432 Quiz 1 and Answer Sketch

Thomas E. Love

Due 2018-03-05 at Noon. Version: 2018-03-05

Contents

0.1 Setup

```
knitr::opts_chunk$set(comment=NA)

library(rms)
library(skimr)
library(broom)
library(tidyverse)
```

1 Question 1. (4 points)

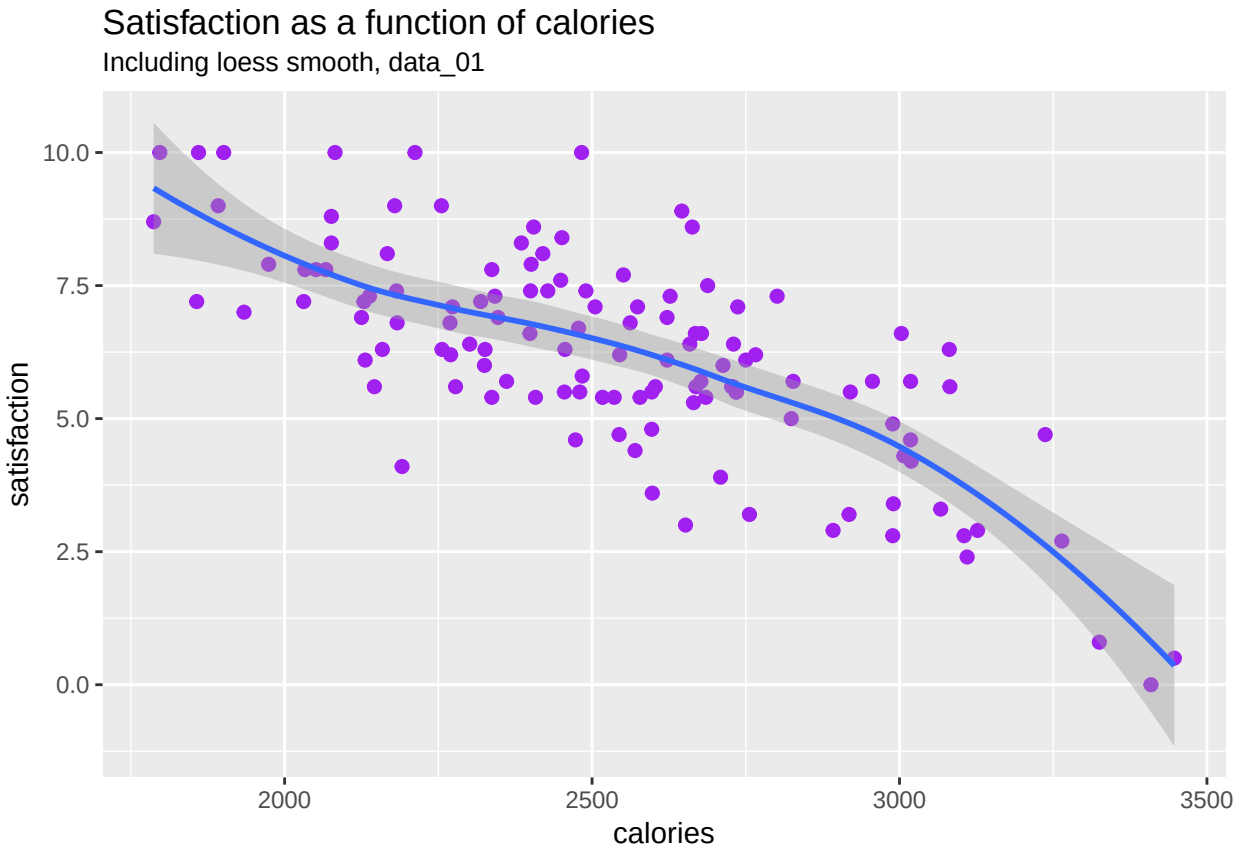
1.1 Setup for Question 1

```
set.seed(43201)

calories <- rnorm(125, mean = 2500, sd = 350)
sat_1 <- 5*(calories/1000) - 4*(calories/1000)^2
err <- rnorm(125, mean = 0, sd = 6)
sat_2 <- 4*((sat_1 + err) - min(sat_1 + err))/(2*sd(sat_1 + err))
satisfaction <- ifelse(sat_2 < 10, sat_2, 10)

data_01 <- data_frame(subject = 1:125,
                      calories = round(calories,0),
                      satisfaction = round(satisfaction,1))

write_csv(data_01, "data/data_01.csv")

rm(calories, sat_1, err, sat_2, satisfaction)
```

1.2 Question 1 Code Attempt

```
ggplot(data_01, aes(x = calories, y = satisfaction)) +
    geom_smooth(method = "loess") +
    labs(title = "Satisfaction as a function of calories",
         subtitle = "Including loess smooth, data_01")
```

1.3 Question 1 Target Plot




```
Saving 6.5 x 4.5 in image
```

The `data_01.csv` data set available to you on our web site will be used for Questions 1-5. Using the `data_01.csv` data set, a student attempted unsuccessfully to generate the Q01 Target Plot shown above, in R, developing the code shown in the Q01 Code Attempt above. Specify how you would FIX the code above to generate the Q01 Target Plot. Note that the points in the Target Plot are not only purple, but double their default size.

1.4 Answer 1 is to add `geom_point(size = 2, col = "purple") +` before the `geom_smooth` line

My final code is repeated below.

```
ggplot(data_01, aes(x = calories, y = satisfaction)) +
    geom_point(size = 2, col = "purple") +
    geom_smooth(method = "loess") +
    labs(title = "Satisfaction as a function of calories",
         subtitle = "Including loess smooth, data_01")
```

1.5 Q01 Results

- 32/41 students got full credit.
- 90% of available points were awarded.

- Some partial credit was awarded for people who guessed incorrectly as to the size of the points, or who tried to jitter the points or do something else that wasn't appropriate.

2 Question 2. (4 points)

Using the `data_01.csv` data set, specify the code required to fit (using `lm`) a model called `m02` that predicts the `satisfaction` score across these subjects using an orthogonal polynomial of degree 3 in the `calories` variable. You can assume that the data have been uploaded properly, and that the tidyverse is loaded, as well.

2.1 Answer 2 is the following code. . .

```
m02 <- lm(satisfaction ~ poly(calories,3), data = data_01)
```

2.2 Q02 Results

- At least 38/41 students got full credit.
- More than 95% of available points were awarded.
- Some partial credit was awarded for people who did something more than what was called for, or who mislabeled the results, or who fit a raw polynomial instead of an orthogonal one.

