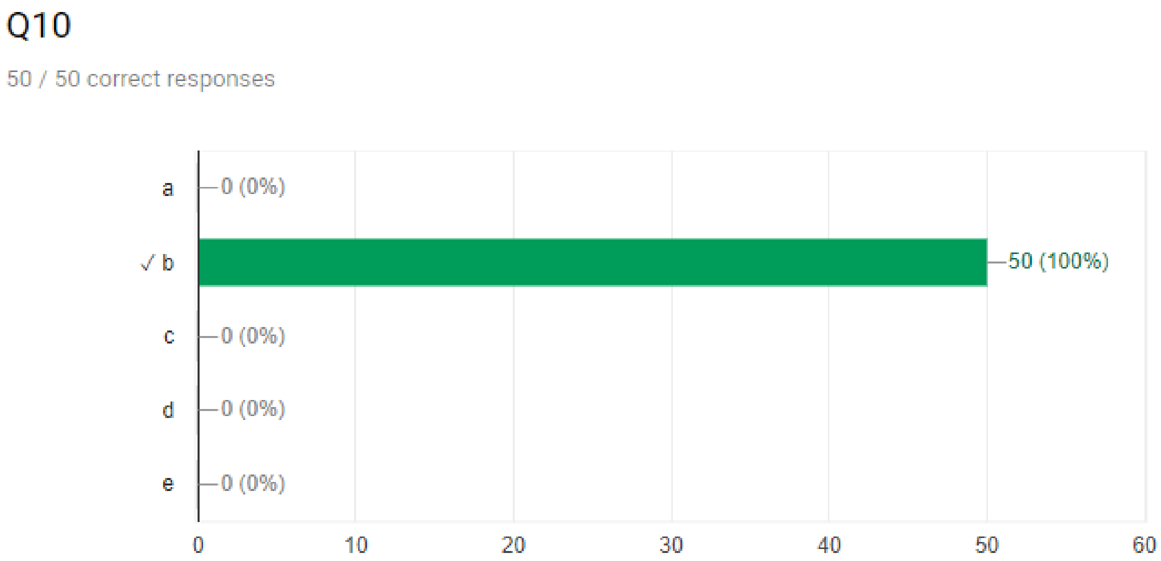
- a. The first thing we should try is dropping a highly influential point.
- b. The first thing we should try is transforming y to improve linearity.
- c. The first thing we should worry about is the assumption of constant variance.
- d. The first thing we should try is to focus on the problem of collinearity.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

10.1 A10 is `b`.

The essential problem here is indicated by the curve in the residuals vs. fitted values plot on the top left. That's a problem with the linearity assumption, and our next step should be to consider transformations of the outcome (and if that doesn't work, perhaps the predictors). That's choice `b`.

I guess I don't actually need to explain this, since everyone got it right. Congratulations!

10.2 Results



Q10

50 / 50 correct responses

11 Q11

(Note that this background information will be used in Q11 - Q15.) Suppose you fit three candidate 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 following results.

```r
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)

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

```
       df      AIC
modelA 10 4380.234
modelB  7 4375.219
modelC  5 5088.652
modelD  3 5329.352
```

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

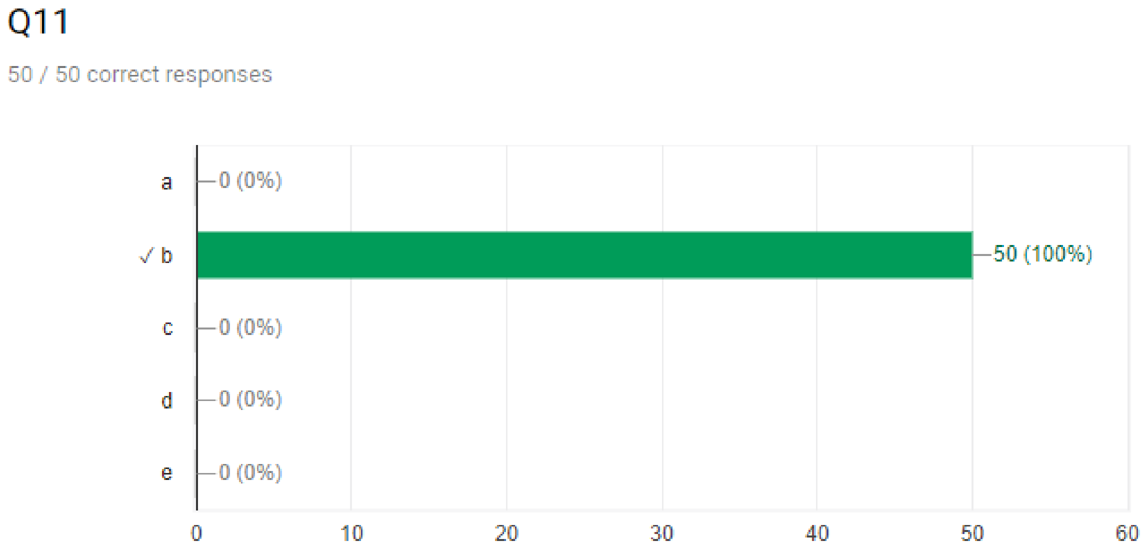
- a. Model A
- b. Model B
- c. Model C
- d. Model D
- e. The output doesn't suggest a "best" choice.

11.1 A11 is `b`.

The model with the smallest AIC is model B. That's our winner, according to this output.

Again, perhaps I don't actually need to explain this, since everyone got it right. Congratulations!

11.2 Results



Q11

50 / 50 correct responses

12 Q12

The following output relates to `modelA` described in Q11.

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

```
   x1    x2    x3    x4    x5    x6    x7    x8 
1.012 1.014 1.107 1.020 1.015 2.610 2.491 1.015 
```

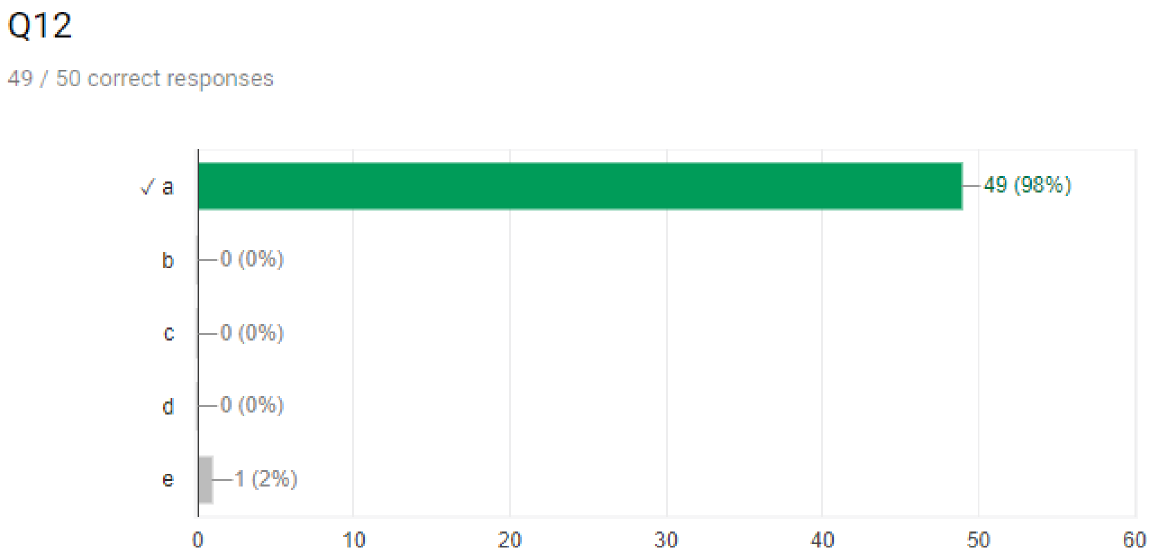
Which of the following statements is the best conclusion from this output?

- a. Model A has no sign of meaningful collinearity.
- b. Model A has a serious problem with collinearity.
- c. Model A's residuals will show no problem with independence.
- d. Model A's residuals will show a serious problem with independence.
- 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.

12.1 A12 is a.

The `vif`, or variance inflation factor, helps us measure collinearity. Since none of the VIF values exceed (or even approach) 5, this indicates that there is no serious collinearity in Model A. The other statements are simply incorrect, and, in the case of `c`, `d`, and `e`, actually unrelated to the VIF.

12.2 Results



Q12

49 / 50 correct responses

13 Q13

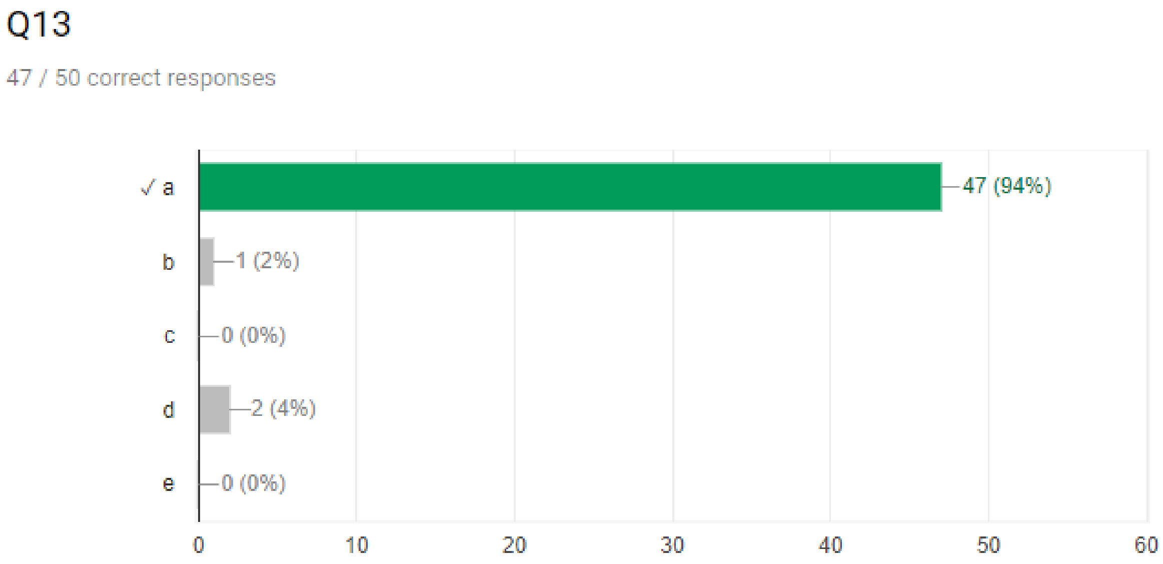
Which of the following R commands would calculate fitted values of `log(predatory.behavior)` using the equation in Model A (from Q11 and Q12), for a new set of data contained in the `newleopard` tibble?

- a. `predict(modelA, newdata = newleopard)`
- b. `glance(modelA, newdata = newleopard)`
- c. `tidy(modelA, newdata = newleopard)`
- d. `augment(modelA)`
- e. `split(modelA, newdata = newleopard)`

13.1 A13 is a.

The `predict` function displayed in choice `a` does this job. The `augment` function from `broom` does this, but within the data set where we fit the model. For out of sample prediction as is called for here, `predict` is the winner.

13.2 Results



Q13

47 / 50 correct responses

14 Q14

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?

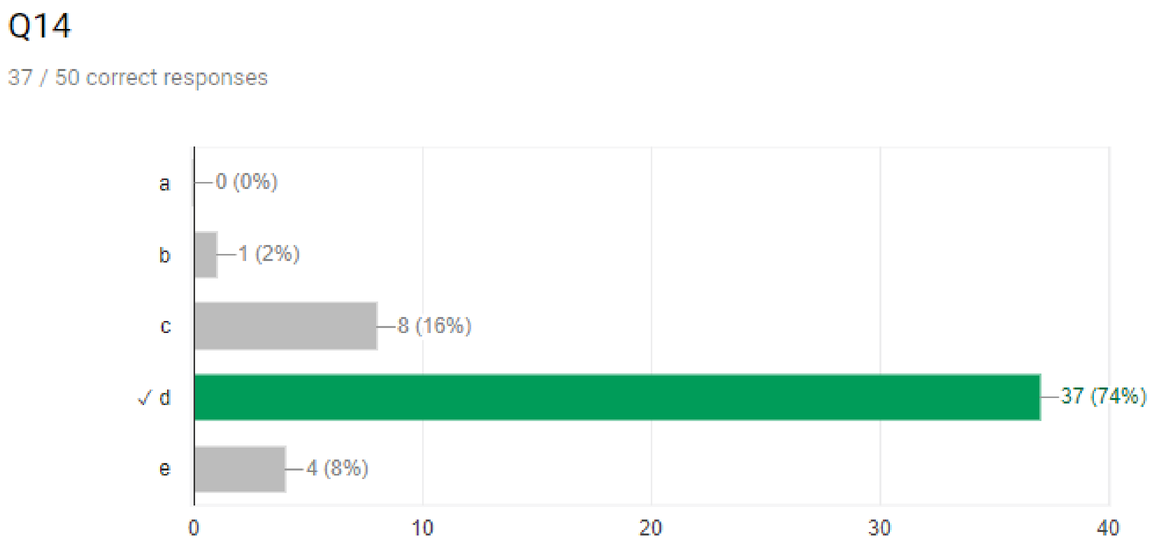
- a. 3 to 4
- b. `log(3)` to `log(4)`
- c. `10*3` to `10*4`
- d. `exp(3)` to `exp(4)`
- e. None of these methods would work.

14.1 A14 is `d`.

To back out of the natural logarithm (`log`) and thus untransform back to our original scale, we'd use the `exp` function. `d` is correct.

For those of you who chose `c`, remember that `log10` in R gives the base-10 logarithm, and not `log`.

14.2 Results

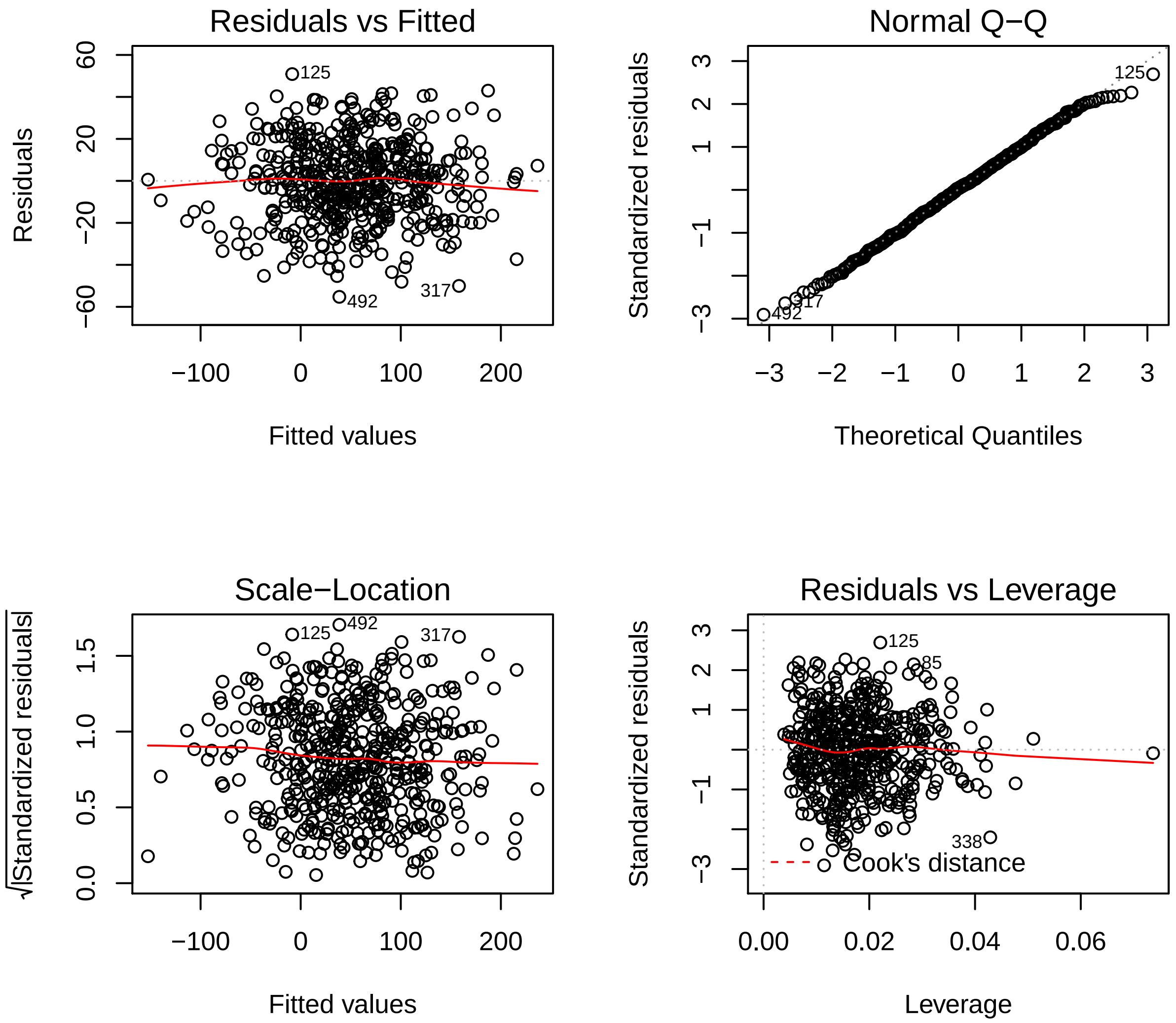


Q14

37 / 50 correct responses

15 Q15

Behold the residual plots for Model A from Q11.



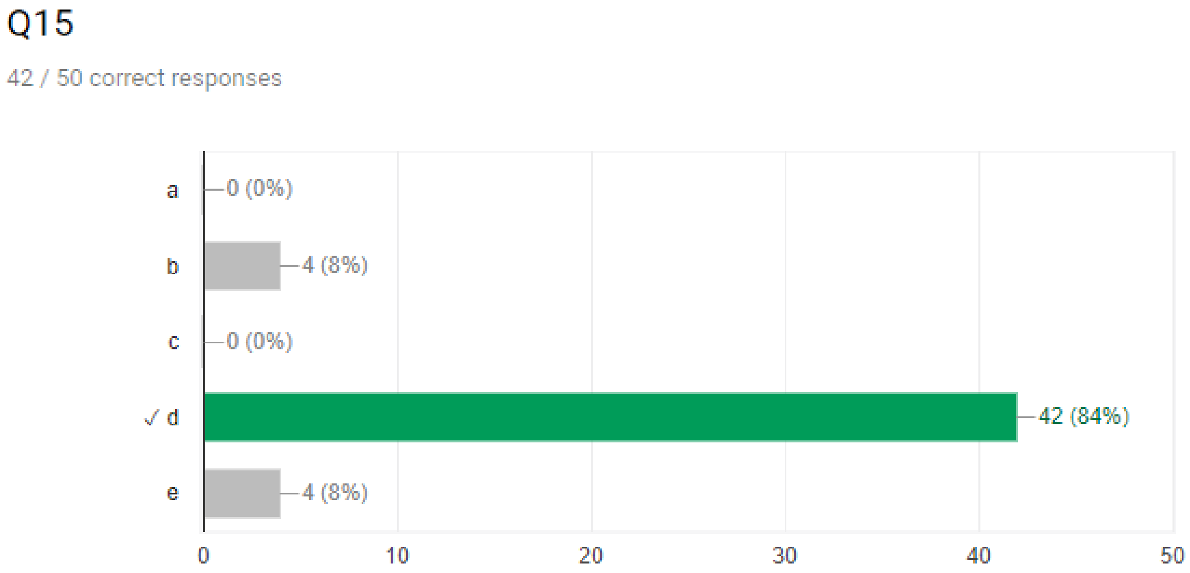
Which of the following conclusions is most appropriate?

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

15.1 A15 is d.

I see no serious curve in the plot of residuals vs. fitted values, so there's no clear problem with *linearity*. I see no serious fan shape in the residuals vs. fitted values, and no clear rise or fall in the scale-location plot, so there's no clear problem with the assumption of constant variance. The Normal Q-Q plot shows no sign of substantial problems with the assumption of Normality. Nor are there any clearly influential points. I see no problems. Choice d is correct.

15.2 Results



Q15

42 / 50 correct responses

16 Q16

A series of 90 models were built by a team of researchers interested in systems biology. 37 of the models showed promising results in an attempt to validate them out of sample. Define the hit rate as the percentage of models built that show these promising results. Which of the following intervals appropriately describes the uncertainty we have around a hit rate estimate in this setting, using a SAIFS approach and permitting a 10% rate of Type I error?

- a. (30.4%, 52.1%)
- b. 0.411 plus or minus 4.1 percentage points.
- c. (32.1%, 50.4%)
- d. (27.0%, 55.4%)
- e. None of these intervals.

16.1 A16 is `c`.

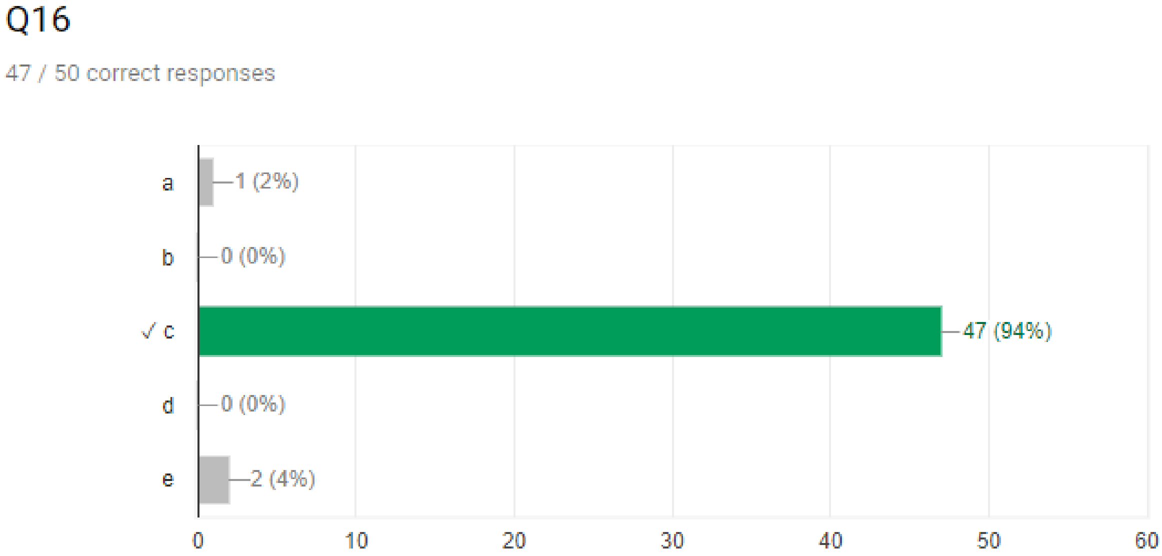
This is just a fancy way of asking for a 90% SAIFS confidence interval, with x = 37 and n = 90.

