432 Quiz 1 Answer Sketch for Fall 2019

Thomas E. Love

due 2019-03-05 at 7 AM. Version 2019-03-12 09:19:09

Grading

Each question is worth 2 points, except for questions 4, 17 and 18 (which are each worth 5 points) and question 8 (which is worth 3 points.)

The maximum possible grade on the quiz was 60 points. The highest score achieved was 55.5/60. The mean was 46 and the median score was 48.5.

Over the break, I will extend two honors opportunities:

1. Successful completion of the first opportunity will increase your score a bit on the quiz, and will be open to anyone who wants to take advantage of it, and in fact, I hope everyone does it.
2. The second opportunity will include the first but also something more extensive, will come with the potential for a larger increase in your quiz score, but will be restricted to those who scored below 50/60 on the quiz initially.

Further details to come.

1 Answer 1 is `c`

The correct answer is `c`. The `sex` variable did not change names, as a result of `clean_names()`, but every other variable did.

- `DBP0` become `dbp0`
- `Age` became `age`
- `Treatment` became `treatment`
- `subjectCode` became `subject_code`
- but `sex` remained `sex`

Incidentally, the `dbptrial` data come from Chen D and Peace KE (2011) *Clinical Trial Data Analysis Using R* Chapman & Hall, Chapter 3.

1.1 Grading 1: 2 points

- Partial Credit: 0.5 points if you listed multiple variables, one of which was `sex`.
- More than 32/37 people provided a correct response. With partial credit, 95.3% of available points were awarded.

2 Answer 2 is a line of code

My favorite option here is:

```
dbptrial <- dbptrial %>% mutate(dbp_diff = dbp0 - dbp4)
```

Other equally acceptable options include:

```
mutate(dbptrial, dbp_diff = dbp0 - dbp4)
dbptrial$dbp_diff <- dbptrial$dbp0 - dbptrial$dbp4
dbptrial['dbp_diff'] = dbptrial['dbp0'] - dbptrial['dbp4']
```