3 Question 3. (2 points)

Summarize the `m02` model you built in Question 2. Which of the following ranges contains the R^2 value for this model?

a. 0 to 0.39
b. 0.4 to 0.59
c. 0.6 to 0.79
d. 0.8 to 1.0

3.1 Answer 3 is b

```
summary(m02)
```

```
Call:
lm(formula = satisfaction ~ poly(calories, 3), data = data_01)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1712 -0.8313 -0.0905  0.7816  3.5178

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.1944     0.1152  53.771   <2e-16 ***
poly(calories, 3)1 -16.0019     1.2880 -12.424   <2e-16 ***
poly(calories, 3)2  -3.0389     1.2880  -2.359   0.0199 *
poly(calories, 3)3  -3.2660     1.2880  -2.536   0.0125 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.288 on 121 degrees of freedom
Multiple R-squared:  0.5789,    Adjusted R-squared:  0.5685
F-statistic: 55.45 on 3 and 121 DF,  p-value: < 2.2e-16
```

The R^2 value is 0.58.

3.2 Q03 Results

- At least 38/41 students got full credit.
- More than 95% of available points were awarded.
- No partial credit was available.

4 Question 4. (3 points)

A new model in R (which I'll call `m04`) was fit to the `data_01` data, now using an orthogonal polynomial of degree 2. The `glance` function applied to `m04` shows an AIC of 428.4 and a BIC of 439.7. Compare these results to `m02`. Which of the following conclusions is most appropriate based on these results?

a. The cubic term in Model `m02` is not helpful according to either AIC or BIC.
b. The cubic term in Model `m02` is helpful according to exactly one of AIC or BIC.
c. The cubic term in Model `m02` is helpful according to both AIC and BIC.
d. None of these conclusions are appropriate.

4.1 Answer 4 is c

Both AIC and BIC are smaller for `m02` with the cubic term than for `m04` without that term. So they both favor Model `m02`.

```
m04 <- lm(satisfaction ~ poly(calories,2), data = data_01)

glance(m04)
```

```
  r.squared adj.r.squared   sigma statistic      p.value df   logLik
1 0.5565444     0.5492746 1.31632  76.55605 2.869842e-22  3 -210.204
      AIC      BIC deviance df.residual
1 428.408 439.7212 211.3891         122
```

```
glance(m02)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.5789213     0.5684813 1.287968  55.45239 1.270124e-22  4 -206.9679
       AIC      BIC deviance df.residual
1 423.9358 438.0773 200.7224         121
```

4.2 Q04 Results

- 37/41 students got full credit.
- 90% of available points were awarded.
- No partial credit was available.

5 Question 5. (3 points)

How many of the subjects in `data_01` have both `calories` above 2500 and `satisfaction` below 5?

5.1 Answer 5 is 24.

The code I used was:

```
data_01 %>% count(calories > 2500, satisfaction < 5)
```

```
# A tibble: 4 x 3
  `calories > 2500` `satisfaction < 5`     n
  <lgl>             <lgl>              <int>
1 F                 F                     58
2 F                 T                      2
3 T                 F                     41
4 T                 T                     24
```

5.2 Q05 Results

- At least 38/41 students got full credit.
- More than 95% of available points were awarded.
- No partial credit was available.

6 Question 6. (3 points)

Suppose we want to use the `geom_label` approach in R to annotate the points in a scatterplot that we're building for an audience beyond ourselves. But two of the points we want to annotate are very close to each other in the plot, so that the labels will likely overlap. There is a package in R that helps you adjust those labels automatically, and it was mentioned and demonstrated in *R for Data Science*. What is the name of that package?

6.1 Answer 6 is `ggrepel`.

If you need a reference, look at section 28.3 of *R for Data Science*.

6.2 Q06 Results

- At least 38/41 students got full credit.
- More than 95% of available points were awarded.
- No partial credit was available.

7 Question 7. (4 points, 2 for part a, 2 for part b)

7.1 Setup for Question 7

```
set.seed(43207)

death <- as.integer(c(rep(1,57), rep(0, 231)))
pitt <- c(rnorm(57, mean = 9.79, sd = 2.5), rnorm(231, mean = 4.84, sd = 2.5))
surgery <- c(rep(1,30), rep(0, 27), rep(1, 170), rep(0, 61))

data_07 <- data_frame(subject = 1:288,
                      death, pss = round(pitt, 1), surgery)

rm(death, pitt, surgery)
```

This question and Q08 are about a retrospective study of 288 patients with esophageal perforation, 57 of whom died during the study follow-up period. The logistic regression model shown above (in the Model for Q07 section) was fit to describe the probability that a patient with esophageal perforation would die (`death` = 1 if died, `death` = 0 if alive) based on their Pittsburgh severity score (`pss`) and whether or not they underwent surgery (`surgery` = 1 if the subject had surgery and 0 if they did not.) Which of the following statements describe this output accurately?

```
m07 <- glm(death ~ pss * surgery,
           family = binomial, data = data_07)

tidy(m07, exponentiate = TRUE, conf.int = TRUE)
```

```
         term  estimate std.error  statistic      p.value     conf.low
1 (Intercept) 0.0046888 1.0631867 -5.0438729 4.562024e-07 0.0004308193
2         pss 1.8774696 0.1319583  4.7736678 1.809007e-06 1.4924048171
3     surgery 0.1531391 1.5075931 -1.2446387 2.132646e-01 0.0075667825
4 pss:surgery 1.1507435 0.1848442  0.7596033 4.474917e-01 0.7923185831
   conf.high
1 0.02944241
2 2.52235475
3 3.17133010
4 1.65872976
```

Rows:

a. At the 5% significance level, the interaction term adds statistically significant value for predicting death.
b. If two patients each have surgery, the model suggests that the patient with the higher PSS will have lower odds of death.

Columns:

1. TRUE
2. FALSE
3. It is impossible to tell from the information provided.

7.2 Answer 7 is `a` is FALSE and `b` is FALSE, too.

- The confidence interval for the odds ratio associated with the interaction term is (0.79, 1.66), which contains 1, so there's no statistically significant interaction effect at the 5% level.