```
saifs.ci(x = 37, n = 90, conf.level = 0.90)
```

```
Sample Proportion                 0.05                 0.95
            0.411                0.321                0.504
```

So the answer (in terms of a percentage rather than a proportion) is (32.1%, 50.4%), or choice `c`.

16.2 Results



Q16

47 / 50 correct responses

17 Q17

The `Pottery` data are part of the `car` package in R. Included are data describing the chemical composition of ancient pottery found at four sites in Great Britain. This data set will also be used in Q18. Here, we will focus here on the Na (Sodium) levels, and our goal is to compare the mean Na levels across the four sites.

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

```
Analysis of Variance Table

Response: Na
          Df  Sum Sq  Mean Sq F value    Pr(>F)    
Site       3 0.25825 0.086082  9.5026 0.0003209 ***
Residuals 22 0.19929 0.009059                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

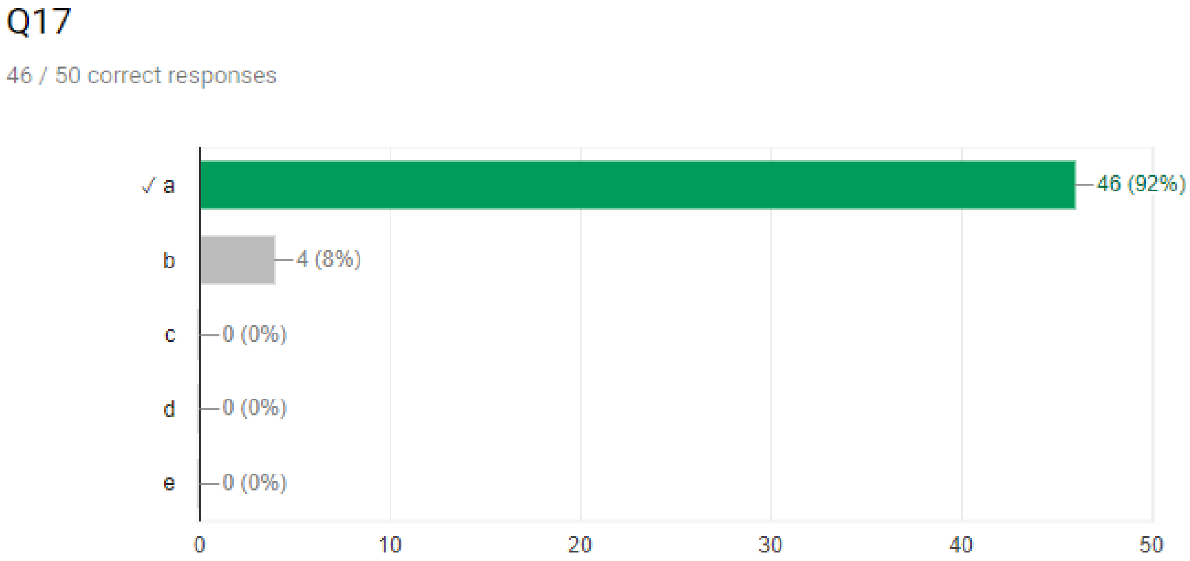
Which of the following conclusions is most appropriate, based on the output above?

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

17.1 A17 is a.

The ANOVA F test gives us a p value well below 0.01, so we can certainly conclude that the population mean Na level in at least one of the four sites is different than the others, at a 1% significance level. That is, in fact, what `a` says. The other statements (`b`, `c` and `d`) show various mistaken versions of what the ANOVA F test might do.

17.2 Results

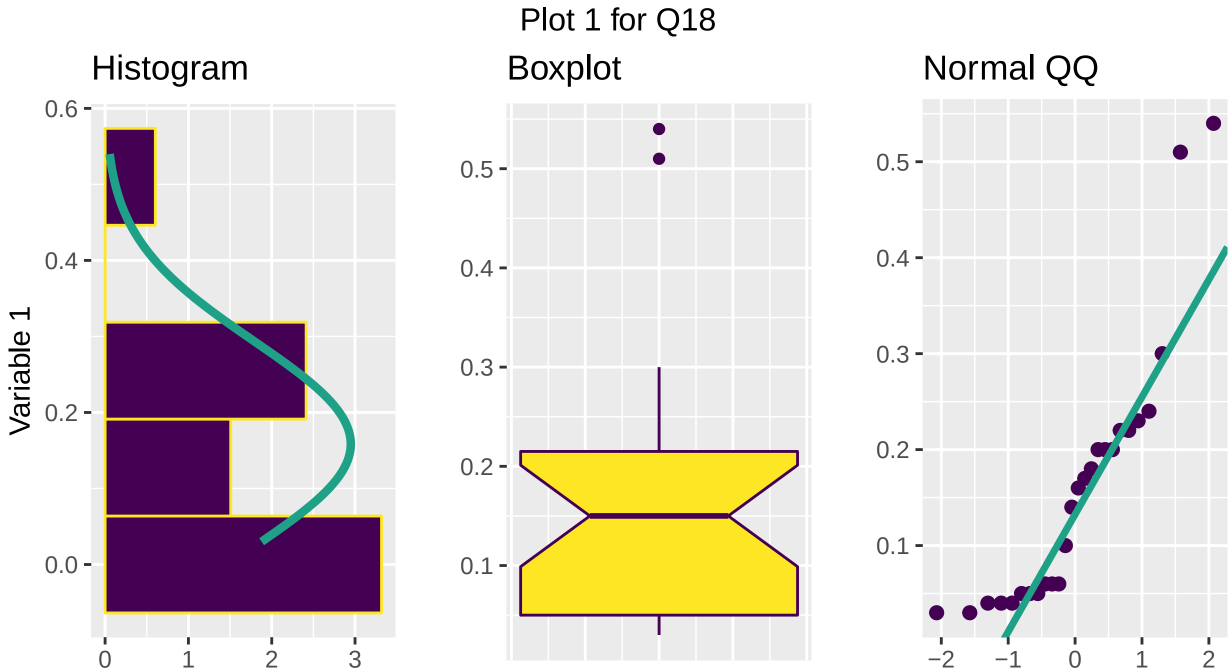


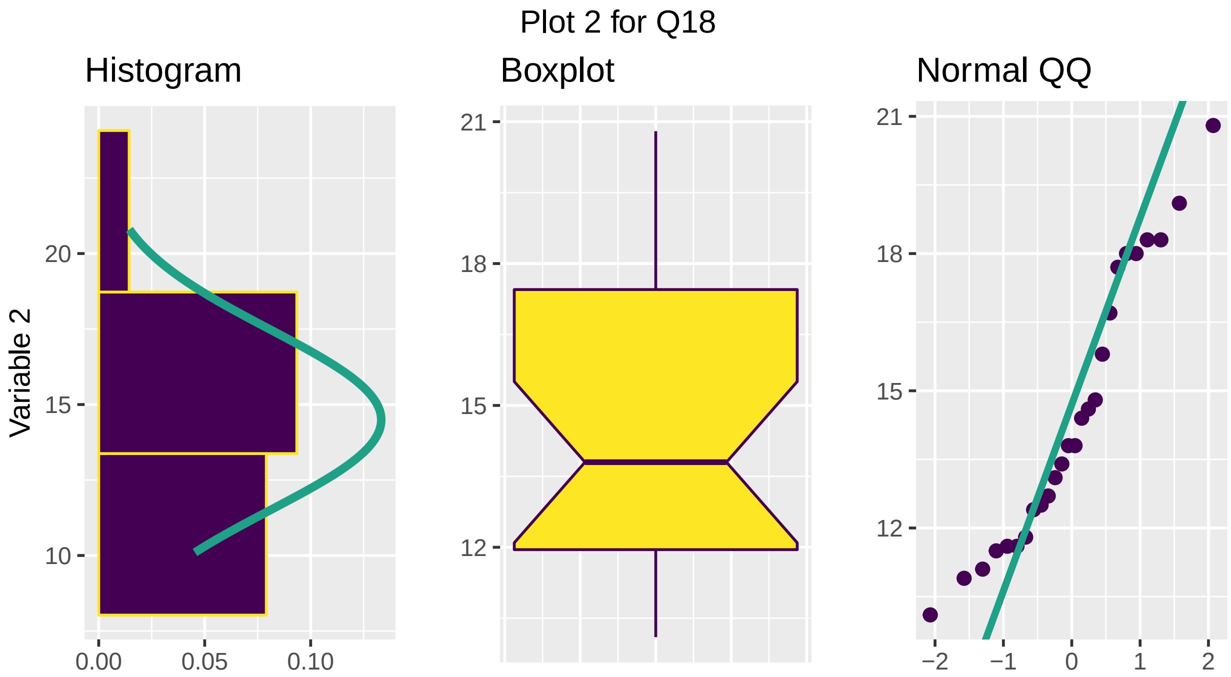
Q17

46 / 50 correct responses

18 Q18

Consider these two sets of plots, generated by the `eda.1sam` function to describe variables from the Pottery data set within the `car` package.