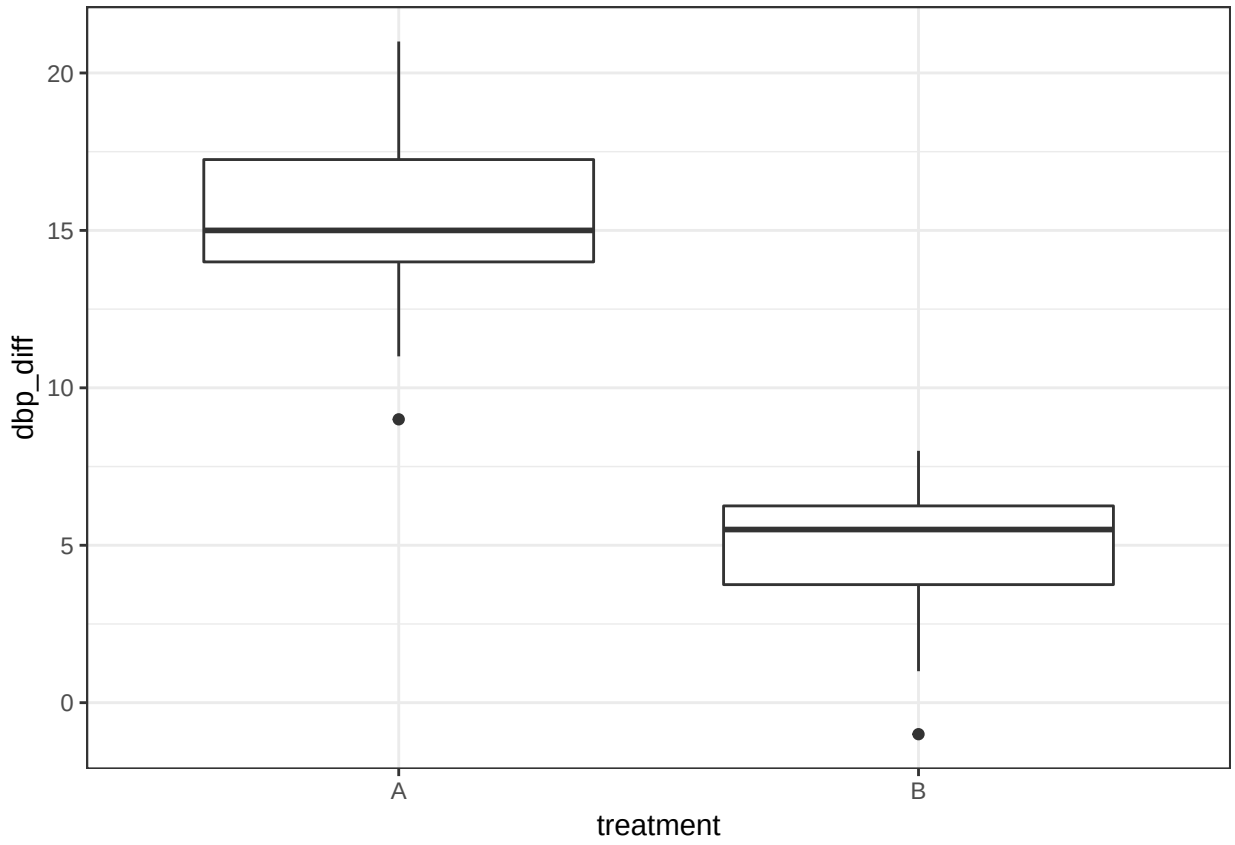
2.1 Grading 2: 2 points

- Partial Credit
  - Some people changed the name of the data, to something like `dbp_trial` or `dbptrial_m` instead of `dbptrial`, or changed the name of the new variable, to something like `dpb_diff`, and thus lost 0.5 point for each such problem.
  - If you created `dbp_diff` as `dbp4` minus `dbp0`, then you received 0.5 point.
  - A `mutate(dbp_diff = dbp0 - dbp4)` command all by itself, without placing it in the `dbp_trial` data set got you 0.5 point.
  - If you misunderstood the question, and created the `dbp_diff` you were supposed to create but then proceeded to assign the value "9 mm Hg More than dbp4" to anyone with a calculated value of 9, then you received 1.5 points.
- Some Notes:
  - In the future, please note that a single line of code means `dbptrial <- dbptrial %>% mutate(dbp_diff = (dbp0 - dbp4))` and not a lot of other extraneous stuff.
  - I didn't take points off for things like adding in a call to `dbptrial` after doing this work, but I probably should have, since that's clearly more than a single line of code.
  - Some people ran `clean_names()` in the middle of this, which was specified as unnecessary, but went unpenalized.
  - We didn't penalize it if your code worked, but we wanted you to use `dplyr` (tidyverse) to do this work.
- 24/37 people provided a correct response. With partial credit, 82.4% of available points were awarded.

3 Answer 3 is a AND e

Here's the boxplot I built.

```
ggplot(dbptrial, aes(x = treatment, y = dbp_diff)) +
    geom_boxplot() + theme_bw()
```



Statements `a` and `e` are true.

- The median change is larger in treatment A (around 15) than treatment B (around 6).
- The height of the boxes (describing the IQRs) are similar in the two groups.
- There is no overlap between the distributions of `dbp_diff` in the two groups. (The smallest A difference is still larger than the largest B difference)
- There is a single outlier candidate in each treatment group.

Note that if you had used `boxplot()` rather than `ggplot`, the outliers wouldn't show up. The two plotting approaches yield different results. That's why I specified the use of `ggplot`.

3.1 Grading 3: 2 points

- Partial Credit
  - If you chose `a` alone, or `e` alone, you got 1 point.
  - If you chose `a`, `e` but also `c`, you got 1 point.
- 29/37 people provided a correct response. With partial credit, 87.8% of available points were awarded.

4 Answer 4 is a sentence, including point estimate 10.40 and 95% CI of either 8.80 and 12.05 mm Hg or 8.85 and 12.10 mm Hg. (5 points)

The `bootdif` function compares B to A, so we need to take the negative of the endpoints provided. The correct response should include a point estimate of 10.40 mm Hg and a 95% confidence interval ranging from **(8.80, 12.05)** mm Hg, and a statement specifying the direction and significance of the result.

```
set.seed(2019)
round(bootdif(dbptrial$dbp_diff, dbptrial$treatment),2)
```

```
Mean Difference           0.025           0.975
         -10.40          -12.05           -8.80
```

But, as it turns out, there is another answer that is just as reasonable, and that is to obtain a 95% confidence interval from **(8.85, 12.10)** mm Hg.

```
dbptrial <- dbptrial %>%
    mutate(treat_A = treatment == "A",
           treat_B = treatment == "B")

set.seed(2019)
round(bootdif(dbptrial$dbp_diff, dbptrial$treat_A),2)
```

```
Mean Difference           0.025           0.975
          10.40            8.85           12.10
```

```
set.seed(2019)
round(bootdif(dbptrial$dbp_diff, dbptrial$treat_B),2)
```

```
Mean Difference           0.025           0.975
         -10.40          -12.05           -8.80
```

4.1 Grading 4: 5 points

A good response would tell us that:

- The point estimate was 10.40 mm Hg, and the 95% CI was **either** (8.80, 12.05) **or** (8.85, 12.10).
- Changes in DBP were statistically significantly larger for those receiving the new treatment (A) than in those receiving the placebo (B).

There were ways to do this that were more artful than others. Here are several of the cleaner student responses:

- The mean difference between the two treatment groups is 10.40 mm Hg with a 95% confidence interval (8.80, 12.05), which tells us that treatment A is significantly more effective at dropping DBP compared to placebo (treatment B) as our confidence interval does not cross 0.
- The mean decrease in diastolic blood pressure in those on the new drug is 10.40 mm Hg more than those on the placebo with a 95% confidence interval of (8.80 to 12.05).
- Patients receiving the new treatment (A) had on average 10.40 mm Hg greater change in diastolic blood pressure (95% CI 8.80, 12.05) during the trial than those receiving placebo, indicating that the change in DBP for patients receiving the new drug was statistically significantly greater than the change observed for those taking placebo.