- If two patients each have surgery, the estimated odds of death will be higher, not lower, for the patient with a higher value of PSS, because both the `pss` main effect and the `pss*surgery` interaction terms are positive.
- Note that these data are simulated, but this question is based on the abstract by Michael Schweigart and others found at http://www.jtcvsonline.org/article/S0022-5223(15)02382-X/fulltext

7.3 Q07a Results

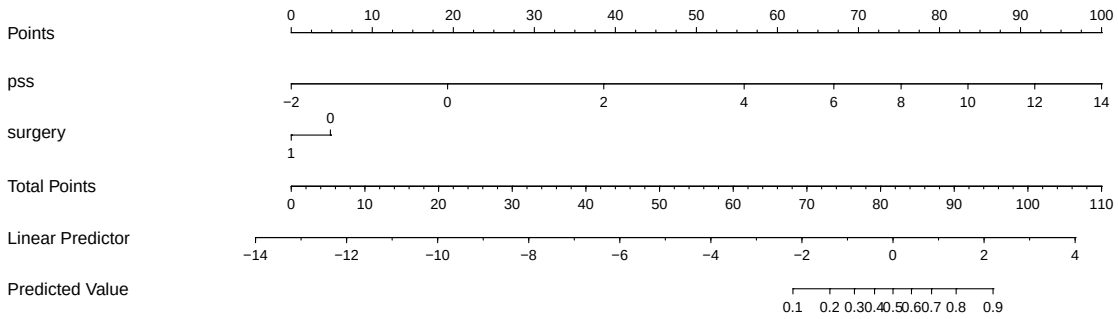
- 37/41 students got full credit.
- 90% of available points were awarded.
- No partial credit was available.

7.4 Q07b Results

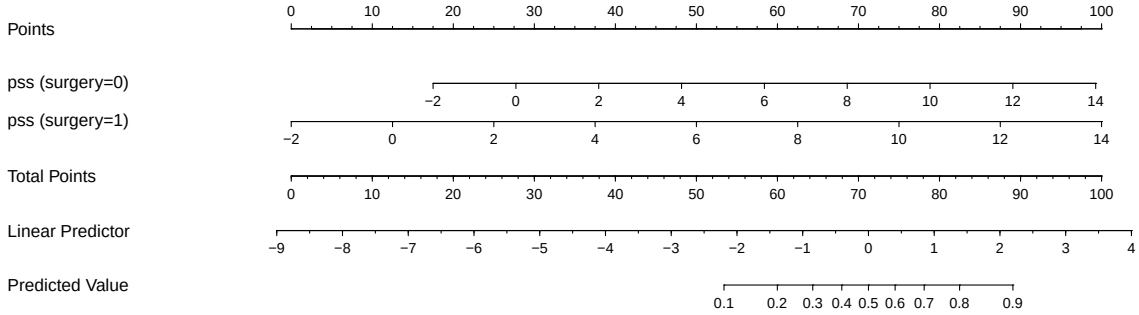
- 38/41 students got full credit.
- 93% of available points were awarded.
- No partial credit was available.

8 Question 8. (4 points, 2 for part a, 2 for part b)

8.1 Nomogram A for Question 8



8.2 Nomogram B for Question 8



An investigator produced two different nomograms using the data in `data_07`, and those nomograms are shown above as Nomogram A and Nomogram B for Question 8. If the images above are hard to see, please note that the two nomograms for Question 8 are also available as PDF files on the Quiz 1 section of our web site. Please answer the two questions posed below, which are: Which of the nomograms shown applies to the model we fit in question 7, and what is the predicted probability of death from that model for a subject named Sam who had surgery and whose Pittsburgh severity score was 8?

Rows:

a. Which nomogram (A or B) accurately describes the model we fit in Question 7?
b. Sam's predicted probability of death is ...

Columns:

1. A
2. B
3. Unknown.

4. less than 0.2
5. between 0.2 and 0.39
6. between 0.4 and 0.59
7. 0.6 or greater

8.3 Answer 8 is `a` goes with 2 (B) and `b` goes with 5 (between 0.2 and 0.39).

Since there's an interaction, it has to be Nomogram B, as Nomogram A doesn't include an interaction term. Using Nomogram B, we get a predicted value between 0.2 and 0.3. Actually, we'd get that same probability (approximately) with either nomogram.

8.4 Q08a Results

- At least 38/41 students got full credit.
- Over 95% of available points were awarded.
- No partial credit was available.

8.5 Q08b Results

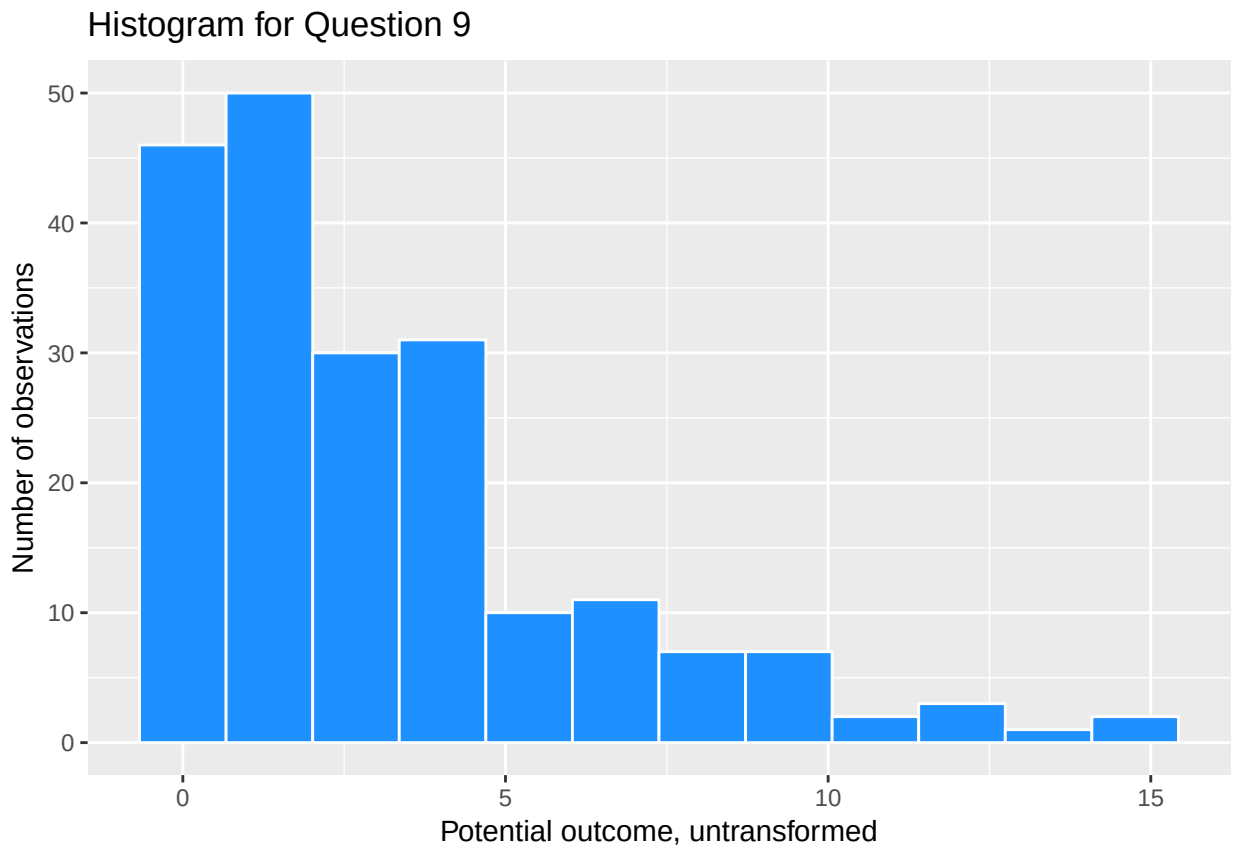
- At least 38/41 students got full credit.
- Over 95% of available points were awarded.
- No partial credit was available.