And now, here are descriptive summary statistics from the `summary` function describing the variables contained in the Pottery data set.

```
          Site          Al              Fe              Mg       
 AshleyRails: 5   Min.   :10.10   Min.   :0.920   Min.   :0.530  
 Caldicot   : 2   1st Qu.:11.95   1st Qu.:1.700   1st Qu.:0.670  
 IsleThorns : 5   Median :13.80   Median :5.465   Median :3.825  
 Llanedyrn  :14   Mean   :14.49   Mean   :4.468   Mean   :3.142  
                  3rd Qu.:17.45   3rd Qu.:6.590   3rd Qu.:4.503  
                  Max.   :20.80   Max.   :7.090   Max.   :7.230  
       Ca               Na        
 Min.   :0.0100   Min.   :0.0300  
 1st Qu.:0.0600   1st Qu.:0.0500  
 Median :0.1550   Median :0.1500  
 Mean   :0.1465   Mean   :0.1585  
 3rd Qu.:0.2150   3rd Qu.:0.2150  
 Max.   :0.3100   Max.   :0.5400  
```

Based on this output, and whatever other work you need to do, which of the statements below is true, about Variable 1 (as shown in Plot 1) and Variable 2 (shown in Plot 2)?

- a. Variable 1 is Sodium (Na), Variable 2 in Plot 2 is Aluminum (Al).
- b. Variable 1 is Calcium (Ca), Variable 2 is Aluminum (Al).
- c. Variable 1 is Iron (Fe), Variable 2 is Magnesium (Mg).
- d. Variable 1 is Iron (Fe), Variable 2 is Calcium (Ca).
- e. Variable 1 is Sodium (Na), Variable 2 is Magnesium (Mg).

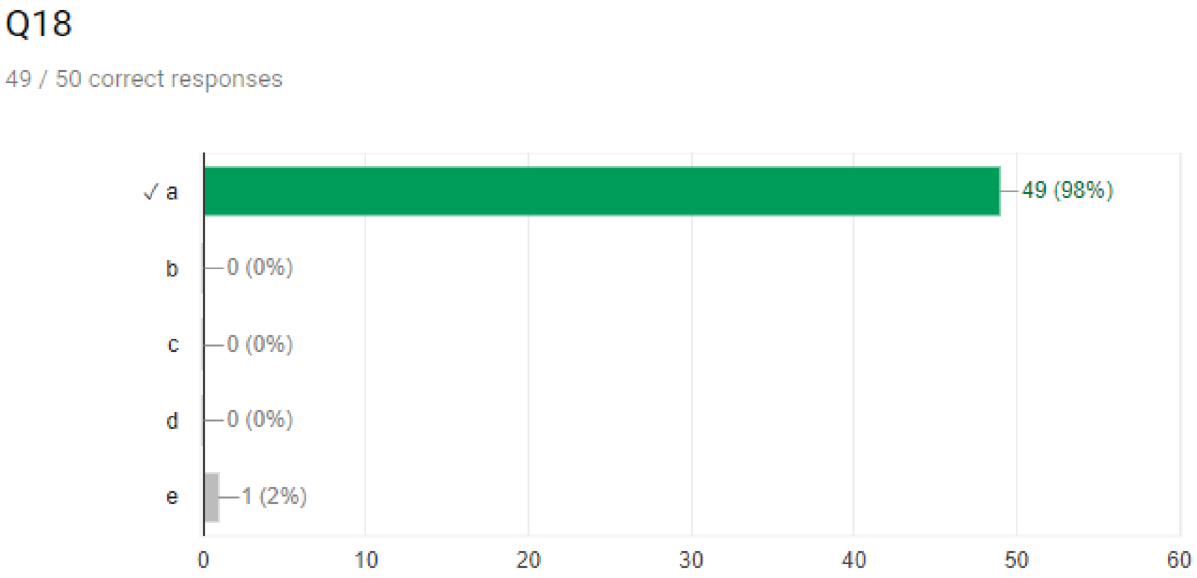
18.1 A18 is a.

The top set of plots shows `Na` and the bottom set of plots shows `Al`.

- From the boxplots, we can easily conclude that Plot 2 must describe `Al`, since that's the only element with a median between 12 and 15, as Plot 2 requires.
- Based again on the median shown in the boxplot in Plot 1, we have two candidates: `Ca` and `Na`, with similar medians. But only `Na` has a maximum value as large as that shown in the plots.

I guess I didn't actually need too much explanation here, as almost everybody got it right.

18.2 Results



Q18

49 / 50 correct responses

19 Q19

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?

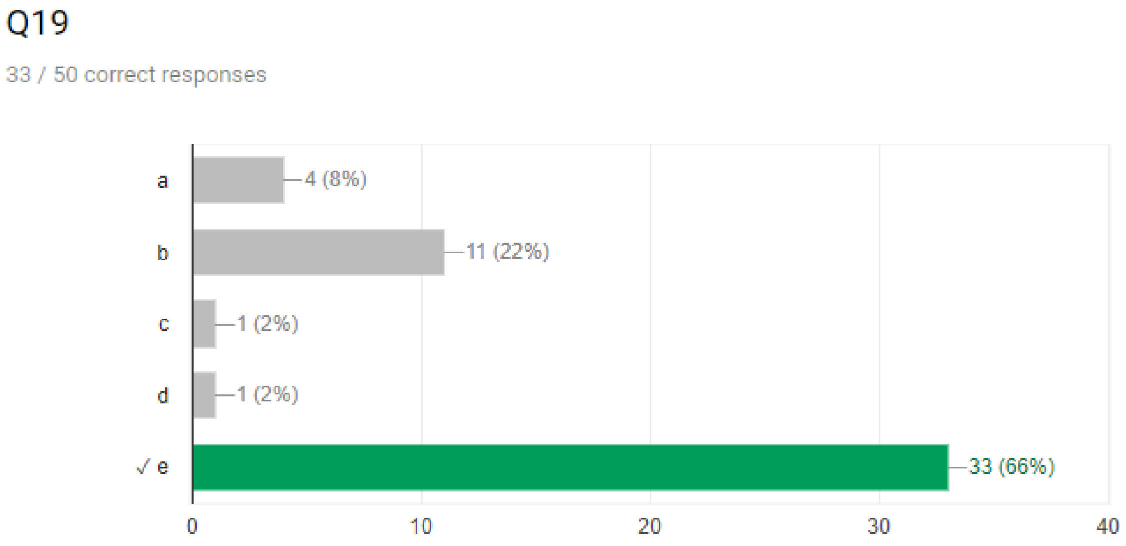
- a. (2.35 - 0.5, 2.35 + 0.5)
- b. (2.35 - 0.1, 2.35 + 0.1)
- c. (2 - 2.35, 2 + 2.35)
- d. (2.35 - 0.35, 2.35 + 0.35)
- e. None of these.

19.1 A19 is `e`.

A good estimate would be to take the estimate and add and subtract 2 times the standard error. That would be `(2.35 - 2*0.5, 2.35 + 2*0.5)`, or (1.35, 3.35). None of our available responses match this result, so we must choose `e`.

The people that chose `b`, I assume, probably just misread the standard error as 0.05, instead of 0.5.

19.2 Results


Q19

33 / 50 correct responses

20 Q20

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

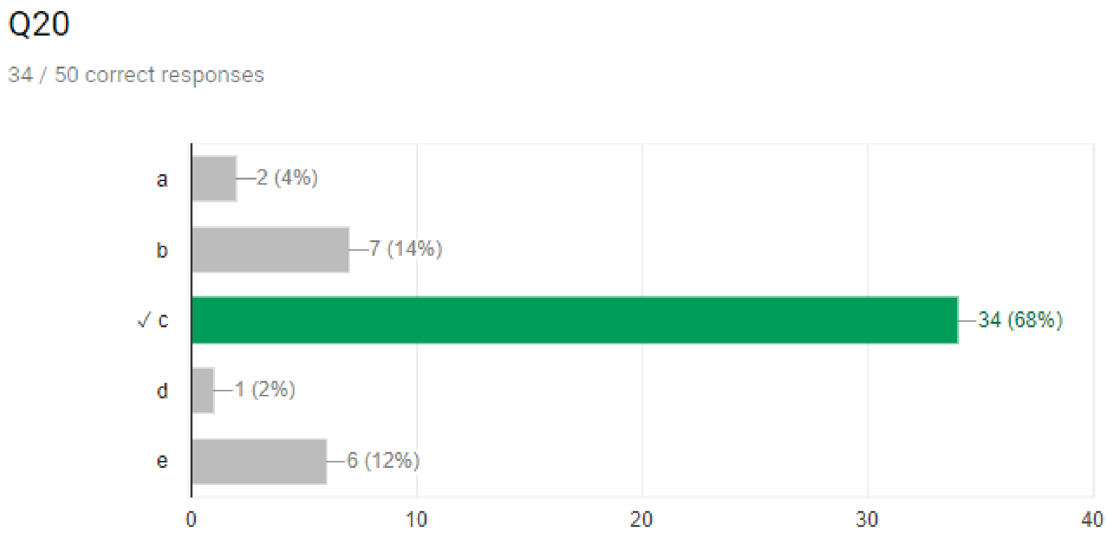
20.1 A20 is `c`.

See, for instance, this quote from Silver in the "Bob the Bayesian" section of Chapter 8.

> The problem with Fisher's notion of hypothesis testing is not with having hypotheses but with the way Fisher recommends that we test them.

Each of the other strategies mentioned (besides `c`) is clearly part of the Bayesian approach, as is explicitly described in the book.

20.2 Results



Q20

34 / 50 correct responses

21 Q21

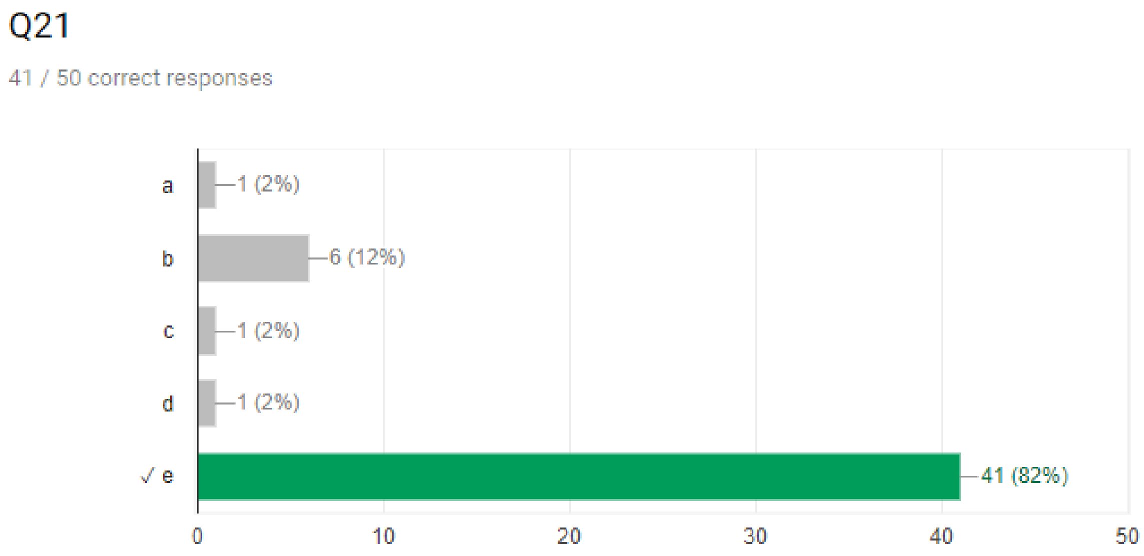
Which of the following statements is **NOT** part of what Silver is trying to tell us in *The Signal and The Noise*? (You may wish to focus on Chapter 13, which summarizes the preceding arguments nicely.)

- 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.

21.1 A21 is e.

Each of the other statements is either an exact quote or a very close paraphrasing of actual text in Chapter 13 of Silver. But statement e is the opposite of what Nate writes, which is "Nature's laws do not change very much."

21.2 Results



Q21

41 / 50 correct responses

22 Q22

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 at https://github.com/thomaselove/431data.

```
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 answered yes to the question “Do you typically check a daily weather report?”

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

22.1 A22 is b.

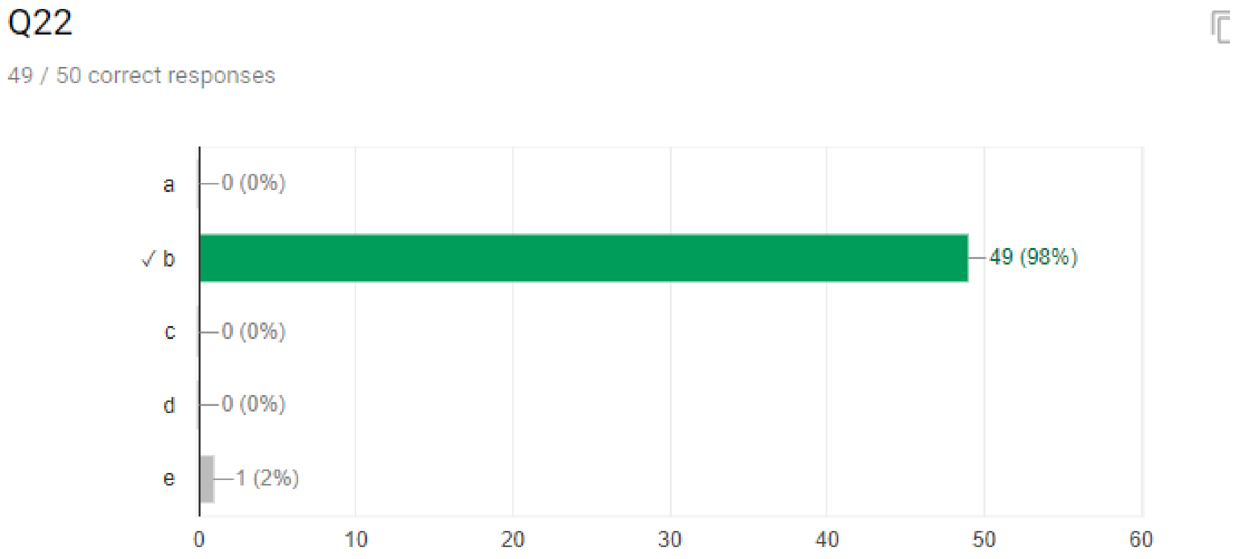
Here’s the code and the resulting table.