- Using a bootstrap approach, the point estimate for the true difference in mean dbp_diff values between the new treatment group and the placebo group is 10.40 mm Hg (95% confidence interval = 8.80 to 12.05 mm Hg), where subjects receiving the new treatment experience a larger drop in dbp on average than the subjects in the placebo group.
- Partial Credit (I was overly generous to most people.)
  - If you had the wrong units, you lost a point.
    - Blood pressure is measured in millimeters of mercury, abbreviated mm Hg, and not mm / Hg, which would imply something like millimeters per mercury. Also, it's mm Hg, not mmHG.
  - If you didn't specify the units, you lost a point.
  - If you specified the direction of effect incorrectly (so that you had the placebo as the winner), you lost 3 points.
  - If you failed to clearly specify (with a word like "larger" or "higher") the direction of the effect, but had the correct result, you lost 1 point.
  - If you didn't specify either the point estimate or the confidence interval, you lost 1.5 points for each thing that was missing.
  - If you failed to specify the confidence level as 95% somehow, you lost a point.
  - If you claimed that you did a bootstrap test, rather than a confidence interval, you lost 1 point.
  - Also, the bootstrap approach has nothing to do with the point estimate, just the CI.
  - If your point estimate was wrong, you lost a point.
  - If either your lower or upper bound of the confidence interval was not one of the two options specified above, you lost 1 point for each.
  - If you didn't describe the result as a mean or average, you lost a point.

Some other problematic examples:

> Ninety-five percent of the time, the mean difference in diastolic blood pressure is 8.8 to 12.05 mmHg greater for subjects receiving treatment A than treatment B, with a point estimate difference of 10.4 mmHg for this bootstrap test.

> The point estimate is 10.40 mmHg with bounds of the confidence interval at 8.80 and 12.05 mmHg which tells us that in 95% of the trials/bootstraps, the confidence interval will contain the true value of the difference in diastolic blood pressure between treatments A and B (where the difference between trial groups is taken as treatment A minus treatment B).

> The point estimate is a difference of 10.40 and the 95% confidence interval is [8.80, 12.05] meaning that if you are to perform this same trial with 100 different samples, 95 of them will have a mean difference in the DBP values that will fall into the range of [8.80, 12.05].

None of these is an appropriate way to interpret a confidence interval. The "95% of the time" or "let's talk about 95 out of 100" seems to be what gets you in trouble here. Don't try to define all confidence intervals somehow when you explain the result in one of them.

Improved: The mean difference in DBP was 10.4 mm Hg, with a 95% CI of 8.8 to 12.05 mm Hg, indicating significantly greater changes for subjects receiving treatment A than treatment B.

> The population mean difference in dbp between treatment group A and B is 10.00, with the 95% bootstrap CI being (8.20, 11.78); this tells us that the true difference in dbp from baseline to visit 4 is ~10, with the true mean difference somewhere between 8.20 and 11.78.

In addition to some calculation issues and failing to explicitly describe the effect's direction, this specifies a population mean difference in the first phrase, which you don't know. You're describing a sample mean. Another trap people commonly fall into is trying to get the words "true difference" in there somehow. That doesn't really work.

Improved: The mean difference in DBP between treatment A and treatment B was 10.4 mm Hg, with a 95% CI of 8.8 to 12.05 mm Hg (or 8.85 - 12.10), indicating significantly greater changes for subjects receiving treatment A.

> At a 95% confidence interval (8.80, 12.05) which does not include 0, the new treatment (A) significantly lowers diastolic blood pressure by 10.40 mm Hg on average in comparison to the placebo (B).

Improved: At a 95% confidence level, treatment A significantly lowered DBP in comparison to the placebo (mean difference: 10.40 mm Hg, 95% CI 8.80, 12.05 or 8.85, 12.10).

- 15/37 people provided a correct response. With partial credit, 76.2% of available points were awarded.

5 Answer 5 is `b`

Statement `b` is true. The interaction is very modest here, as indicated by the essentially parallel lines joining the treatment means by sex. Of the four groups, females had larger mean changes (quite possibly not significantly larger, but certainly larger in the data set) in DBP for Treatment A and for Treatment B.

5.1 Grading 5: 2 points

- I awarded no partial credit here. The most common incorrect responses were `e` and `a`.
- 21/37 people provided a correct response. 56.8% of available points were awarded.

6 Answer 6 using the Main Effects only is 79.7% or 79.8%, depending on the rounding you did.

```
anova(lm(dbp_diff ~ treatment + sex, data = dbptrial))
```

```
Analysis of Variance Table

Response: dbp_diff
          Df  Sum Sq Mean Sq  F value    Pr(>F)
treatment  1 1081.60 1081.60 145.3001 2.231e-14 ***
sex        1    2.98    2.98   0.3997    0.5311
Residuals 37  275.42    7.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plot suggests that the interaction term would be very modest, so we fit a model without it. The η^2 value is the sum of squares accounted for by the `treatment` and `sex` factors, combined, divided by the total sum of squares (which adds in the residual sum of squares). That turns out to be 79.7%

```
(1081.60 + 2.98) / (1081.60 + 2.98 + 275.42)
```

```
[1] 0.7974853
```

If you'd instead fit a model including the interaction, your table would have been:

```
anova(lm(dbp_diff ~ treatment * sex, data = dbptrial))
```

```
Analysis of Variance Table

Response: dbp_diff
              Df  Sum Sq Mean Sq  F value    Pr(>F)
treatment      1 1081.60 1081.60 141.4927 4.958e-14 ***
sex            1    2.98    2.98   0.3893    0.5366
treatment:sex  1    0.23    0.23   0.0305    0.8624
Residuals     36  275.19    7.64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And the η^2 would have been 79.8%.

```
(1081.60 + 2.98 + 0.23) / (1081.60 + 2.98 + 0.23 + 275.19)
```

```
[1] 0.7976544
```