9 Question 9. (3 points)

9.1 Setup for Question 9

```
set.seed(43209)
data_09 <- data_frame(subject = 1:200,
                      outcome = 0.2 + rchisq(200, df = 1, ncp = 2))
```

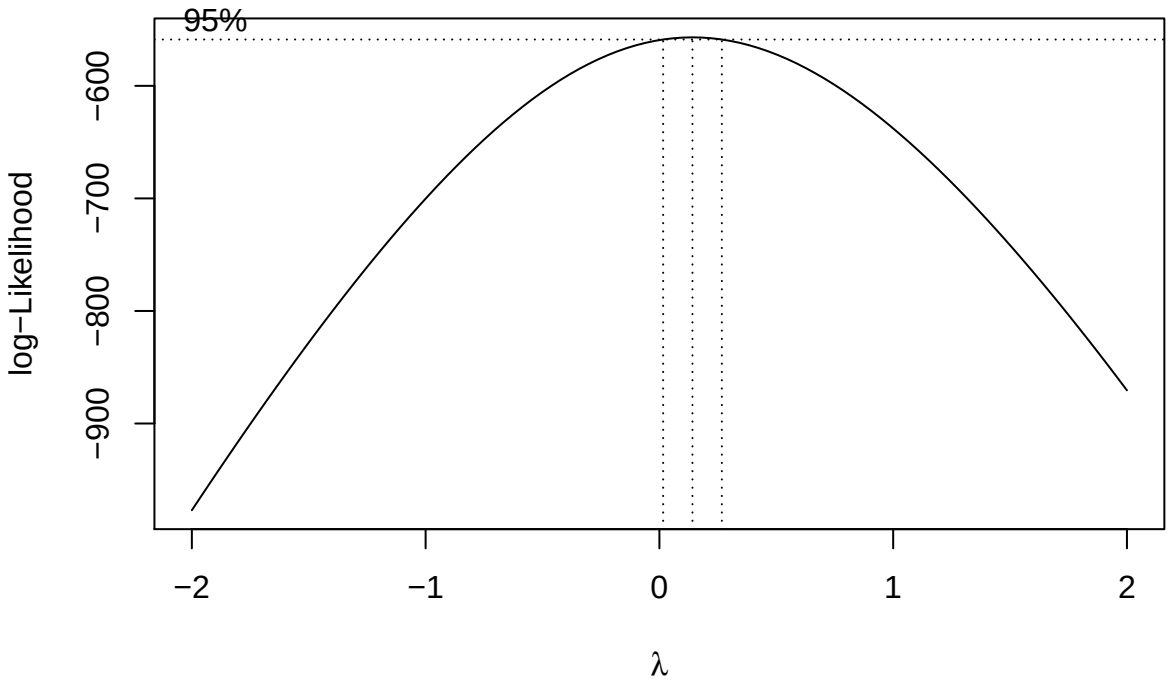
9.2 Histogram for Q09




```
Saving 6.5 x 4.5 in image
```

9.3 Box-Cox plot for Q09

```
MASS::boxcox(lm(outcome ~ 1, data = data_09))
```



9.4 "Skim" Results for Q09

```
data_09 %>% select(outcome) %>% skim
```

```
Skim summary statistics
 n obs: 200
 n variables: 1

Variable type: numeric
 variable missing complete   n mean   sd  p0  p25 median  p75  p100
  outcome       0      200 200 3.17 3.13 0.2 0.73    2.2 4.45 14.96
     hist
 <U+2587><U+2583><U+2582><U+2582><U+2581><U+2581><U+2581><U+2581>
```

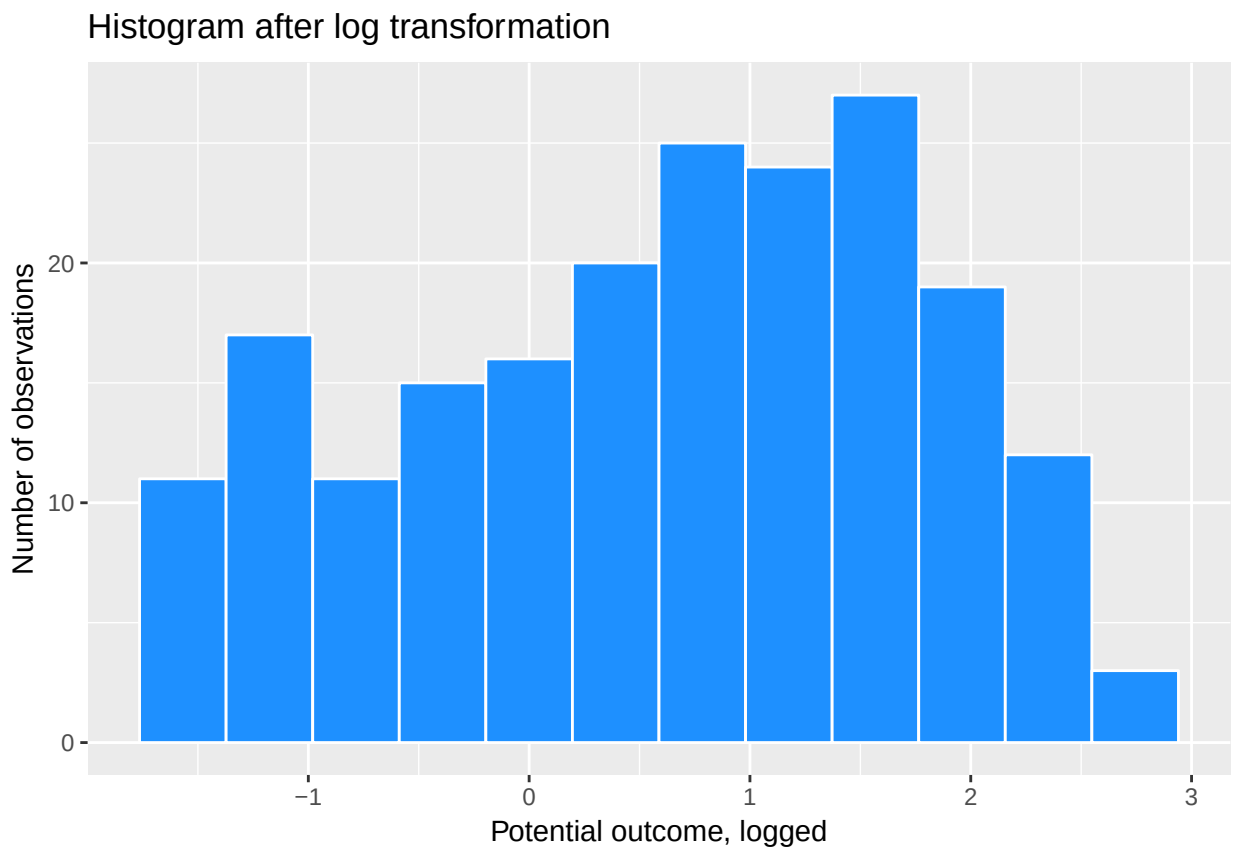
Consider the information provided above (histogram, Box-Cox plot + skim results) on the distribution of a potential outcome variable in a linear regression model to be built using the `data_09` data set. Which of the following transformations of the `outcome` data would be most appropriate in this setting?