```
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",
```

```r
                             "No Check" = "FALSE")) %>%
  mutate(female = fct_relevel(female, "Female"),
         ck_weather = fct_relevel(ck_weather, "Check"))

wctable <- table(wc$female, wc$ck_weather, wc$age)

wctable
```

```
, ,  = 18 - 29


         Check No Check
  Female    79       30
  Male      41       26

, ,  = 30 - 44


         Check No Check
  Female   105       28
  Male      56       15

, ,  = 45 - 59


         Check No Check
  Female   115       29
  Male     119       15

, ,  = 60+


         Check No Check
  Female   121       20
  Male     103       14
```

22.2 Results



Q22

49 / 50 correct responses

23 Q23

Perform an appropriate test to see if the odds ratio for a Yes answer 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 describes an appropriate test.

23.1 A23 is `c.`

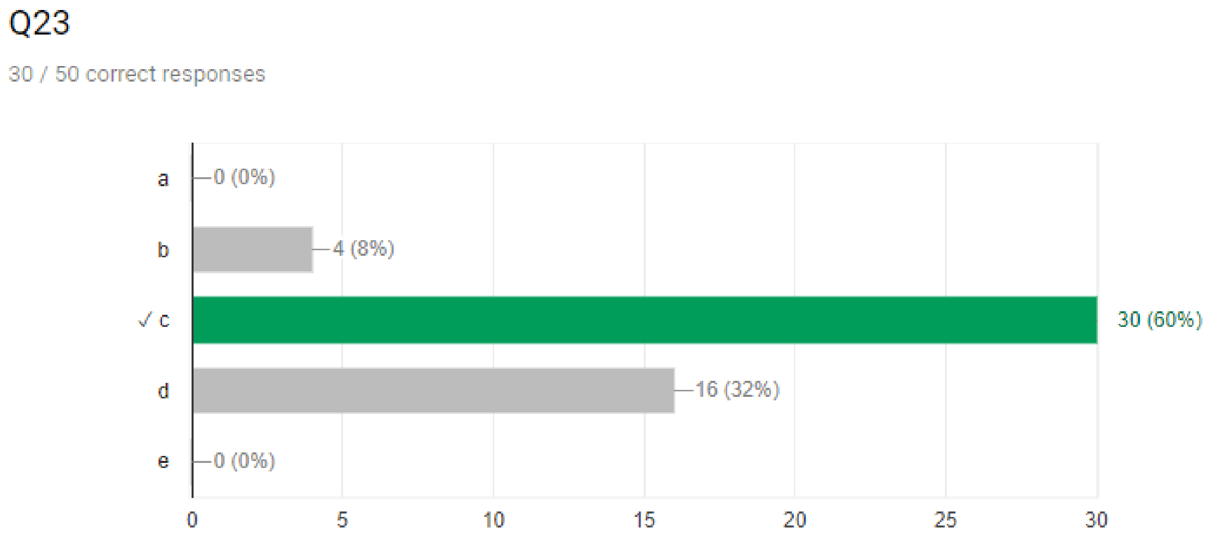
We need to run a Woolf test on the table we just built. Our conclusion matches `c`. There is no significant effect at the 5% level.

```r
vcd::woolf_test(wctable)
```

```

	Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)

data:  wctable
X-squared = 6.5613, df = 3, p-value = 0.08728
```

A surprising number of people had trouble fitting the `woolf_test`, because they forgot that it works on a table, and not a data frame. The people choosing `d` failed to correctly link their result to its interpretation, I think.

23.2 Results



Q23

30 / 50 correct responses

24 Q24

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 accurate.

24.1 A24 is `c`.

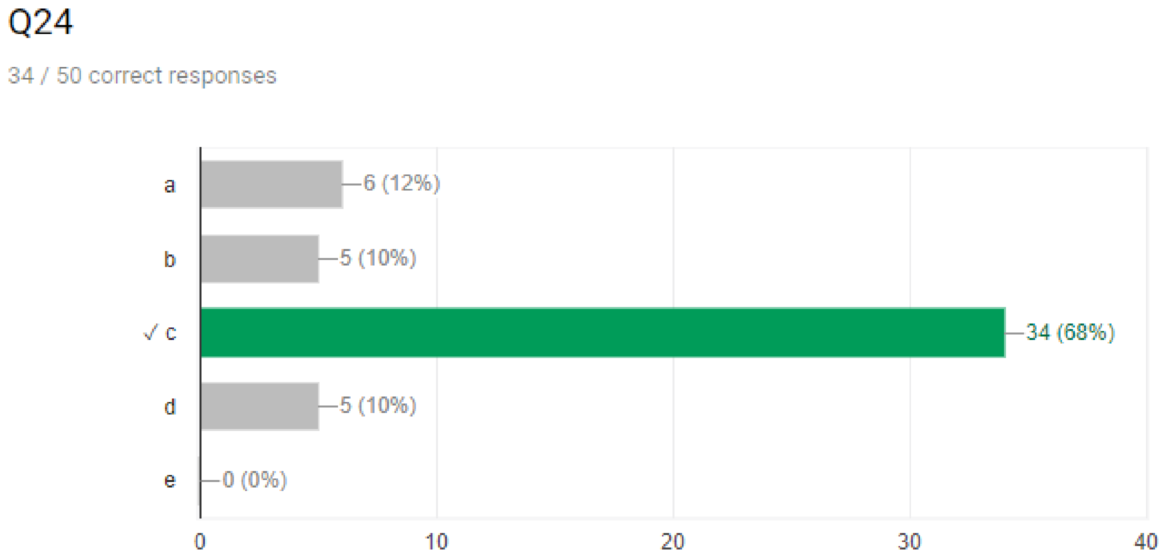
Now we will run the Cochran-Mantel-Haenszel test.

```
mantelhaen.test(wctable)
```

```

    Mantel-Haenszel chi-squared test with continuity correction

data:  wctable
Mantel-Haenszel X-squared = 0.17397, df = 1, p-value = 0.6766
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.6533502 1.2870384
sample estimates:
common odds ratio
        0.9169988
```

The odds ratio is 0.92 for Females vs. Males on checking the weather. That's less than 1 but the confidence interval certainly includes 1.

24.2 Results



Q24

34 / 50 correct responses

25 Q25

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.

25.1 A25 is a.

Approach `a` accomplishes the specified task. The others do not.

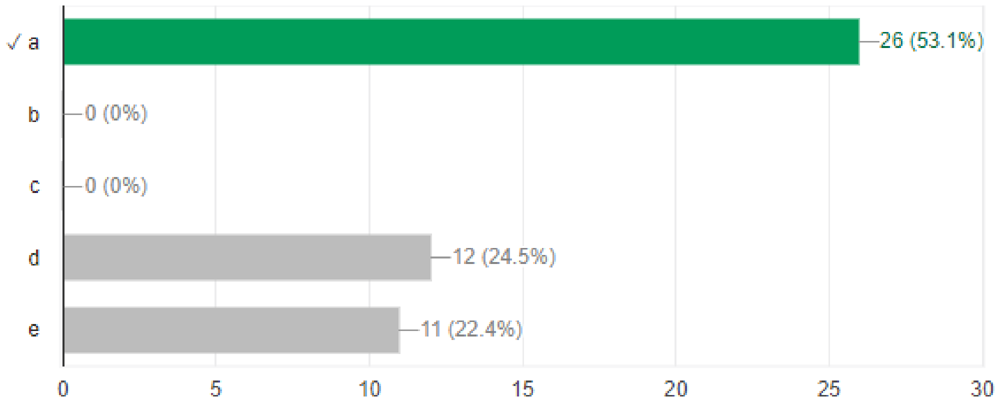
Option `d` was popular. That would create a test sample with 80% of the observations and a training sample with the remaining 20%, which is the opposite of what we're trying to do.

25.2 Results

Note that one person skipped this question. Not a good idea on a quiz like this, where leaving something blank is always wrong, but even a blind guess should be right about 20% of the time.

Q25

26 / 49 correct responses



26 Q26

The southern white rhinoceros (*Ceratotherium simum simum*) has a wild population exceeding 20,000, making them the most abundant rhino species in the world, according to Wikipedia. Your outcome of interest is a measure of size. Suppose that you want to compare those living in an area of southern Africa subject to serious problems from poaching (you have data on 20 rhinos living near Watering Hole A) and those living in an area of southern Africa more than 1,000 km away with a less serious poaching problem (you have data on 20 rhinos living near Watering Hole B). Your interest is to understand how exposure to poaching is associated with average rhino size. An enormous amount of output follows, from R. The data are, alas, simulated. **Some of the output is useful, some is not**.

26.1 Specifying the data frames

```
rhino1 <- data.frame(B, A)
rhino2 <- gather(rhino1, key = "Location", value = "Size") %>%
  mutate(Location = factor(Location))

summary(rhino1)
```

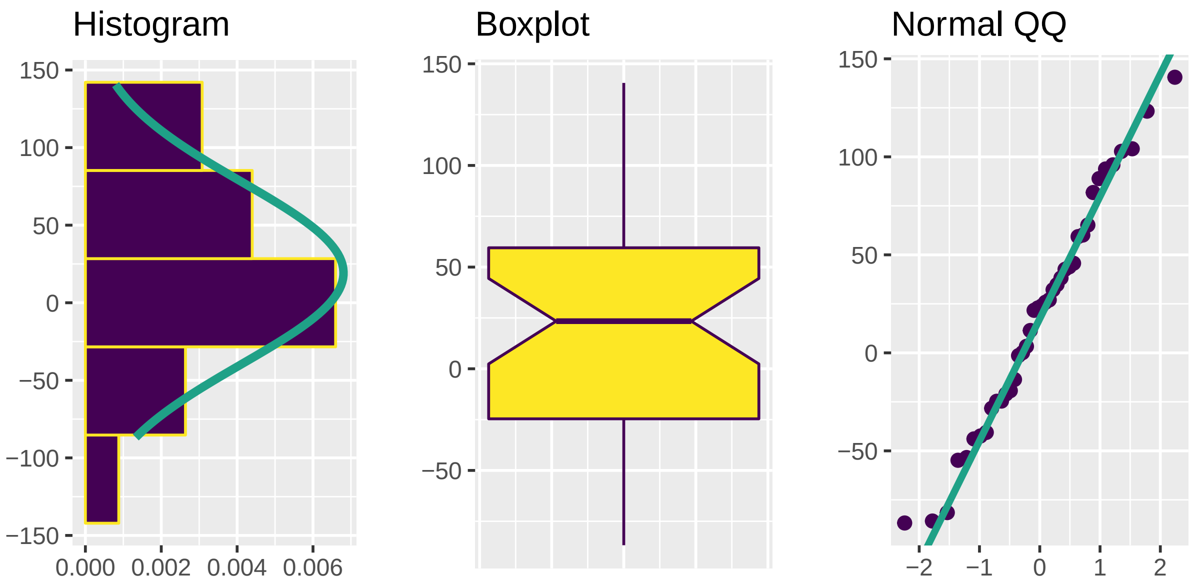
```
       B               A        
 Min.   :477.0   Min.   :396.7  
 1st Qu.:498.2   1st Qu.:466.6  
 Median :521.5   Median :502.1  
 Mean   :520.3   Mean   :501.1  
 3rd Qu.:536.2   3rd Qu.:534.4  
 Max.   :586.0   Max.   :617.8  
```

```
summary(rhino2)
```

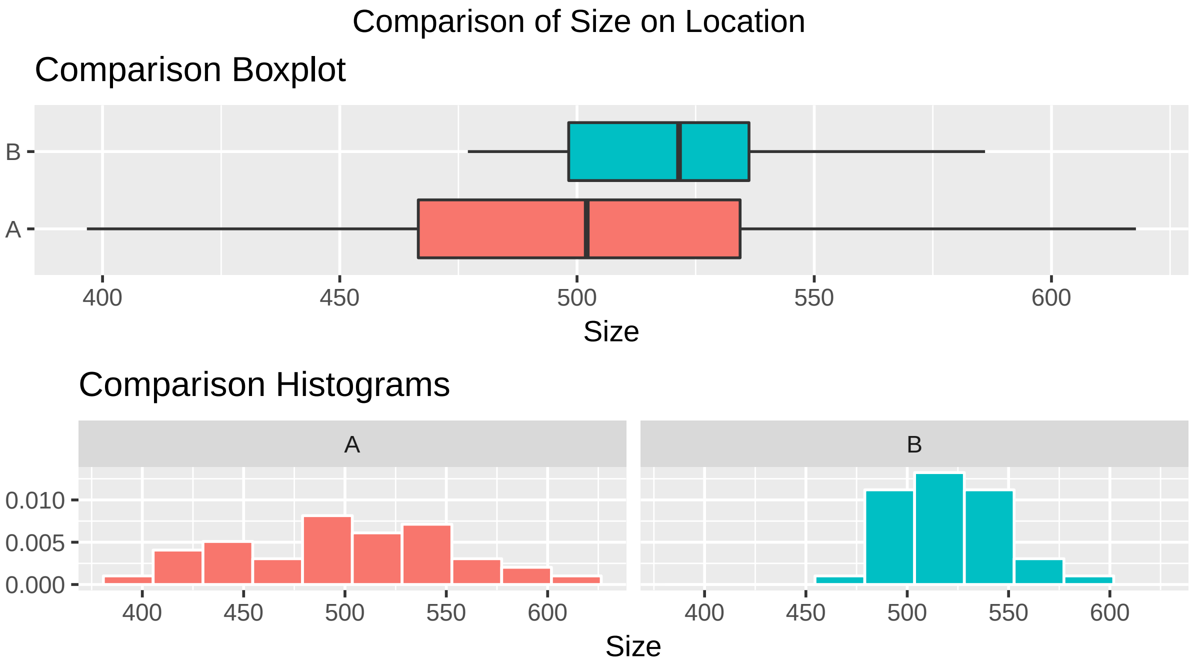
```
 Location      Size      
 A:40     Min.   :396.7  
 B:40     1st Qu.:489.2  
          Median :515.6  
          Mean   :510.7  
          3rd Qu.:535.9  
          Max.   :617.8  
```

26.2 Plotting the data from each data frame