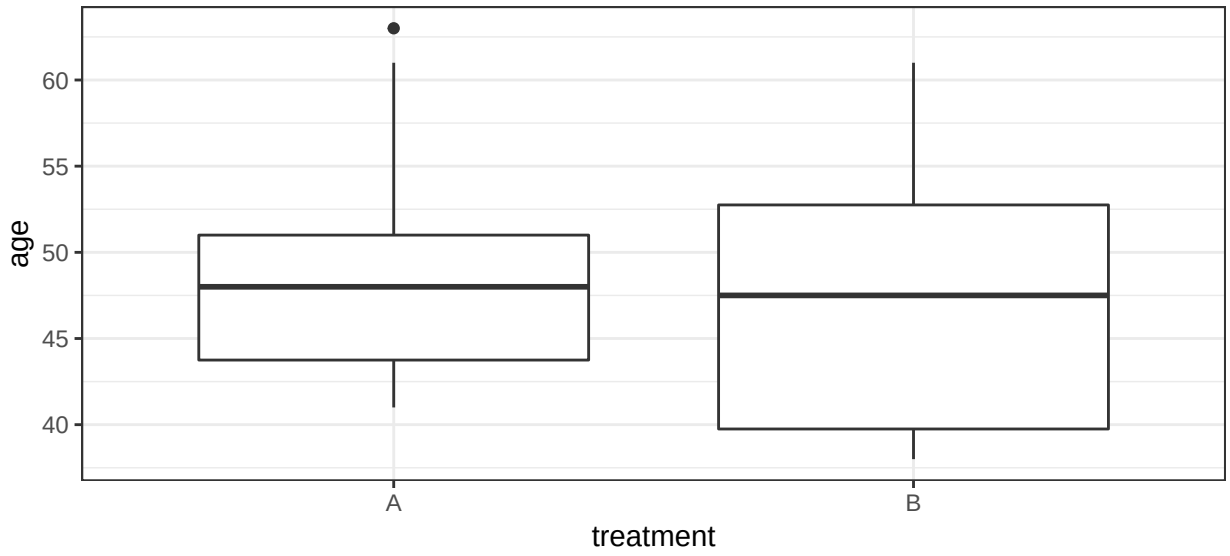
6.1 Grading 6: 2 points

- Partial Credit
  - I gave full credit for either 79.8% or 79.7%. I awarded no other partial credit.
- 29/37 people provided a correct response. With partial credit, 78.4% of available points were awarded.

7 Answer 7 is d

Yes it is consistent, because treatment shows no significant or substantial association with age. We'd expect in a RCT that treatment assignment would be unrelated to age or sex.

```
ggplot(dbptrial, aes(x = treatment, y = age)) +
    geom_boxplot() + theme_bw()
```



According to this boxplot, there is no substantial difference in the distribution of `age` across the two treatment groups. We could also consider doing a statistical test, like a t test.

```
t.test(age ~ treatment, data = dbptrial)
```

```
    Welch Two Sample t-test

data:  age by treatment
t = 0.73245, df = 36.336, p-value = 0.4686
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.740468  5.840468
sample estimates:
mean in group A mean in group B
          48.60           47.05
```

I guess you could also consider a bootstrap confidence interval for the difference in mean ages between the two treatment groups...

```
set.seed(2019)
bootdif(dbptrial$age, dbptrial$treatment)
```

```
Mean Difference           0.025           0.975
       -1.55000        -5.80125         2.40000
```

7.1 Grading 7: 2 points

- No partial credit. More than 32/37 were correct. 97.3% of available points were awarded.

8 Answer 8 is `c` (3 points)

The relevant p value would be the 0.6418, associated with the comparison of Model 1 (which includes the sex-treatment interaction) and Model 2 (which does not.) By this standard, the `STATEMENT` appears to be true. There is no significant sex-treatment interaction here, after adjusting for age.

The most common responses that were not `c` were `g` and `i`.

- Statement `g` suggests comparing Model 2 to Model 3. Model 2 contains an interaction between age and treatment and one between age and sex, while Model 3 contains no interactions. That's not the comparison we're looking for, which is focused on whether the treatment-sex interaction is important.
- It is certainly possible to tell what to conclude about the STATEMENT, so `i` isn't correct, but what worries me about this item is that people were perhaps thinking that the STATEMENT refers to `gender` and the variable in `dbptrial` refers to the patient's `sex` instead. Those are different things, and I shouldn't have been so sloppy about it in writing the question.

8.1 Grading 8: 3 points

- Because of my sloppiness with `gender` and `sex`, I decided that responses `c` (my intended response) and `i` (the correct response when you read `gender` as being importantly different from `sex` here) were each appropriate. Otherwise, there was no partial credit.
- 10/37 people chose `c`, and an additional 15/37 chose `i`, so 25/37 provided a correct response. 67.6% of available points were awarded.

9 Answer 9 is `a` and `c`

`a` and `c` are quantitative, (`c` is continuous within an interval). But b and d are categorical.