a. No transformation is needed. Fit the model to the raw outcome.
b. A log transformation is likely to be helpful.
c. Squaring the data would be helpful.
d. We should use a restricted cubic spline.
e. We should center the data.
f. It is impossible to tell from the information provided.

9.5 Answer 9 is b.

These `outcome` data are substantially right skewed, and we'd like them to be more symmetric, and better approximated by a Normal distribution. That is usually best accomplished by a log transformation. The Box-Cox plot also suggests a transformation with power near 0 (the logarithm.) In this case, the transformation works reasonably well.

```
ggplot(data_09, aes(x = log(outcome))) +
    geom_histogram(fill = "dodgerblue", col = "white", bins = 12) +
    labs(x = "Potential outcome, logged", y = "Number of observations",
         title = "Histogram after log transformation")
```



- Squaring the data would only make them more right skewed.
- A restricted cubic spline isn't appropriate for an outcome transformation.
- Centering the data will do nothing to the shape.

9.6 Q09 Results

- At least 38/41 students got full credit.
- Over 95% of available points were awarded.
- No partial credit was available.

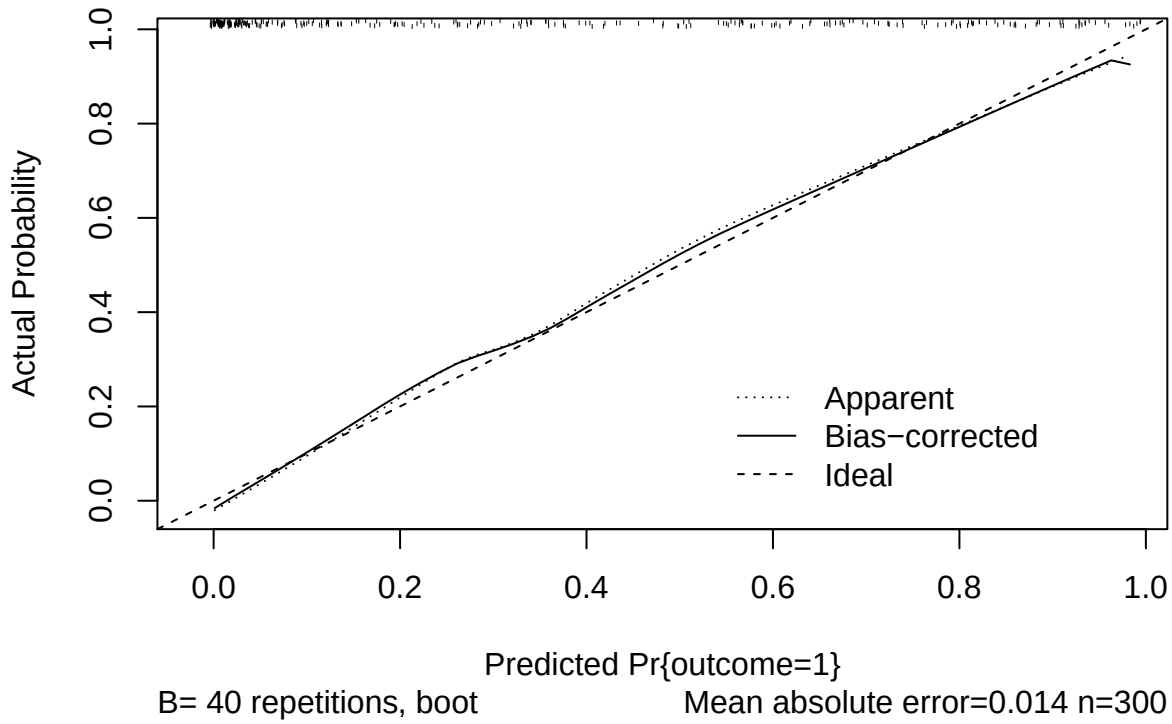
10 Question 10. (4 points)

10.1 Setup for Question 10.

```
set.seed(43210)
outcome <- c(rep(1,100), rep(0, 200))
sev <- c(rnorm(100, mean = 90, sd = 10), rnorm(200, mean = 65, sd = 15))
fem <- c(rep(1, 50), rep(0, 50), rep(1, 103), rep(0, 97))
com <- c(rpois(100, lambda = 4), rpois(200, lambda = 2))
soc <- c(rchisq(100, df = 1), rnorm(200, 5, 2))

data_10 <- data_frame(
    subject = 1:300,
    outcome = outcome,
    severity = ifelse(sev > 0, sev, 0),
    female = fem,
    comorbidities = ifelse(com > 0, com, 0),
    social.support = ifelse(soc > 0, soc, 0)
)

rm(outcome, sev, fem, com, soc)

d = datadist(data_10)
options(datadist = "d")
modelA <- lrm(outcome ~ severity + female, data = data_10, x = TRUE, y = TRUE)
modelB <- lrm(outcome ~ severity + female + comorbidities + social.support, data = data_10, x = TRUE, y
```