```
rhino1$diff <- rhino1$B - rhino1$A
eda.1sam(dataframe = rhino1, variable = rhino1$diff)
```



```
eda.ksam(outcome = rhino2$Size, group = rhino2$Location, notch = FALSE)
```



26.3 Inference attempts using the rhino1 data frame

```
t.test(rhino1$B - rhino1$A)
```

```

	One Sample t-test

data:  rhino1$B - rhino1$A
t = 2.0691, df = 39, p-value = 0.04521
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  0.4298566 37.9301434
sample estimates:
mean of x 
    19.18 
```

```
wilcox.test(rhino1$B - rhino1$A, conf.int = TRUE)
```

```

	Wilcoxon signed rank test with continuity correction

data:  rhino1$B - rhino1$A
V = 554.5, p-value = 0.05292
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 -0.3999593 38.6999512
sample estimates:
(pseudo)median 
      19.43694 
```

26.4 Inference attempts using the `rhino1` data frame

```
t.test(Size ~ Location, data = rhino2)
```

```

	Welch Two Sample t-test

data:  Size by Location
t = -2.0468, df = 56.567, p-value = 0.04534
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -37.9480188  -0.4119812
sample estimates:
mean in group A mean in group B 
        501.145         520.325 
```

```
wilcox.test(Size ~ Location, data = rhino2, conf.int = TRUE)
```

```

	Wilcoxon rank sum test with continuity correction

data:  Size by Location
W = 599, p-value = 0.05368
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -39.0999437   0.5999544
sample estimates:
difference in location 
             -18.25373 
```

26.5 Now, finally, here is the question for Q26

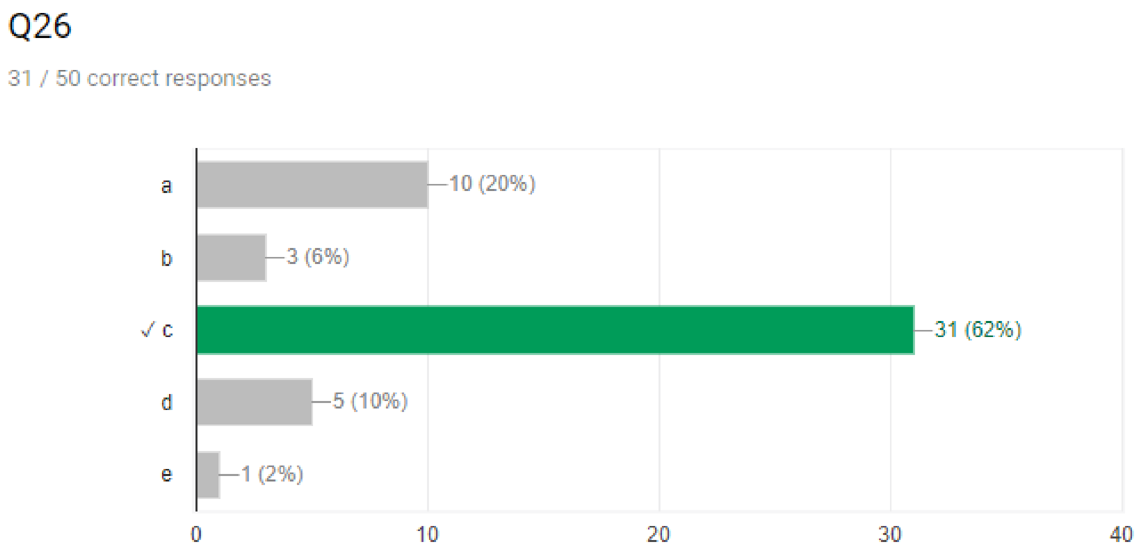
Consider all of the provided output, remembering that **some of it is useless**. Your job is to specify the correct study design, and the resulting 95% confidence interval for the true mean difference in size, where the difference is defined as location B - location A.

- a. Paired samples, and the best estimate shown is (0.43, 37.93)
- b. Paired samples, and the best estimate shown is (-0.40, 38.70)
- c. Independent samples, and the best estimate shown is (0.41, 37.94)
- d. Independent samples, and the best estimate shown is (-0.60, 39.10)
- e. None of these statements are correct.

26.6 A26 is c.

This is an independent samples study. We want to compare means. We're left with the two-sample Welch's t test. That's the confidence interval shown in c. It looks appropriate (we're probably not willing to assume equal variances, but we should be willing to assume Normality based on the eda.ksam result), and it compares means. The paired samples material, looking at paired differences, is of no use.

26.7 Results



Q26

31 / 50 correct responses

27 Q27

Building on our story in Q26, we saw that in location B, where poaching is modest, 29 of the 40 rhinos were of size 500 or greater, while in location A, only 21/40 met that standard.

Suppose that a new set of tough regulations on rhino poaching is to be put into effect in a third location much like location A. We are attempting to calculate the power of a test to detect a change in this new third location. Assume the new location starts at a level of 52.5% without the change in regulations and suppose we assume that changing the regulations will bring the percentage to 70%, which begins to approach the level we saw in location B? Assume a two-sided comparison with a 5% significance level, and that you will be able to study a sample of 100 rhinos in this third location.

Which of the following statements contains the power of this new study?

a. 0 to 20%
b. 21 to 40%
c. 41 to 60%
d. 61 to 80%
e. More than 80%.

27.1 A27 is d.

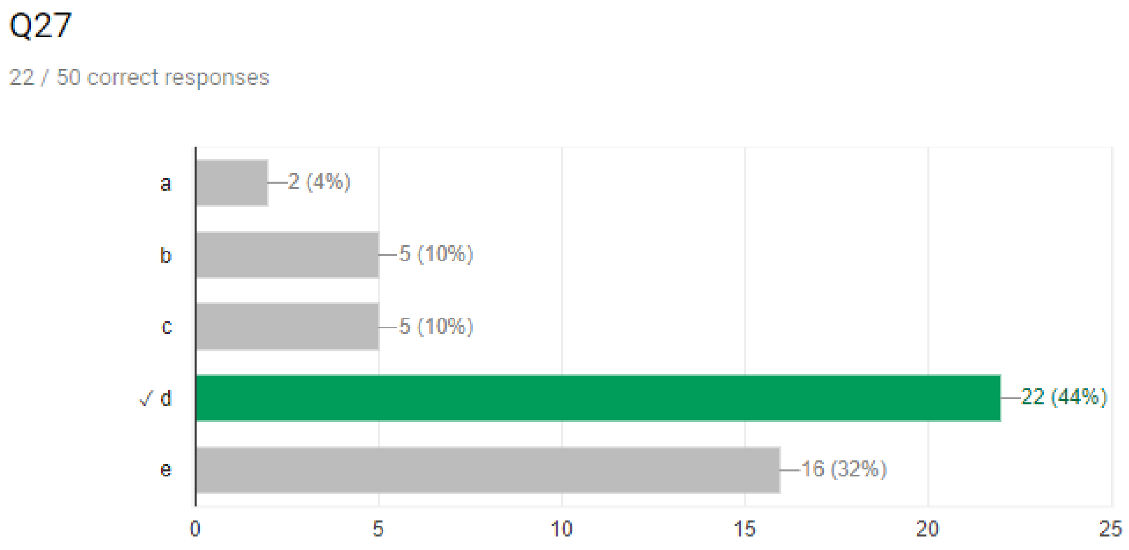
```
power.prop.test(n = 100, p1 = 0.525, p2 = 0.7, sig.level = 0.05)
```

```

     Two-sample comparison of proportions power calculation 

              n = 100
             p1 = 0.525
             p2 = 0.7
      sig.level = 0.05
          power = 0.7222795
    alternative = two.sided