Everyone got `a` right, but there was some confusion on the others.

9.1 Grading 9: 2 points

- Partial Credit
  - Here, you were making four binary choices. I gave you 0.5 point for each one you got right.
  - So if you chose `a` and `b`, for example, you made 2 correct decisions (`a` and `d`) so you got one point.
  - If you chose `a` only, then you made 3 correct decisions (on `a`, `b` and `d`) for 1.5 points.
- 18/37 people provided a correct response. With partial credit, 86.5% of available points were awarded.

10 Answer 10 is `e`

The Box-Cox plot suggests a power of 0.5 for the model predicting y using x. This means we transform the y variable by taking it to the power 0.5, i.e. we take the square root of y. That's `e`.

- Another option here would have been `lm(y ~ x^2)` if I'd made that option available to you.

10.1 Grading 10: 2 points

- There was no partial credit here. The most common incorrect response was, unsurprisingly, `c`. It's the `y` that Box-Cox is transforming, not `x`.
- 31/37 people provided a correct response. 83.8% of available points were awarded.

11 Answer 11 is `c`, `d` and `e`

The data come from Riffenburgh RH *Statistics in Medicine* Second Edition, DB1.

```
dd <- datadist(riff1)
options(datadist = "dd")

model_11 <- lrm(biopsy ~ age + dre + tru + vol + psa,
                data = riff1, x = TRUE, y = TRUE)

model_11
```

```
Logistic Regression Model

 lrm(formula = biopsy ~ age + dre + tru + vol + psa, data = riff1,
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           301    LR chi2      56.81    R2       0.241    C       0.738
  0            206    d.f.             5    g        1.380    Dxy     0.477
  1             95    Pr(> chi2) <0.0001    gr       3.976    gamma   0.477
 max |deriv| 5e-09                          gp       0.203    tau-a   0.207
                                            Brier    0.177

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -0.7839 1.1310 -0.69  0.4882
 age       -0.0045 0.0171 -0.26  0.7943
 dre        0.2863 0.2975  0.96  0.3358
 tru        0.6879 0.2795  2.46  0.0138
 vol       -0.0295 0.0095 -3.10  0.0019
 psa        0.1063 0.0260  4.10  <0.0001
```

The predictors that show significant effects are those with p values in the Wald tests below 0.05. That's `tru`, `vol` and `psa`, which corresponds to choices `c`, `d` and `e`.

11.1 Grading 11: 2 points

- There was no partial credit here. There were scattered incorrect responses.
- 32/37 people provided a correct response. 86.5% of available points were awarded.

12 Answer 12 is c

Three predictors remain. You'll have to use `glm` to fit a logistic regression here.

```
model_12 <- glm(biopsy ~ age + dre + tru + vol + psa,
                data = riff1, family = binomial())
step(model_12)
```

```
Start:  AIC=330.55
biopsy ~ age + dre + tru + vol + psa

       Df Deviance    AIC
- age   1   318.62 328.62
- dre   1   319.48 329.48
<none>      318.55 330.55
- tru   1   324.66 334.66
- vol   1   329.64 339.64
- psa   1   351.40 361.40

Step:  AIC=328.62
biopsy ~ dre + tru + vol + psa

       Df Deviance    AIC
- dre   1   319.50 327.50
<none>      318.62 328.62
- tru   1   324.67 332.67
- vol   1   330.08 338.08
- psa   1   351.48 359.48

Step:  AIC=327.5
biopsy ~ tru + vol + psa

       Df Deviance    AIC
<none>      319.50 327.50
- tru   1   326.32 332.32
- vol   1   333.27 339.27
- psa   1   353.19 359.19


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

Coefficients:
(Intercept)          tru          vol          psa
   -0.84241      0.71667     -0.03149      0.10558

Degrees of Freedom: 300 Total (i.e. Null);  297 Residual
Null Deviance:      375.4
Residual Deviance: 319.5    AIC: 327.5
```

12.1 Grading 12: 2 points

- No partial credit. Those who got it wrong picked `b`. More than 32/37 were correct. 94.6% of available points were awarded.

13 Answer 13 is c

And now, we'll refit this smaller model with `lrm` to get the C statistic.

```
model_13 <- lrm(biopsy ~ tru + vol + psa,
                data = riff1, x = TRUE, y = TRUE)

model_13
```

```
Logistic Regression Model

 lrm(formula = biopsy ~ tru + vol + psa, data = riff1, x = TRUE,
     y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           301    LR chi2      55.86    R2        0.238    C         0.742
  0            206    d.f.             3    g         1.364    Dxy       0.484
  1             95    Pr(> chi2) <0.0001    gr        3.912    gamma     0.484
 max |deriv| 2e-09                          gp        0.201    tau-a     0.210
                                            Brier     0.177

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -0.8424 0.3576 -2.36  0.0185
 tru        0.7167 0.2760  2.60  0.0094
 vol       -0.0315 0.0092 -3.41  0.0007
 psa        0.1056 0.0256  4.12  <0.0001
```

The model's C statistic is 0.742, it appears. Note that the Nagelkerke R^2 is 0.238.

13.1 Grading 13: 2 points

- Everyone got it right. Yay!

14 Answer 14 is b

The C statistic indicates poor discrimination. The R^2 is only modest, but certainly not zero.