10.2 Calibration Plot for Model A

```
plot(calibrate(modelA))
```

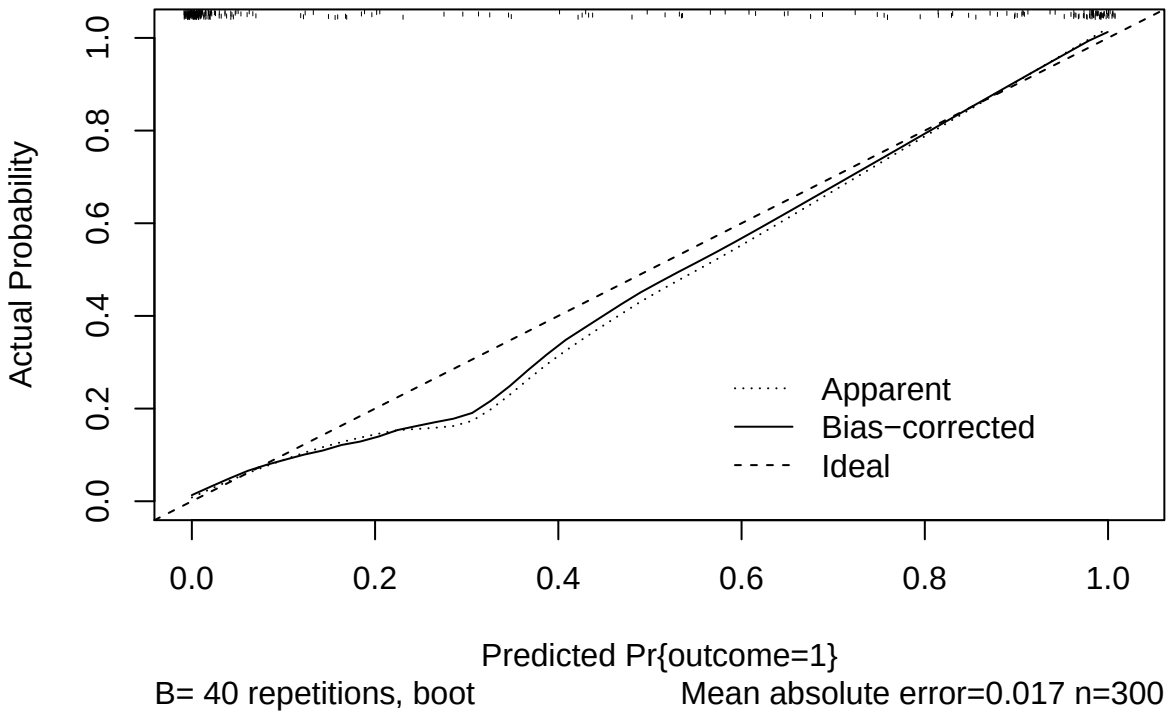



```
n=300   Mean absolute error=0.014   Mean squared error=0.00026
0.9 Quantile of absolute error=0.024
```

Note that the Nagelkerke R-squared for Model A is 0.596

10.3 Calibration Plot for Model B

```
plot(calibrate(modelB))
```




```
n=300   Mean absolute error=0.017   Mean squared error=0.00059
0.9 Quantile of absolute error=0.035
```

Note that the Nagelkerke R-squared for Model B is 0.865

Two models, called `modelA` and `modelB` were fit to predict the same binary outcome using logistic regression on the same data set of 300 subjects. The calibration plots and Nagelkerke R-squared values from the `rms` package are shown for each model, above. Based on those results, which of the following best describes the conclusions we should draw?

a. Model A shows better calibration but worse discrimination.
b. Model A shows worse calibration but better discrimination.
c. Model A shows better calibration and better discrimination.
d. Model A shows worse calibration and worse discrimination.
e. None of these conclusions are appropriate.

10.4 Answer for Question 10 is a

Model A shows better calibration (the bias-corrected estimates are closer to the ideal line), but worse discrimination than Model B as measured by the Nagelkerke R-squared.

10.5 Q10 Results

- At least 38/41 students got full credit.
- Over 95% of available points were awarded.

- No partial credit was available.

11 Question 11. (2 points)

11.1 Setup for Questions 11-14

The same setting will apply to questions Q11 - Q14. In attempting to measure the complex relationships between four potential treatments and primary insurance on a summary measure of health obtained after treatment among 360 Northeast Ohio residents, two linear models were developed, called Model C and Model D. Each of the 360 subjects received exactly one of the four Treatments (although Treatments A and B were selected more often than C or D), and the sample was obtained to include equal numbers of Medicare, Medicaid and Commercially insured subjects.

```
set.seed(43211)

tre = c(rep("A", 120), rep("B", 120), rep("C", 60), rep("D", 60))
ins = c(rep("Medicare", 40), rep("Medicaid", 40), rep("Commercial", 40), rep("Medicare", 40), rep("Medi
hea = c(rnorm(40, 200, 50), rnorm(40, 180, 50), rnorm(40, 190, 50), rnorm(40, 185, 50), rnorm(40, 165, 5

data_11 <- data_frame(subject = 1:360,
                      treatment = tre,
                      insurance = ins,
                      health = hea)
#data_11

modelC <- lm(health ~ treatment + insurance, data = data_11)

modelD <- lm(health ~ treatment * insurance, data = data_11)

remove(tre, ins, hea, d, wrong)
```

11.2 Output for Q11

```
anova(modelC)
```

```
Analysis of Variance Table

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

```
anova(modelC, modelD)
```

```
Analysis of Variance Table

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

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What was included in modelD but not included in modelC?

11.3 Answer 11 is "the interaction of treatment and insurance"

Model D is `lm(health ~ treatment * insurance)` according to the ANOVA table comparing Models C and D.

11.4 Q11 Results

Everyone successfully answered Q11.

12 Question 12. (2 points)

Did the additional piece in modelD that was added to modelC account for statistically significant predictive value for health?

a. Yes, at the 5% significance level.
b. No, at the 5% significance level.
c. It is impossible to tell from the output provided.