NOTE: n is number in *each* group
```

So the answer is 72%, and that's in `d`.

A reasonable argument could be made that we aren't really doing a two-sample comparison here, so that this power analysis (and any power.prop.test analysis) isn't correct. **So I threw out the item, and just gave everyone 3 points for any response to it.**

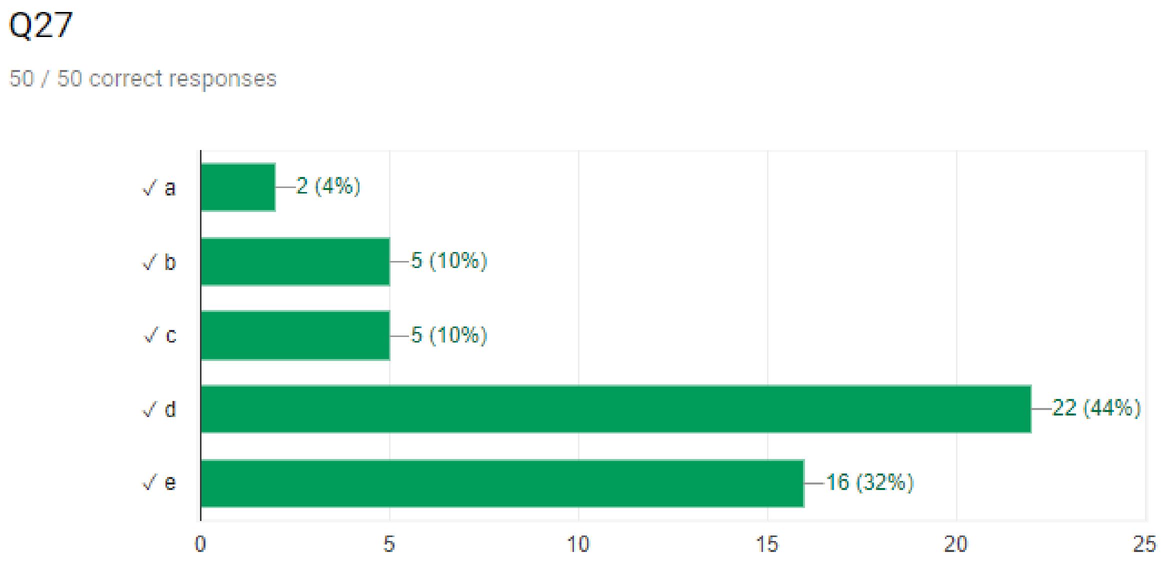
27.2 Original Results for Q27


Q27

22 / 50 correct responses

27.3 Final Results for Q27

Again, I threw out this question, and just gave everyone the 3 points.


Q27

50 / 50 correct responses

28 Q28

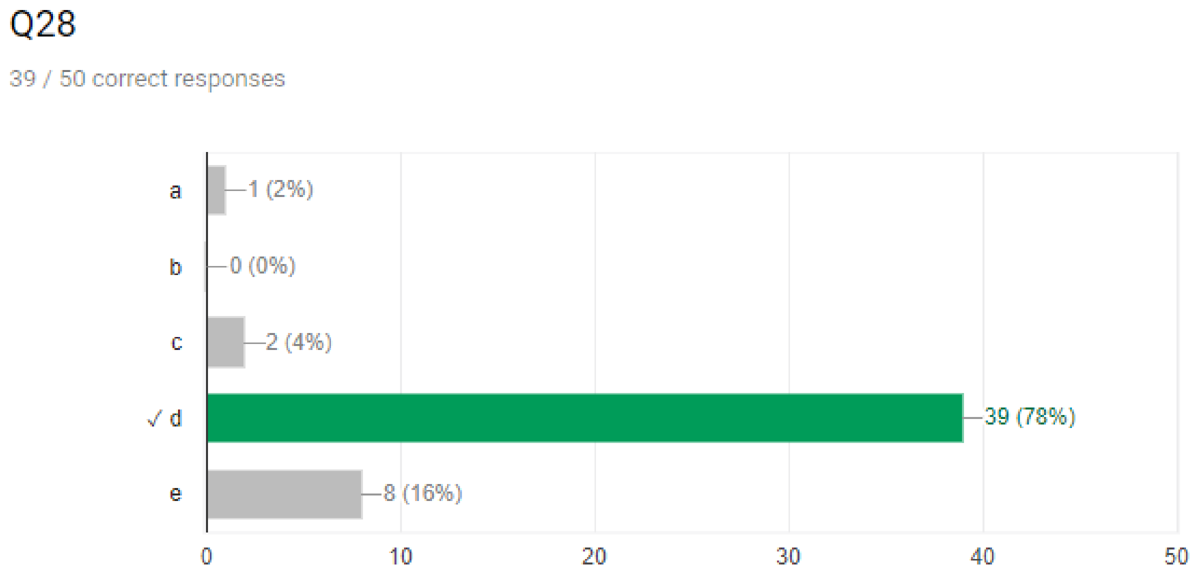
You are comparing two regression models for the same outcome, which you built using a training sample of data. You then use each model to predict data in a test sample, that was not used to calculate the original regression equations. Which of the following summaries will **NOT** be useful to you in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

- a. The mean of the squared prediction errors.
- b. The mean of the absolute values of the prediction errors.
- c. The maximum of the absolute values of the prediction errors.
- d. The AIC statistic calculated in the training sample.
- e. All of these measures would be useful to you.

28.1 A28 is `d`.

Knowing the in-sample performance of AIC isn't going to help you judge out-of-sample validation. The other options (`a`, `b`, and `c`) are all reasonable.

28.2 Results



Q28

39 / 50 correct responses

29 Q29

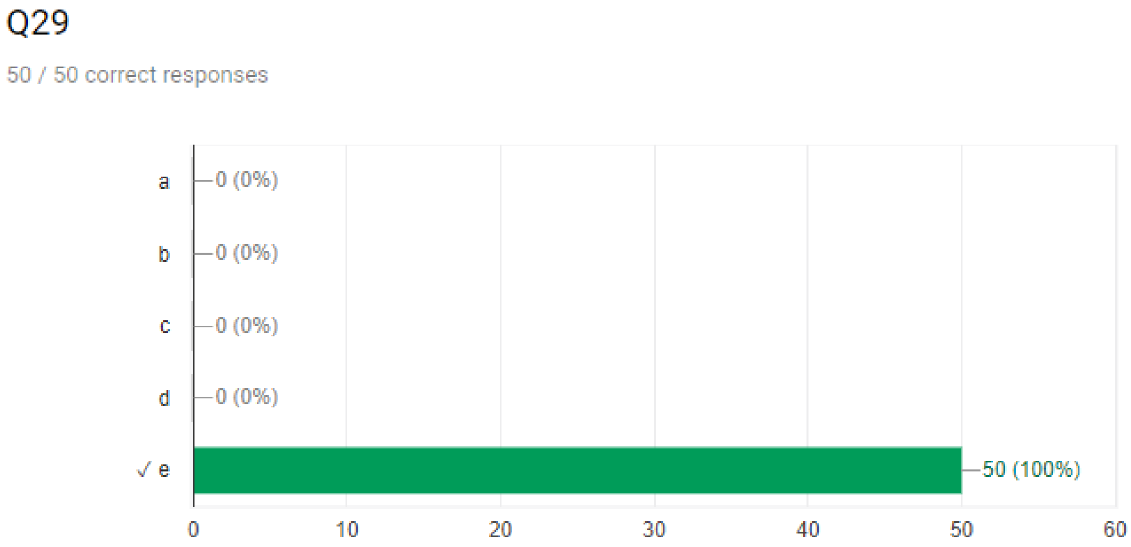
According to Jeff Leek in *The Elements of Data Analytic Style*, which of the following is **NOT** a good reason to create graphs for data exploration?

- 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.

29.1 A29 is e.

If this isn't clear, take a look at Chapter 5 of Jeff's book.

29.2 Results



Q29

50 / 50 correct responses

30 Q30

According to Jeff Leek in *The Elements of Data Analytic Style*, which of the following is **NOT** a good idea in creating graphs you will share with other people to describe your work?

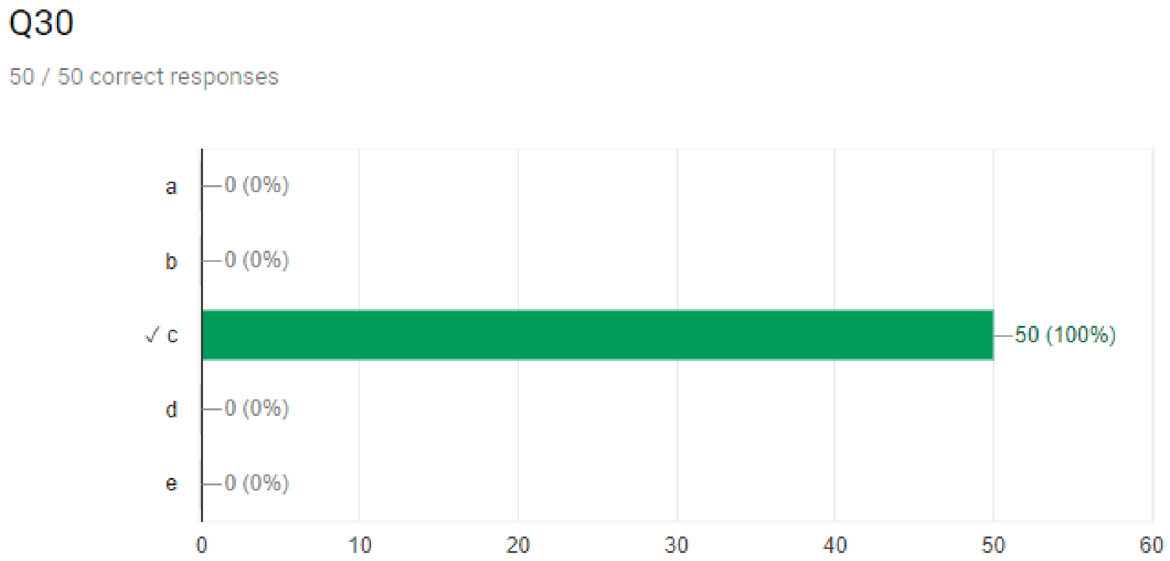
- a. If you have multiple plots to compare, use the same scale on the vertical axis.

- b. Axis labels should be large, easy to read, in plain language.
- c. Add a third dimension, perhaps with animation.
- d. Include units in figure labels and legends.
- e. Use color and size to help communicate information, for instance to point out confounding.

30.1 A30 is c.

If this isn't clear, take a look at Chapter 10 of Jeff's book.

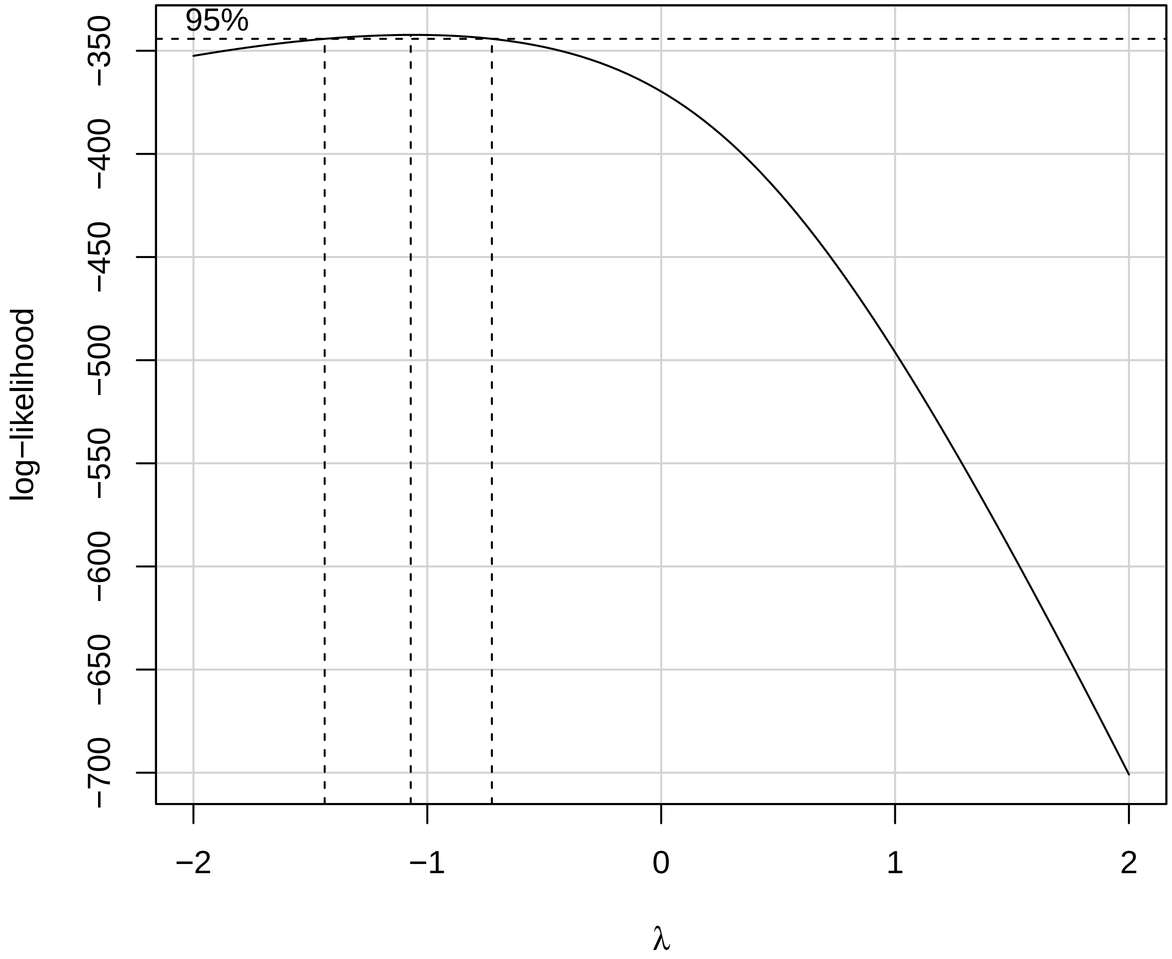
30.2 Results


Q30

50 / 50 correct responses

31 Q31

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



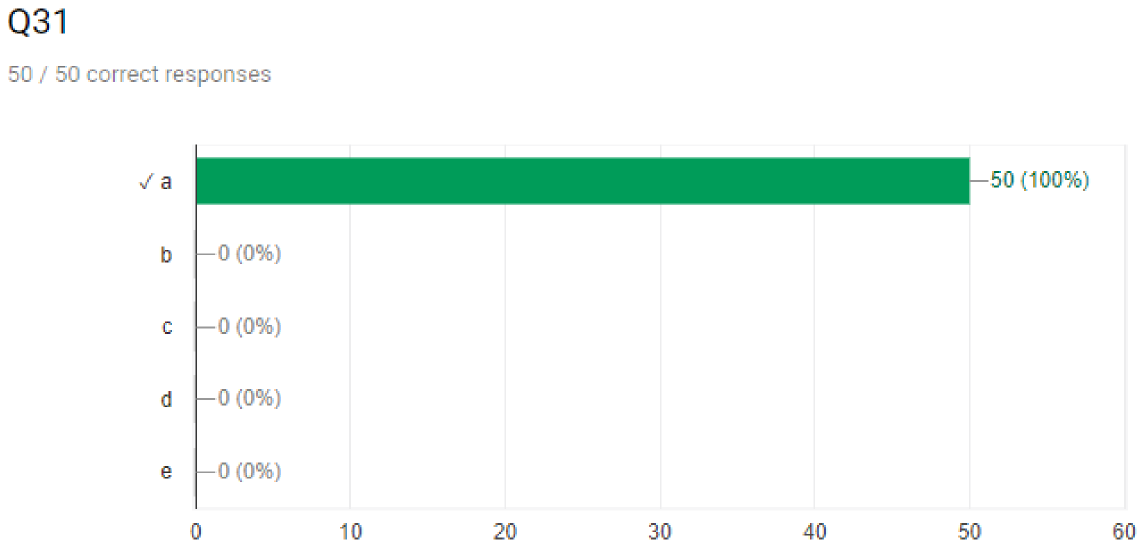
What transformation of our response does this plot suggest?

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

31.1 A31 is a.

The suggested power is clearly near -1, which indicates the inverse of our outcome.

31.2 Results



Q31

50 / 50 correct responses

32 Q32

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?

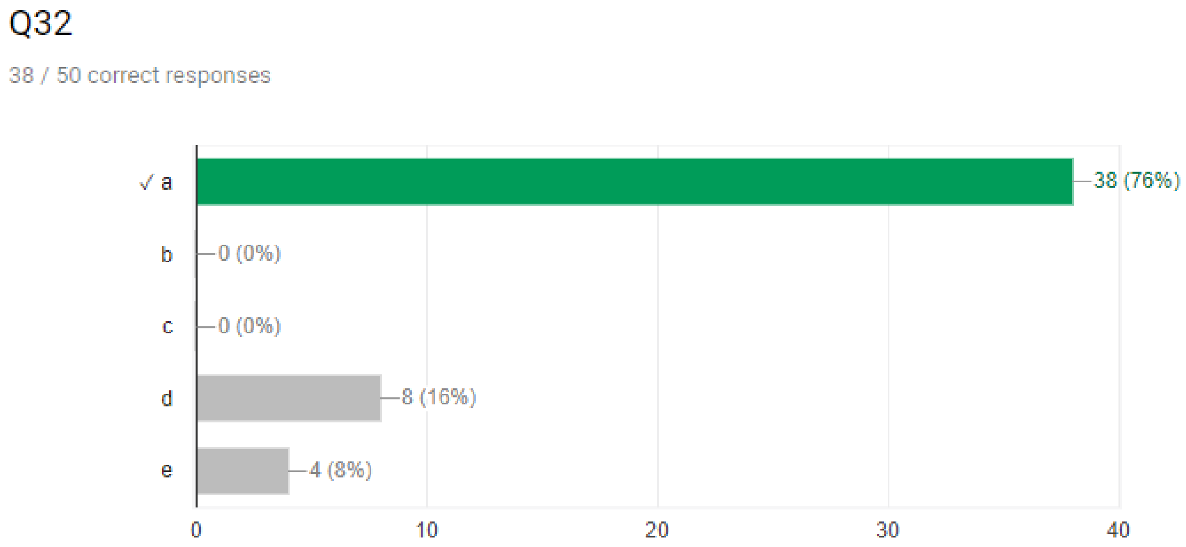
- a. Model A
- b. Model B
- c. Model C
- d. Model D
- e. It is impossible to tell from the information provided.

32.1 A32 is `a`.

Model A has a much smaller gap between its Multiple R^2 and its Adjusted R^2 than do its comparison models.

Choice `d` was a popular error. Model D has a nominal R^2 value of 0.55, so while its Adjusted R^2 is larger than, say, Model A, it is still a good distance away from the nominal value.

32.2 Results



Q32

38 / 50 correct responses

33 Q33

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 has Dr. Love used to accomplish such a plot?

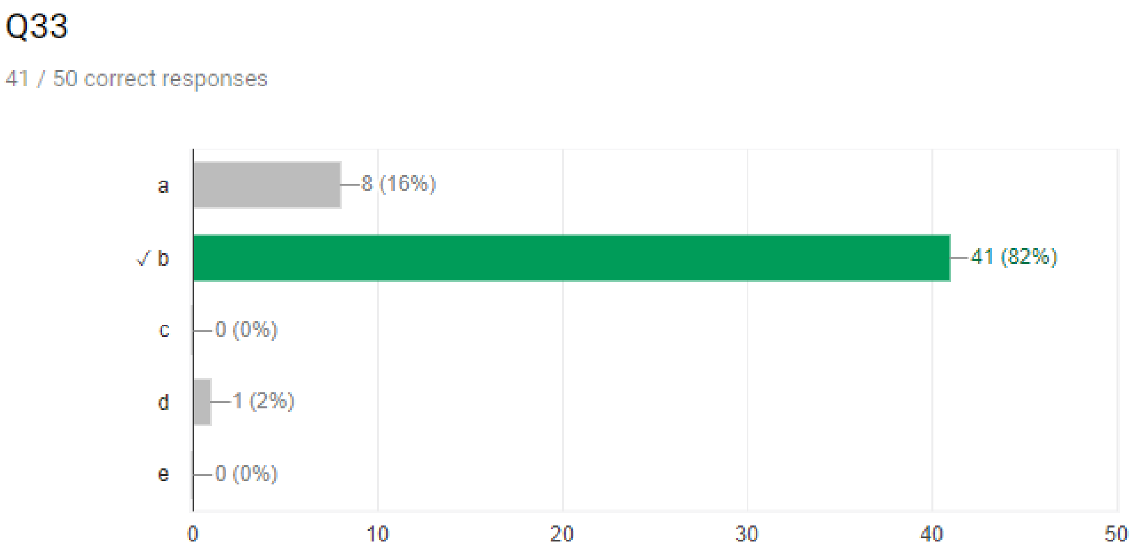
- a. `fct_reorder` and `facet_wrap`
- b. `fct_relevel` and `facet_wrap`
- c. `fct_collapse` and `facet_wrap`
- d. `fct_reorder` and `group_by`
- e. `fct_collapse` and `group_by`

33.1 A33 is `b`.

`fct_relevel` lets you specify a new order for factor levels "by hand" which is what you'd need to do here, and `facet_wrap` is the easier approach to getting the individual subsetted histograms in this context.

Choice `a` was popular, too, but `fct_reorder` lets you reorder the factor levels by sorting along another variable's values, and that's not what we're doing here.

33.2 Results



Q33

41 / 50 correct responses

33.3 Setup for Q34-40

For Q34 - Q40, consider the data I have provided in the `hospsim.csv` file at https://github.com/thomaselove/431data. The data describe 750 patients at a metropolitan hospital. They are simulated. Available are:

- `subject.id` = Subject Identification Number (not a meaningful code)
- `age` = the patient's age, in years (all subjects are between 21 and 75)
- `sex` = the patient's sex (FEMALE or MALE)
- `a1c` = the patient's hemoglobin A1c level (in %)
- `ldl` = the patient's LDL cholesterol level (in mg/dl)
- `sbp` = the patient's systolic blood pressure (in mm Hg)
- `bmi` = the patient's body mass index (in kg/square meter)
- `statin` = does the patient have a prescription for a statin medication (YES or NO)
- `insurance` = the patient's insurance type (MEDICARE, COMMERCIAL, MEDICAID, UNINSURED)
- `hsgrads` = the percentage of adults in the patient's home neighborhood who have at least a high school diploma (this measure of educational attainment is used as an indicator of the socio-economic place in which the patient lives)
- `clinic.type` = whether the patient goes to a newly built clinic or an old clinic

34 Q34

Using the `hospsim` data, what is the 95% confidence interval for the odds ratio which compares the odds of receiving a statin if you are MALE divided by the odds of receiving a statin if you are FEMALE. Do **NOT** use a Bayesian augmentation here.

- a. Odds Ratio is 0.48, CI is (0.33, 0.70)
- b. Odds Ratio is 0.86, CI is (0.81, 0.93)
- c. Odds Ratio is 1.16, CI is (1.07, 1.24)
- d. Odds Ratio is 2.07, CI is (1.42, 3.01)
- e. None of these answers are correct.

34.1 A34 is `d`.

The correct odds ratio is 2.07, with 95% CI (1.42, 3.01).

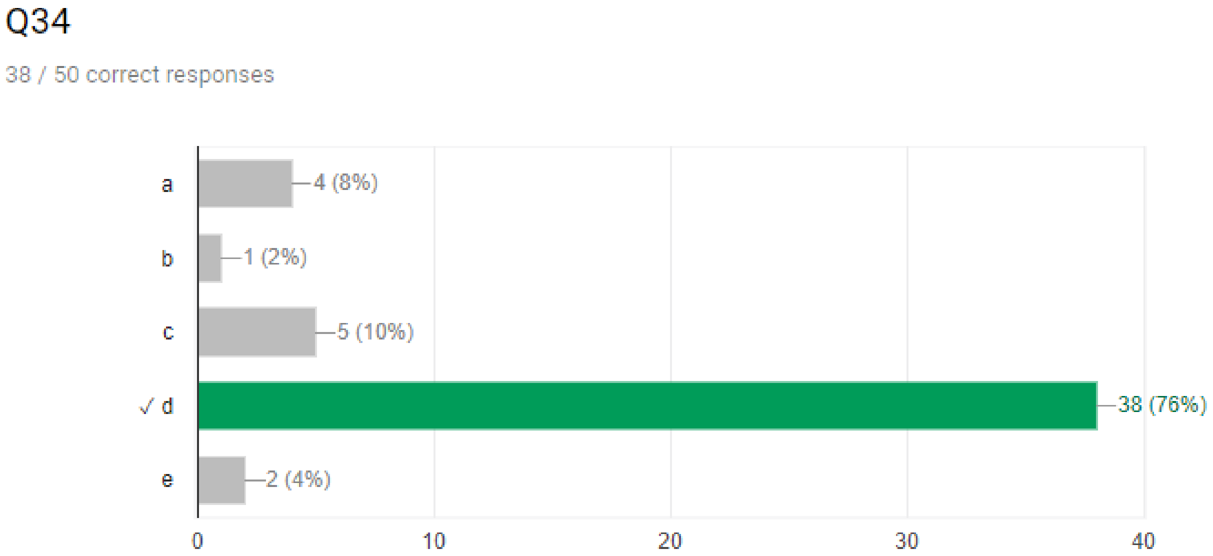
```r
hospsim <- read.csv("data/hospsim.csv") %>% tbl_df