- I'd clarified the way to interpret "poor" discrimination in a note to everyone, so this should have been a bit more straightforward than it was.
- Similarly, a Nagelkerke R^2 of 0.238 isn't indicative of "essentially no predictive value".

14.1 Grading 14: 2 points

- There was no partial credit awarded here. The most common incorrect response was `c`, but some people chose other things as well.
- 28/37 people provided a correct response. 75.7% of available points were awarded.

15 Answer 15 is 0.27

There are two ways to get where we're going in R. We can use the `lrm` model:

```
predict(model_13, newdata = tibble(psa = 9.6, vol = 60, tru = 1), type = "fitted")
```

```
        1
0.2686113
```

Or we can build the model as a `glm` model, in which case we would have:

```
model_16 <- glm(biopsy ~ tru + vol + psa,
                data = riff1, family = binomial())

predict(model_16, newdata = tibble(psa = 9.6, vol = 60, tru = 1), type = "response")
```

```
        1
0.2686113
```

15.1 Grading 15: 2 points

- Partial Credit
  - I gave 1.5 points to people who gave answers that were insufficiently rounded, like 0.269, but that if rounded properly would have been correct.
  - I gave 1 point to people who rounded too much and gave 0.3 as an answer, because I was less sure they'd done things correctly.
- 25/37 people provided a correct response. With partial credit, 77.7% of available points were awarded.

16 Answer 16 is `0.714`.

The index-corrected estimate of Dxy is 0.4273, but we want the index-corrected estimate of the C statistic. Remembering that C = 0.5 + Dxy/2, we have C = 0.5 + (0.4273/2) = 0.71365, and after rounding to three decimal places, that's 0.714.

16.1 Grading 16: 2 points

- Partial Credit
  - I gave 1.5 points to people who rounded incorrectly to 0.713.
  - Several people gave incorrect responses of 0.738 or something close to that, which would be what the original data gives (from `index.orig`) but not the validated result in `index.corrected`.
- 25/37 people provided a correct response. With partial credit, 71.6% of available points were awarded.

17 Answer 17 is a sentence - the odds ratio estimate is 1.08 or 1.0845. (5 points)

The odds ratio is exp(0.08113) or 1.0845119 = 1.0845.

Two reasonable ways to do this that I thought of in advance were:

- If Harry's `psa` is 1 ng/ml larger than Chuck's but Harry and Chuck have the same result on the transurethral ultrasound, then the estimated odds of a positive biopsy are 8.45% larger for Harry than they are for Chuck.
- If Subject A's `psa` is 1 ng/ml larger than Subject B's and A and B have the same `tru` result, the model's estimated odds of a positive biopsy are 1.0845 times as large for A as for B.

17.1 Grading 17: 5 points

Prior to grading, I worried that this would be a bloodbath. And it was. My quest over the next week will be to figure out a way to teach this (and drill this, if needed) so that when confronted 1, 10 or 100 months from now with a logistic regression model, you will interpret the effects correctly.

For full credit, your response needed to not say anything untrue, and also accomplish these five things very clearly:

1. correctly exponentiate the coefficient of `psa`, (*I penalized more heavily if you blew this.*)
2. specify the impact of a change in `psa` (specifying the units as ng/ml)
3. specify the impact of that change in terms of a change in the estimated **odds** (not risk, or rate, or probability, or percentage chance or anything like that)
4. specify the impact on the odds of a positive biopsy (thus specifying the **outcome**)
5. specify that this was only true for making a comparison while holding the transurethral ultrasound value at a consistent level. (*This - the ceteris paribus requirement - was the thing most often missed.*)

Partial Credit was awarded to people who did some of the things above correctly, but not all of them. Most people in that situation got 2 or 3 points out of 5.

- Some people lost points for including something incorrect along with generally correct stuff.
- Many more people lost points for not doing one or more of the things above clearly enough.
- Some people talked about significance, but not the direction or size of the effect. That's not responsive to the question.
- Some folks didn't bother to explain the odds ratio value or percentage value in context, just stated what it was. Also not sufficient.
- Not catching the ratio part - writing, for instance, that the odds are 1.08 greater rather than 1.08 times greater or 8% greater, was a common problem.
- Another common problem was saying that person A was 1.08 times more likely to have a positive result, which isn't right. The odds are 1.08 times higher, not the probability or likelihood. They're not interchangeable.
- A small point - it is an **odds** ratio, not an *odd* ratio.
- Just 3/37 people provided a response I was satisfied with, and awarded 5 points. With partial credit, 50.3% of available points were awarded.

This is hard, and it didn't go very well. The three student responses that worked for me were:

- An increase of 1 ng/ml of prostate-specific antigen level represents increased odds of a positive cancer biopsy by 8.45% all other variables being constant.
- Holding `tru` at a fixed value, we will see 8.45% increase in the odds of getting positive `biopsy` for a 1 ng/ml increase in `psa` since exp(0.08113) = 1.0845.
- Holding all else constant, for every 1 ng/ml increase in PSA level, the odds of having a positive result in the biopsy statistically significantly increases by 8 percent.