12.1 Answer 12 is a.

Yes. The p value is 0.0058

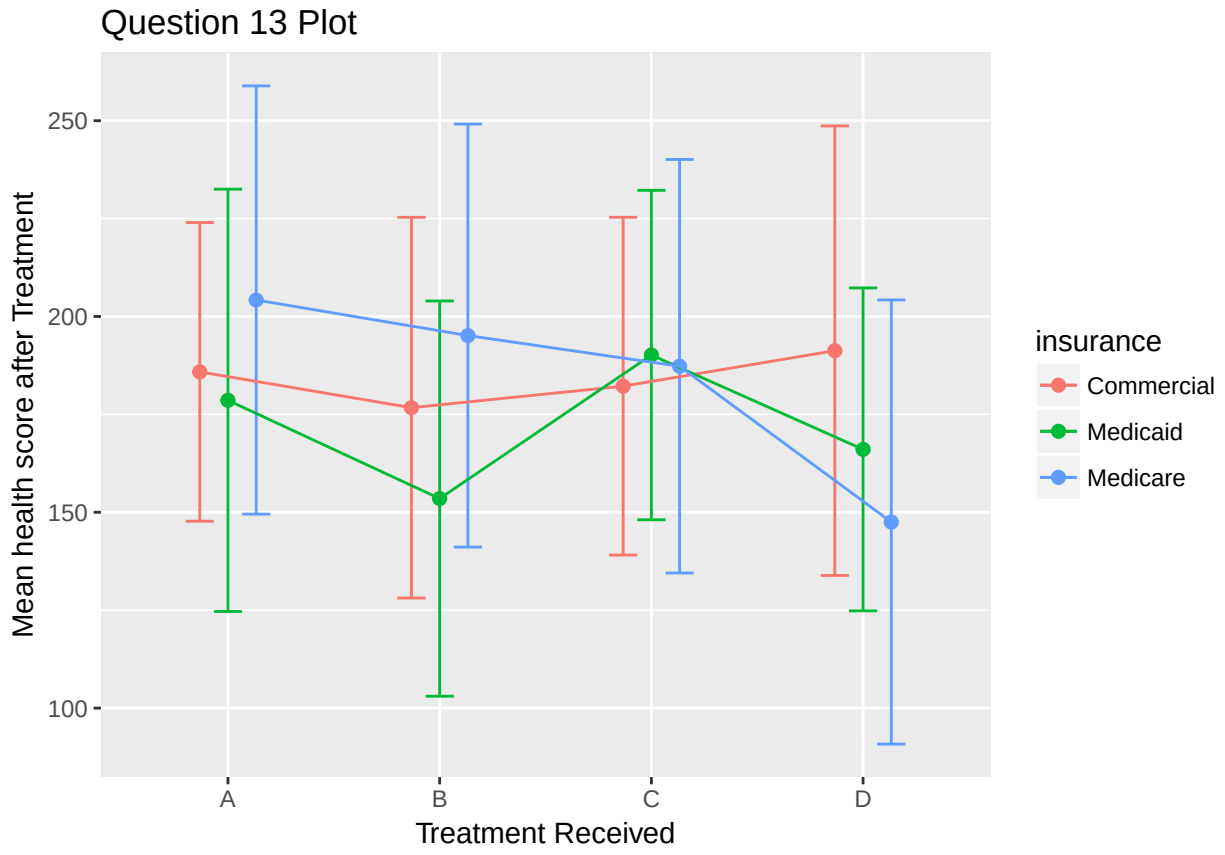
12.2 Q12 Results

- At least 38/41 students got full credit.
- Over 95% of available points were awarded.
- No partial credit was available.

13 Question 13. (3 points)

13.1 Plot for Q13

```
data_11sum <- data_11 %>%
    group_by(treatment, insurance) %>%
    summarize(mean.health = mean(health), sd.health = sd(health))
pd <- position_dodge(0.4)

ggplot(data_11sum, aes(x = treatment, y = mean.health, col = insurance)) +
    geom_errorbar(aes(ymin = mean.health - sd.health,
                      ymax = mean.health + sd.health),
                  width = 0.4, position = pd) +
    geom_point(size = 2, position = pd) +
    geom_line(aes(group = insurance), position = pd) +
    labs(y = "Mean health score after Treatment", x = "Treatment Received",
         title = "Question 13 Plot")
```




```
ggsave("figures/fig13.png")
```

```
Saving 6.5 x 4.5 in image
```

What does the Plot for Q13 shown above (of means with intervals indicating one standard deviation in either direction) suggest about the best choice of model, comparing `modelC` to `modelD`?

a. `modelC` seems like the better choice.
b. `modelD` seems like the better choice.

c. This plot does not help us make the decision.

13.2 Answer 13 is b.

There's clearly an interaction in the plot of means. The lines joining the group means intersect, quite a bit. Which treatment is best seems inexorably linked to insurance. Model D includes the interaction term.

13.3 Q13 Results

- 34/41 students got full credit.
- 83% of available points were awarded.
- No partial credit was available.

Response	a	b	c
%	2	83	15

14 Question 14. (4 points, 1 point each for a, b, c, and d)

14.1 Output for Q14

```
TukeyHSD(aov(health ~ treatment + insurance, data = data_11))
```

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

Fit: aov(formula = health ~ treatment + insurance, data = data_11)

$treatment
          diff       lwr        upr     p adj
B-A -14.441051 -31.39497  2.5128687 0.1255680
C-A  -3.000425 -23.76465 17.7638020 0.9822740
D-A -21.272656 -42.03688 -0.5084291 0.0422930
C-B  11.440627  -9.32360 32.2048534 0.4862733
D-B  -6.831604 -27.59583 13.9326223 0.8307770
D-C -18.272231 -42.24869  5.7042326 0.2024593