table(hospsim$sex, hospsim$statin)
```

```
          NO YES
  FEMALE 107 305
  MALE    49 289
```

```
twobytwo(289, 49, 305, 107, "Male", "Female", "Yes", "No")
```

```
Loading required package: Epi


Attaching package: 'Epi'

The following object is masked from 'package:base':

    merge.data.frame

2 by 2 table analysis:
------------------------------------------------------
Outcome   : Yes
Comparing : Male vs. Female

       Yes  No    P(Yes) 95% conf. interval
Male   289  49    0.8550    0.8133   0.8887
Female 305 107    0.7403    0.6958   0.7803

                                   95% conf. interval
             Relative Risk: 1.1550    1.0747   1.2413
         Sample Odds Ratio: 2.0691    1.4229   3.0088
Conditional MLE Odds Ratio: 2.0672    1.4034   3.0763
    Probability difference: 0.1147    0.0571   0.1704

             Exact P-value: 1e-04
        Asymptotic P-value: 1e-04
------------------------------------------------------
```

34.2 Results



Q34

38 / 50 correct responses

35 Q35

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.

35.1 A35 is c.

```
anova(lm(hsgrads ~ insurance, data = hospsim))
```

```
Analysis of Variance Table

Response: hsgrads
           Df Sum Sq Mean Sq F value    Pr(>F)    
insurance   3   3580 1193.46  10.794 5.992e-07 ***
Residuals 746  82481  110.56                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

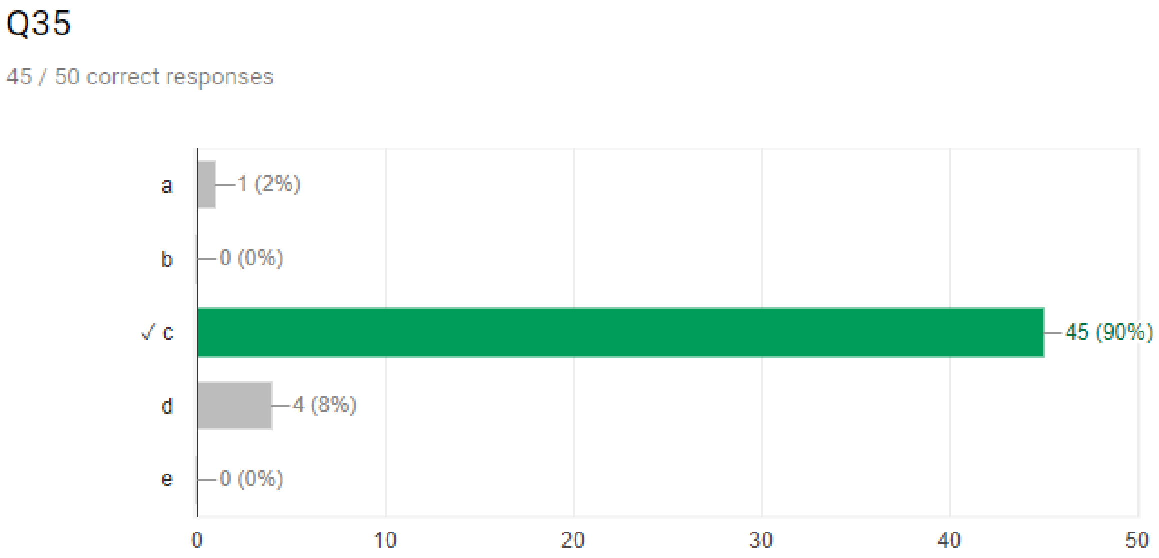
```
TukeyHSD(aov(hsgrads ~ insurance, data = hospsim))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = hsgrads ~ insurance, data = hospsim)

$insurance
                          diff       lwr       upr     p adj
MEDICAID-COMMERCIAL  -6.4368450 -9.467453 -3.406237 0.0000004
MEDICARE-COMMERCIAL  -0.8032317 -2.989483  1.383020 0.7799896
UNINSURED-COMMERCIAL -3.7118288 -8.960222  1.536564 0.2641848
MEDICARE-MEDICAID     5.6336133  2.563514  8.703713 0.0000163
```

```
UNINSURED-MEDICAID    2.7250163 -2.948734  8.398767 0.6036857
UNINSURED-MEDICARE   -2.9085971 -8.179893  2.362698 0.4867175
```

Medicaid is significantly lower than either Commercial or Medicare. That's option `c`.

35.2 Results



Q35

45 / 50 correct responses

36 Q36

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.

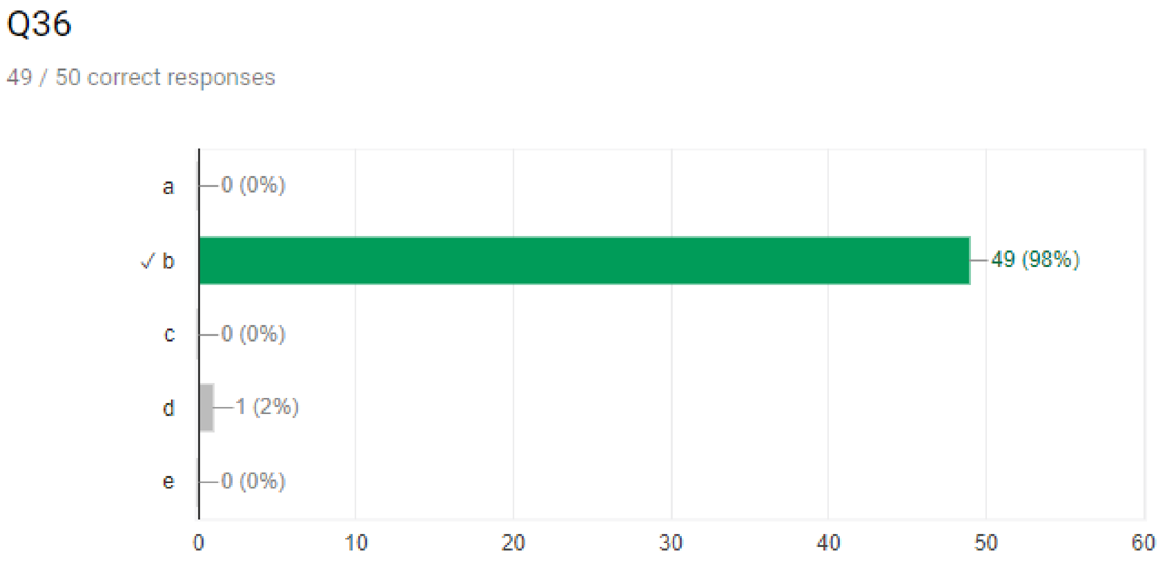
36.1 A36 is `b`.

```r
model_q36 <- lm(ldl ~ clinic.type + age + sex +
                  insurance + hsgrads + a1c + bmi +
                  sbp + statin, data = hospsim)

summary(model_q36)
```

```

Call:
lm(formula = ldl ~ clinic.type + age + sex + insurance + hsgrads +
    a1c + bmi + sbp + statin, data = hospsim)

Residuals:
    Min      1Q  Median      3Q     Max
-80.537 -25.332  -5.444  19.466 144.280

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         76.17788   21.39231   3.561 0.000393 ***
clinic.typeOLD       9.30018    3.10478   2.995 0.002832 **
age                 -0.20805    0.17414  -1.195 0.232594
sexMALE             -7.76691    2.75510  -2.819 0.004945 **
insuranceMEDICAID   -2.24584    4.55044  -0.494 0.621775
insuranceMEDICARE   -4.00606    3.48198  -1.151 0.250306
insuranceUNINSURED  12.57047    7.19072   1.748 0.080854 .
hsgrads              0.07277    0.12912   0.564 0.573221
a1c                  2.47784    0.69390   3.571 0.000379 ***
bmi                 -0.37129    0.18590  -1.997 0.046157 *
sbp                  0.26646    0.09204   2.895 0.003905 **
statinYES          -10.39065    3.46619  -2.998 0.002811 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.38 on 738 degrees of freedom
Multiple R-squared:  0.09209,   Adjusted R-squared:  0.07856
F-statistic: 6.805 on 11 and 738 DF,  p-value: 5.715e-11
```

Yes. Older Clinics have significantly higher LDL cholesterol after accounting for all other variables. The overall R-square is 9.2%, well less than 20%. So the correct answer is `b`.

36.2 Results



Q36

49 / 50 correct responses

37 Q37

Run a backwards elimination stepwise procedure. After doing so, how many of the original nine regression inputs (`clinic.type`, `age`, `sex`, `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

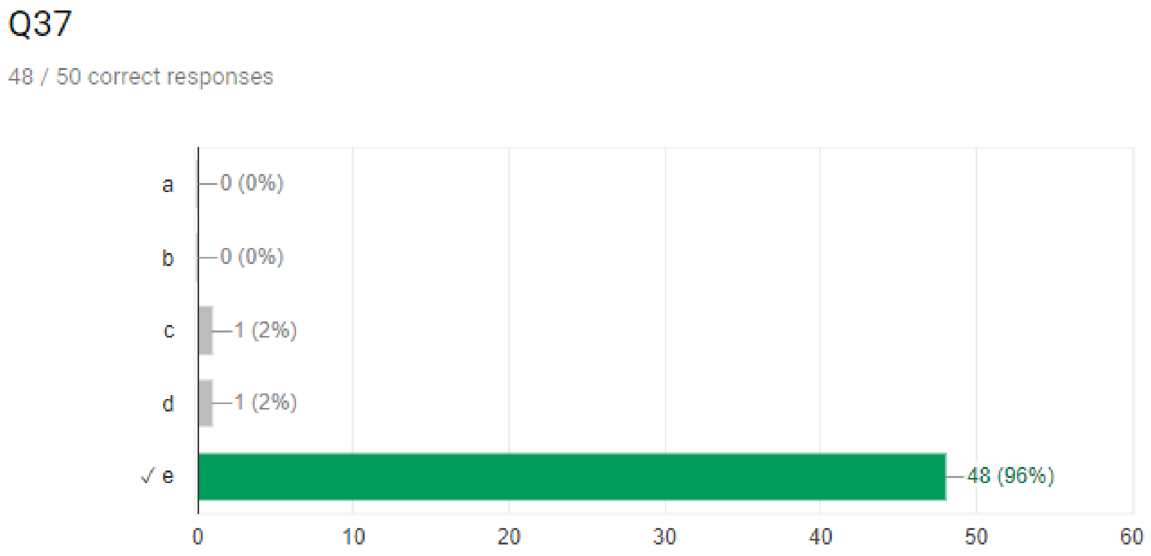
37.1 A37 is `e`.

Seven inputs remain. The only two that drop out are `hsgrads` and `age`.

```r
step(model_q36)
```

```
Start:  AIC=5402.89
ldl ~ clinic.type + age + sex + insurance + hsgrads + a1c + bmi + 
    sbp + statin

              Df Sum of Sq    RSS    AIC
- hsgrads      1     420.3 977110 5401.2
- age          1    1888.9 978579 5402.3
- insurance    3    7305.6 983995 5402.5
<none>                     976690 5402.9
- bmi          1    5279.5 981969 5404.9
- sex          1   10517.7 987207 5408.9
- sbp          1   11091.1 987781 5409.4
- clinic.type  1   11874.6 988564 5410.0
- statin       1   11892.7 988582 5410.0
- a1c          1   16875.3 993565 5413.7

Step:  AIC=5401.21
ldl ~ clinic.type + age + sex + insurance + a1c + bmi + sbp + 
    statin

              Df Sum of Sq    RSS    AIC
- age          1    1820.3 978930 5400.6
- insurance    3    7411.8 984522 5400.9
<none>                     977110 5401.2
- bmi          1    5217.7 982328 5403.2
- sex          1   10176.9 987287 5407.0
- sbp          1   10808.7 987919 5407.5
- clinic.type  1   11501.0 988611 5408.0
- statin       1   12105.1 989215 5408.4
- a1c          1   16683.1 993793 5411.9

Step:  AIC=5400.61
ldl ~ clinic.type + sex + insurance + a1c + bmi + sbp + statin

              Df Sum of Sq    RSS    AIC
<none>                     978930 5400.6
- bmi          1    4078.3 983009 5401.7
- insurance    3   12120.4 991051 5403.8
- sex          1    9592.1 988522 5405.9
- sbp          1    9789.2 988720 5406.1
- clinic.type  1   13593.0 992523 5408.9
- statin       1   15069.2 994000 5410.1
- a1c          1   19161.7 998092 5413.1


Call:
lm(formula = ldl ~ clinic.type + sex + insurance + a1c + bmi +
    sbp + statin, data = hospsim)

Coefficients:
       (Intercept)      clinic.typeOLD             sexMALE   insuranceMEDICAID
           70.9329              9.7166             -7.3483             -1.2067
 insuranceMEDICARE  insuranceUNINSURED                 a1c                 bmi
           -6.1510             12.8761              2.5997             -0.3167
               sbp           statinYES
            0.2470            -11.3866
```

37.2 Results



Q37

48 / 50 correct responses

38 Q38

Compare your initial "kitchen sink" model with all 9 inputs to the model generated by the stepwise approach in Q37 using adjusted R^2, AIC and BIC. What conclusions can you draw?