18 Answer 18 is a sentence - the calculation is (1.24, 3.64). (5 points)

The 95% confidence interval is calculated as `exp(0.75402 - 2 x 0.26941), exp(0.75402 + 2 x 0.26941)` or 1.2401099 to 3.6431183, which rounds to (1.24, 3.64). Since this interval is entirely above 1, it means that if two people have the same `psa`, but only one has a positive transurethral ultrasound, then that person has significantly higher odds of a positive biopsy.

I didn't ask you to specify the point estimate (which is `exp(0.75402)`, or 2.13) for the odds ratio for `tru`, but some people did.

18.1 Grading 18: 5 points

For full credit, your response needed to not say anything untrue, and also accomplish these seven things very clearly:

1. correctly calculate the odds ratio's confidence interval
2. specify the impact of a change in `tru` (specifying the units as ng/ml)
3. specify the impact of that change in terms of a change in the estimated **odds** (not risk, or rate, or probability, or percentage chance or anything like that)
4. specify the impact on the odds of a positive biopsy (thus specifying the **outcome**)
5. specify that this was only true for making a comparison while holding the PSA value at a consistent level,
6. specify that the effect was statistically significant, and
7. specify the direction of the effect, specifically that a positive transurethral ultrasound was associated with **higher** odds of a positive biopsy.

That was a lot to do and no one succeeded to my satisfaction, but it's incredibly fundamental to understanding what the model is telling you.

- Partial Credit was generally not awarded here to people who missed the confidence interval badly. People who gave a confidence interval without converting to an odds ratio had to be very specific about what that result actually meant, and most people didn't succeed in that regard.
- Otherwise, partial credit worked in a similar fashion to question 17, in that most people who failed to do one or two of the things above received 1-3 of the 5 available points.
- Again, the most common problem was forgetting to say something about *ceteris paribus* (with other conditions remaining the same, or "when all else is equal") - specifically, you needed to say something about `psa` being constant across the imaginary subjects you were comparing. Most of the people who missed that got 3/5.
- Another common problem was specifying that there was a significant difference but not its direction. That usually got you at most 2/5.
- 0/37 people provided a completely correct response. With partial credit, 39.5% of available points were awarded.

19 Answer 19 is `a`

We're looking at variance inflation factors, and those help us assess collinearity. None of the VIFs are large enough to give us too much concern here, as they are all well below 5.

19.1 Grading 19: 2 points

- There was no partial credit available.
- More than 32/37 people provided a correct response. 94.6% of available points were awarded.

20 Answer 20 is `h`

This plot can be useful in several ways, but of this list of options, the only one is that it is a check of the influence of each observation on our model. None of the points appear highly influential.

20.1 Grading 20: 2 points

- Everybody got this right. Great!

21 Answer 21 is `f`

This is a plot to assess the calibration of the model - how well the predicted values obtained by the model match the observed (true) values of our outcome.

21.1 Grading 21: 2 points

- There was no partial credit available.
- More than 32/37 people provided a correct response. 97.3% of available points were awarded.

22 Answer 22 is `d` (first) and `a` (second)

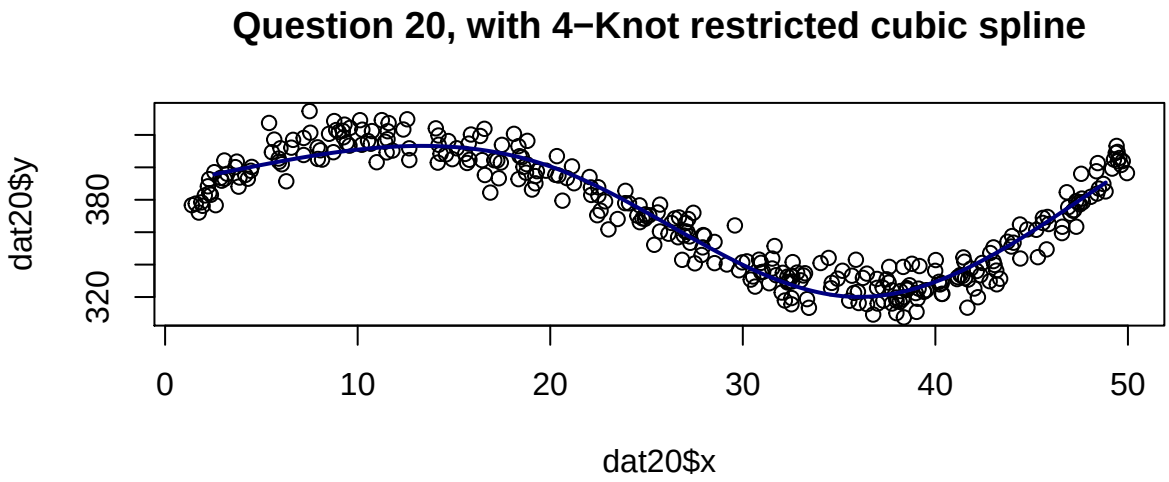
The most important non-linear term to consider would involve the `height` variable, which is quantitative, so a restricted cubic spline in `height` is the best option. The next most important thing to do is include a restricted cubic spline in `age`, since that's next up, and is also quantitative.