The smaller model (stepwise result from Q37) is . . .

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

38.1 A38 is `e`.

Here model 1 is the original model, and model 2 is the stepwise result. The Adjusted R^2 is slightly higher (better) in Model 2, and the AIC and BIC are each lower (better) in Model 2.

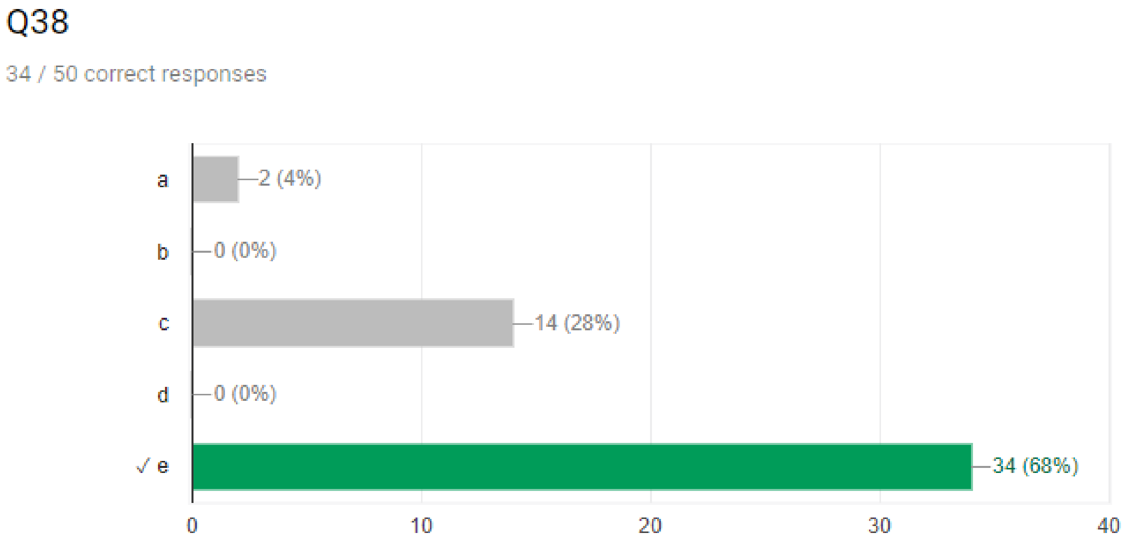
```
select(glance(model_q36), adj.r.squared, AIC, BIC)
```

```
  adj.r.squared      AIC      BIC
1    0.07856249 7533.296 7593.357
```

```
model_q38 <- lm(ldl ~ clinic.type + sex + insurance + a1c +
                  bmi + sbp + statin, data = hospsim)

select(glance(model_q38), adj.r.squared, AIC, BIC)
```

```
  adj.r.squared      AIC      BIC
1    0.07894474 7531.015 7581.836
```

38.2 Results



Q38

34 / 50 correct responses

39 Q39

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

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

39.1 A39 is a.

Both sex and insurance add significant value, but the R^2 is quite low.

```r
model_q39 <- lm(a1c ~ sex + insurance, data = hospsim)

summary(model_q39)
```

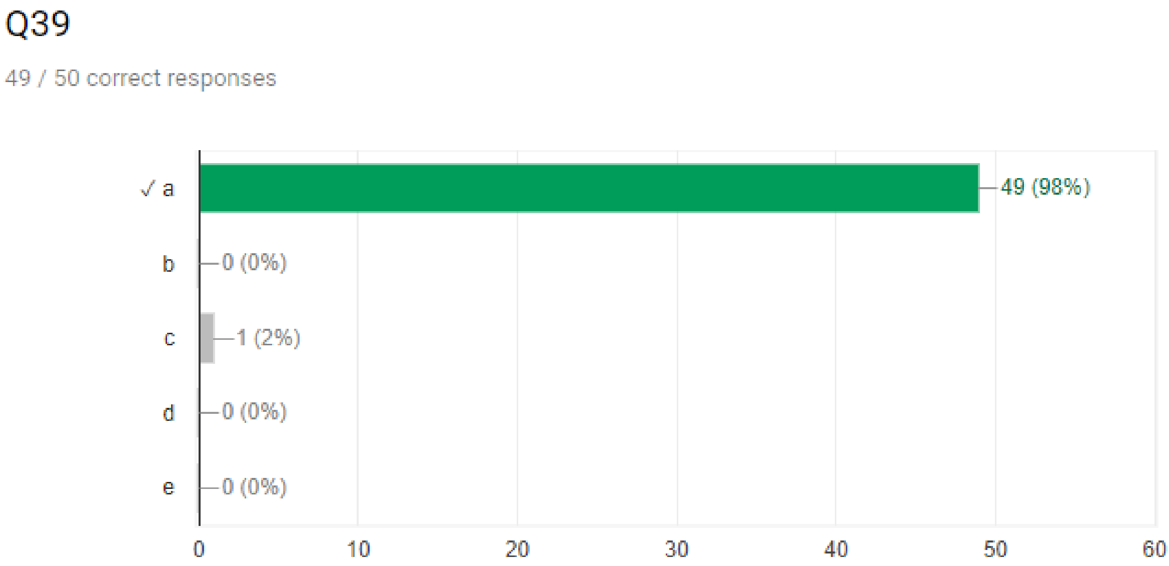
```

Call:
lm(formula = a1c ~ sex + insurance, data = hospsim)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5484 -1.3584 -0.4914  0.9067  8.3067 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)          7.7914     0.1298  60.005  < 2e-16 ***
sexMALE              0.2867     0.1441   1.990  0.04691 *  
insuranceMEDICAID    0.2703     0.2189   1.235  0.21727    
insuranceMEDICARE   -0.4981     0.1584  -3.143  0.00174 ** 
insuranceUNINSURED   0.8672     0.3790   2.288  0.02240 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.955 on 745 degrees of freedom
Multiple R-squared:  0.03769,   Adjusted R-squared:  0.03253 
```

```
F-statistic: 7.295 on 4 and 745 DF,  p-value: 9.157e-06
```

```r
anova(model_q39)
```

```
Analysis of Variance Table

Response: a1c
           Df  Sum Sq Mean Sq F value    Pr(>F)    
sex         1   21.20 21.2006  5.5477   0.01876 *  
insurance   3   90.31 30.1041  7.8776 3.528e-05 ***
Residuals 745 2847.01  3.8215                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

39.2 Results



Q39

49 / 50 correct responses

40 Q40

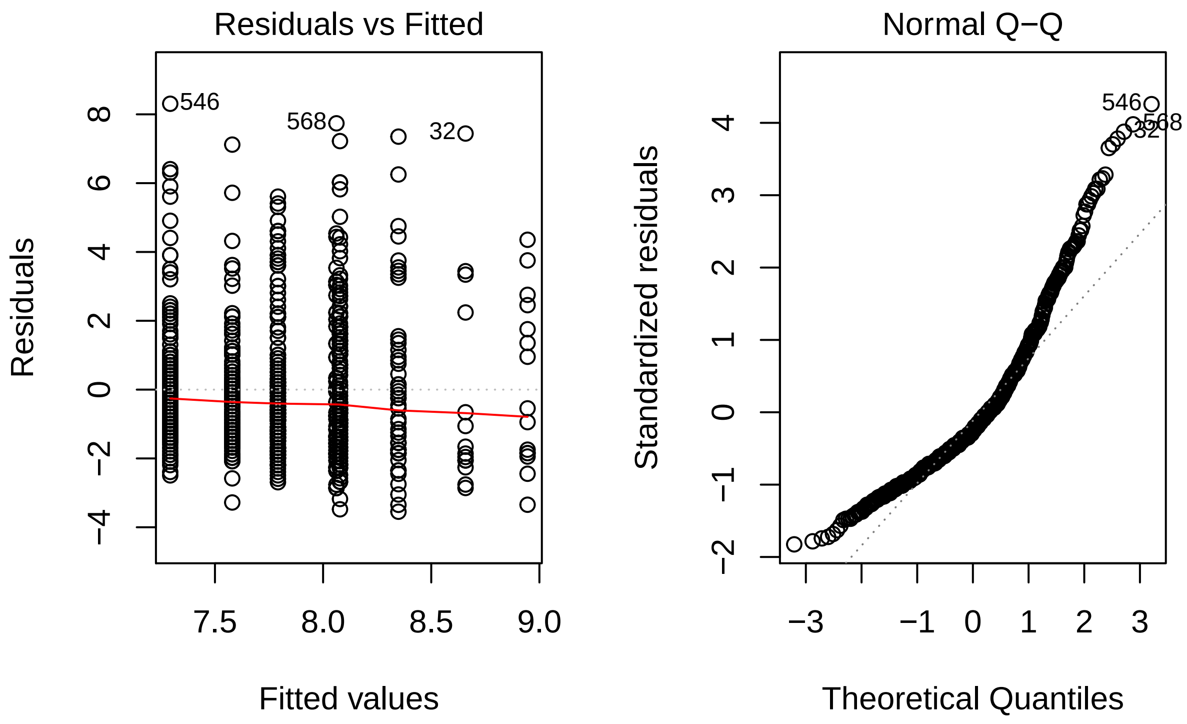
In your model for Q39, identify the subject with the largest residual. Which of the following characteristics best describes this subject?

- a. This is a female Medicare patient visiting a new clinic.
- b. This is a female Medicare patient visiting an old clinic.
- c. This is a male Medicare patient visiting a new clinic.
- d. This is a male Medicare patient visiting an old clinic.
- e. None of these accurately describe the subject in question.

40.1 A40 is `b`.

She's the patient in the 546th row of the data. She's a 28-year old Female Medicare patient in an Old clinic.

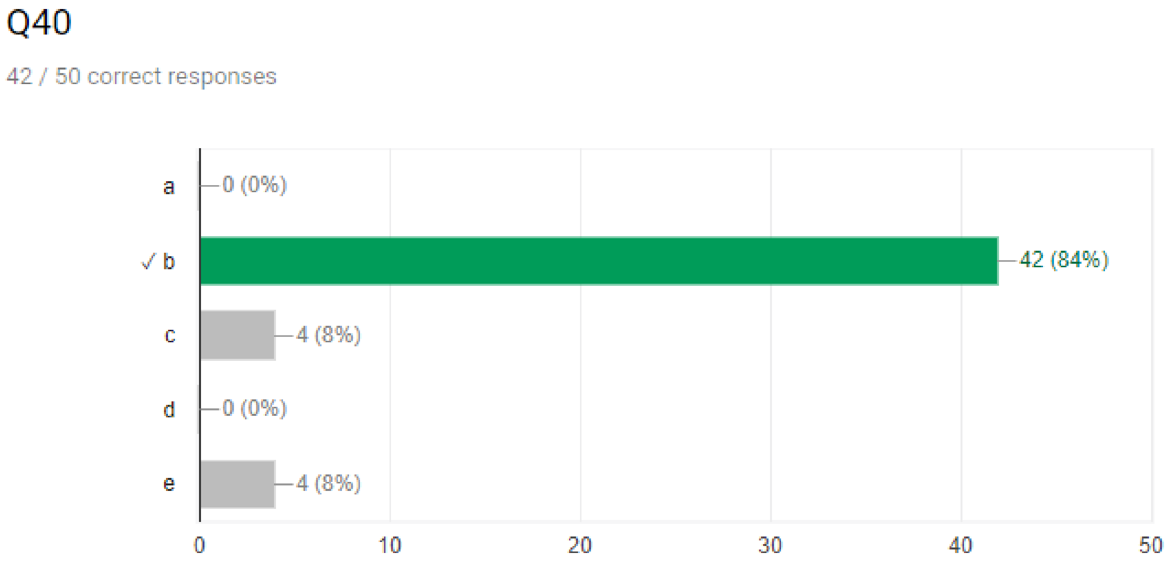
```r
par(mfrow=c(1,2))
plot(model_q39, which = 1:2)
```



```r
par(mfrow=c(1,1))

hospsim %>% slice(546)
```

```
# A tibble: 1 x 11
  subject.id clinic.type   age    sex insurance hsgrads   a1c   ldl   bmi   sbp
      <fctr>      <fctr> <int> <fctr>    <fctr>   <dbl> <dbl> <int> <dbl> <int>
1      X1546         OLD    28 FEMALE  MEDICARE    89.9  15.6   179  25.1   118
# ... with 1 more variables: statin <fctr>
```

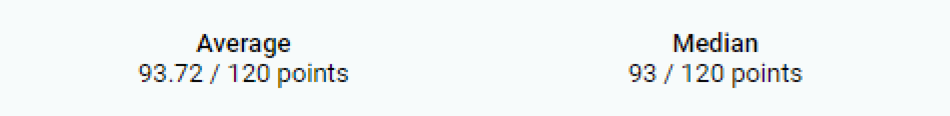
40.2 Results



Q40

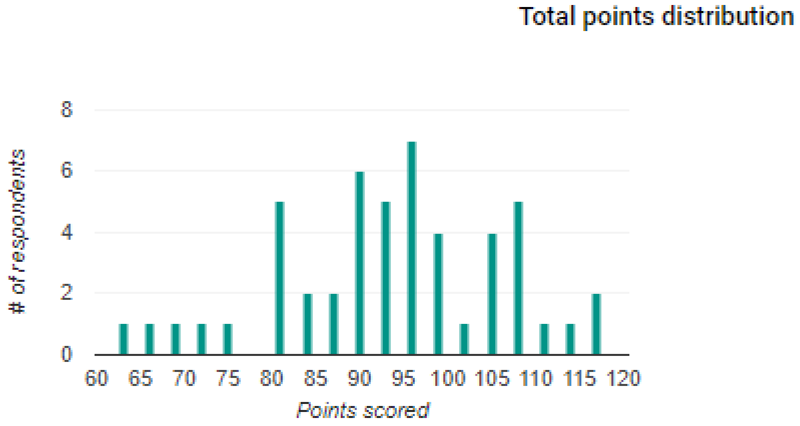
42 / 50 correct responses

41 Overall Achievement



Average
93.72 / 120 points

Median
93 / 120 points


Total points distribution
# of respondents
Points scored