22.1 Grading 22: 2 points

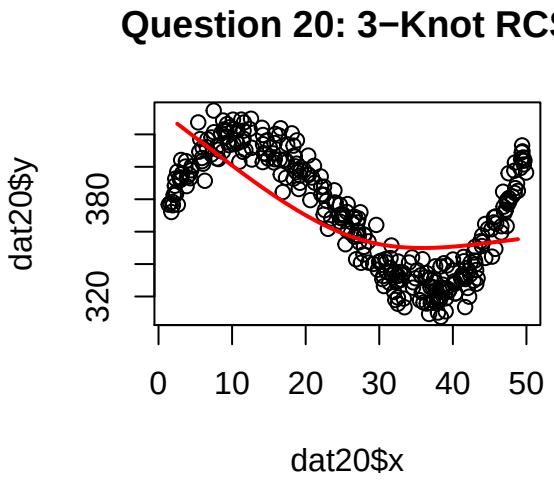
- You got one point for each correct answer.
- If you had something other than `d` in the first option, but used `d` as your second option, you got 1 point.
- More than 32/37 people provided a correct response. 97.3% of available points were awarded.

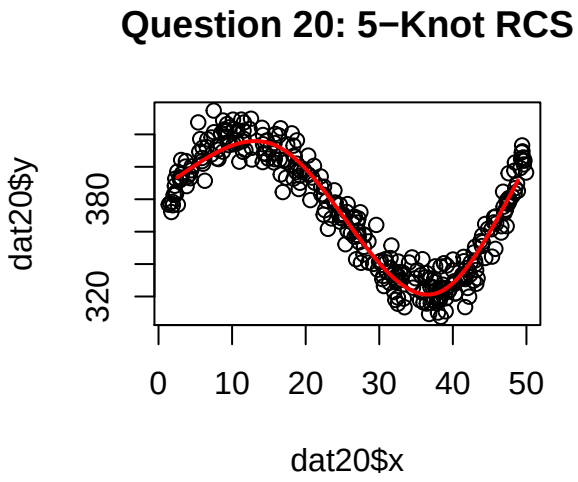
23 Answer 23 is c.

We need two "bends" to match the data well, which requires three degrees of freedom, so four knots.



The four-knot RCS is both necessary and sufficient here - there's no meaningful gain by adding a fifth knot, nor is a model with just three knots sufficient.





23.1 Grading 23: 2 points

- There was no partial credit available.
- 30/37 people provided a correct response. 81.1% of available points were awarded.

24 Answer 24 is that adjusted R^2 and bias-corrected AIC recommend `c, d, e, f` and that Cp and BIC recommend `c, d`

- Adjusted R^2 and bias-corrected AIC each recommend the model with 5 inputs (intercept, `c`, `d`, `e`, and `f`)
- Cp and BIC recommend the model with 3 inputs (intercept, `c` and `d`)

24.1 Grading 24: 2 points

- Partial Credit: You got 0.5 point for specifying the correct model for a statistic. So if you got the right choice 3 times out of the 4 statistics, you'd have received 1.5 points.
  - 28 people got the right model for adjusted R^2
  - 27 got the right model for bias-corrected AIC
  - 29 got the right model for BIC
  - only 19 got the right model for Mallows' Cp, with several people picking the model with all predictors, having apparently forgotten that Cp will always be equal to k when the entire model is fit. That's how Cp works.
- 18/37 people provided a correct response for all 4 statistics. With partial credit, 69.6% of available points were awarded.

25 Answer 25 is `c, d, e, f`

10-fold cross-validation of `q24_m3`

```
data24 <- readRDS("data/data24.Rds")

set.seed(2019)

cv_q24_m3 <- data24 %>%
  crossv_kfold(k = 10) %>%
  mutate(model = map(train, ~ lm(outcome ~ c + d,
                                 data = .)))

cv_q24_m3_pred <- cv_q24_m3 %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

cv_q24_m3_results <- cv_q24_m3_pred %>%
    summarise(Model = "q24_m3 (c and d)",
              RMSE = sqrt(mean((outcome - .fitted) ^2)),
              MAE = mean(abs(outcome - .fitted)))
```

10-fold cross-validation of `q24_m5`

```
set.seed(2019)

cv_q24_m5 <- data24 %>%
  crossv_kfold(k = 10) %>%
  mutate(model = map(train, ~ lm(outcome ~ c + d + e + f,
                                 data = .)))
```

```
cv_q24_m5_pred <- cv_q24_m5 %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

cv_q24_m5_results <- cv_q24_m5_pred %>%
    summarise(Model = "q24_m5 (c, d, e, f)",
              RMSE = sqrt(mean((outcome - .fitted) ^2)),
              MAE = mean(abs(outcome - .fitted)))
```

```
bind_rows(cv_q24_m3_results, cv_q24_m5_results)
```

```
# A tibble: 2 x 3
  Model                RMSE   MAE
  <chr>               <dbl> <dbl>
1 q24_m3 (c and d)     5.11  4.15
2 q24_m5 (c, d, e, f)  5.10  4.12
```

So the larger model, with `c`, `d`, `e` and `f` is the winner.

25.1 Grading 25: 2 points

- If your candidate models from Question 24, didn't include `cdef`, but did include `cdf`, I gave a point here if you picked `cdf`.
- If your candidate models from Question 24, didn't include `cdef` or `cdf`, but did include `cd` or `cde`, I gave a point here if you picked `cd` or `cde`.
- 21/37 people provided a correct response. With partial credit, 77.0% of available points were awarded.