$insurance
                         diff        lwr       upr     p adj
Medicaid-Commercial -13.04163 -28.500264  2.417002 0.1172331
Medicare-Commercial   5.80960  -9.649033 21.268233 0.6504308
Medicare-Medicaid    18.85123   3.392598 34.309864 0.0120740
```

Using `modelC`, the Output for Q14 shown above was developed. If larger values of the health outcome are better, then, based on the output above, which of the following conclusions can you draw, at a global 95% confidence level?

ROWS:

a. Treatment D looks significantly better than Treatment A.
b. Treatment B looks significantly better than Treatment A.
c. Medicare looks significantly better than Medicaid.
d. Medicare looks significantly better than Commercial.
e. None of these statements are true.

COLUMNS:

- TRUE
- FALSE

14.2 Answer 14 is `a` is FALSE, `b` is FALSE, `c` is TRUE, `d` is FALSE.

- Note that D actually looks significantly worse than A by the table of Tukey comparisons.
- There is no significant difference in the table for B - A.
- Medicare does look significantly better than Medicaid
- Medicare does not look signifcantly better than Commercial.

14.3 Q14 Results

Question	a	b	c	d
Received full credit ($n = 41$)	36	41	> 37	> 37
% of available points awarded	88	100	> 95	> 95

- No partial credit was available.

15 Question 15. (3 points)

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

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

15.1 Answer 15 is b and c

Leek: "The #1 way to be a jerk reviewer is ... [to ask for either major or minor revisions] ... even if you think the paper is uninteresting and you wouldn't accept it even if they did everything you said."

15.2 Q15 Results

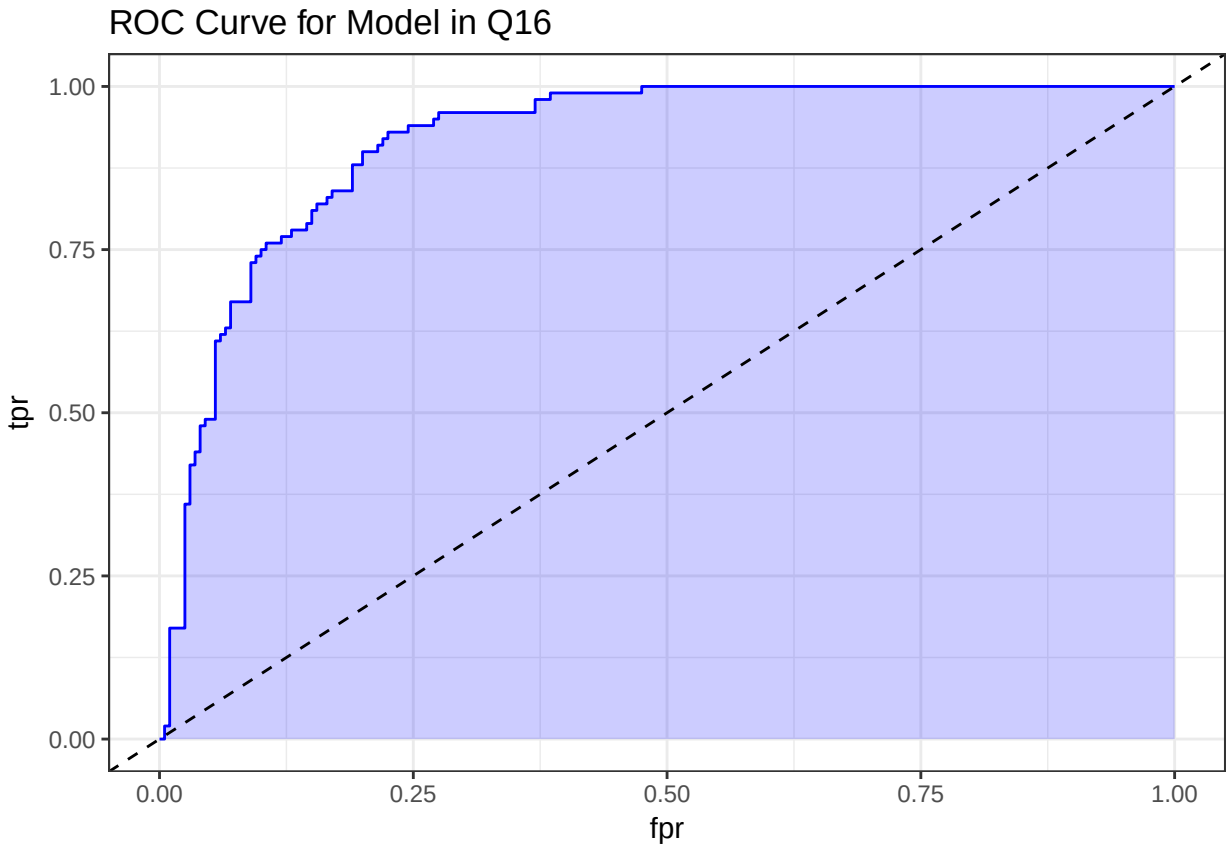
- 24/41 students got full credit.
- 59% of available points were awarded.
- No partial credit was available, even if you got one of the two right, or if you got both of these, but also indicated other options.

Selection	% of respondents
b and c	**24**
b only	6
a, b and c	5
a, c and d	2
a only	1
a and b	1
a and c	1
d only	1

16 Question 16. (3 points)

16.1 Plot for Q16

```
Loading required package: gplots


Attaching package: 'gplots'

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

    lowess
```




```
Saving 6.5 x 4.5 in image
```

The ROC curve plotted above was developed for a logistic regression model with multiple predictors. Which of the following C statistics is associated with this curve?

a. C = 0.52
b. C = 0.62
c. C = 0.72
d. C = 0.92

16.2 Answer 16 is d

As it turns out, it's just `modelB` from Q10. The C statistic for that model is 0.916. Here's the proof.

```
modelB
```

```
Logistic Regression Model

 lrm(formula = outcome ~ severity + female + comorbidities + social.support,
     data = data_10, x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           300    LR chi2     292.30    R2       0.865    C       0.985
  0            200    d.f.             4    g        6.115    Dxy     0.970
  1            100    Pr(> chi2) <0.0001    gr     452.817    gamma   0.970
 max |deriv| 5e-05                          gp       0.432    tau-a   0.432
                                            Brier    0.043

                Coef     S.E.   Wald Z Pr(>|Z|)
 Intercept      -12.6631 2.2647 -5.59  <0.0001
 severity         0.1553 0.0264  5.88  <0.0001
 female          -0.4257 0.5614 -0.76  0.4483
 comorbidities    0.9573 0.2109  4.54  <0.0001
 social.support  -0.9397 0.1587 -5.92  <0.0001
```

16.3 Q16 Results

- 37/41 students got full credit.
- 90% of available points were awarded.
- No partial credit was available, and if you thought this was C = 0.72, you need to recalibrate yourself.

17 Question 17. (4 points)

17.1 Code Attempt for Q17

```
data_17 <- data_17 %>%
    mutate(gov_ins = factor(insurance,
                            Medicare or Medicaid = Yes,
                            Commercial or Uninsured = No))
```

On the Quiz 1 web page, there is a file called `data_17.csv` containing insurance data on thousands of subjects, each of whom is classified as falling into one of four different insurance categories, specifically Medicare, Commercial, Medicaid, and Uninsured. Some of the subjects (less than 5%) have missing data on this `insurance` variable. Assume that the tidyverse has been loaded in R, and that the data have been loaded into a tibble called `data_17`. Suppose you now want to create a variable called `gov_ins` within the `data17` tibble that (a) is a factor, and (b) which takes the value Yes if the subject's insurance is provided by the goverment (Medicare or Medicaid) but No otherwise, while (c) retaining NA for the missing values. Your first attempt is as shown in the Code Attempt for Q17. Fix the call to the mutate function in that code so that your resulting code will actually do what is required.

17.2 Answer for Q17 is a line of code

The code you need is

```
mutate(gov_ins = fct_recode(insurance,
                            Yes = "Medicare",
                            Yes = "Medicaid",
                            No = "Commercial",
                            No = "Uninsured"))
```

Here's the proof that this works.

```
data_17 <- read.csv("data/data_17.csv") %>% tbl_df

data_17 <- data_17 %>%
    mutate(gov_ins = fct_recode(insurance,
                            Yes = "Medicare",
                            Yes = "Medicaid",
                            No = "Commercial",
                            No = "Uninsured"))

data_17 %>% count(gov_ins, insurance)
```

```
# A tibble: 5 x 3
  gov_ins insurance      n
  <fct>   <fct>      <int>
1 No      Commercial  1786
2 No      Uninsured    413
3 Yes     Medicaid    1345
4 Yes     Medicare    1085
5 <NA>    <NA>         171
```

17.3 Q17 Results

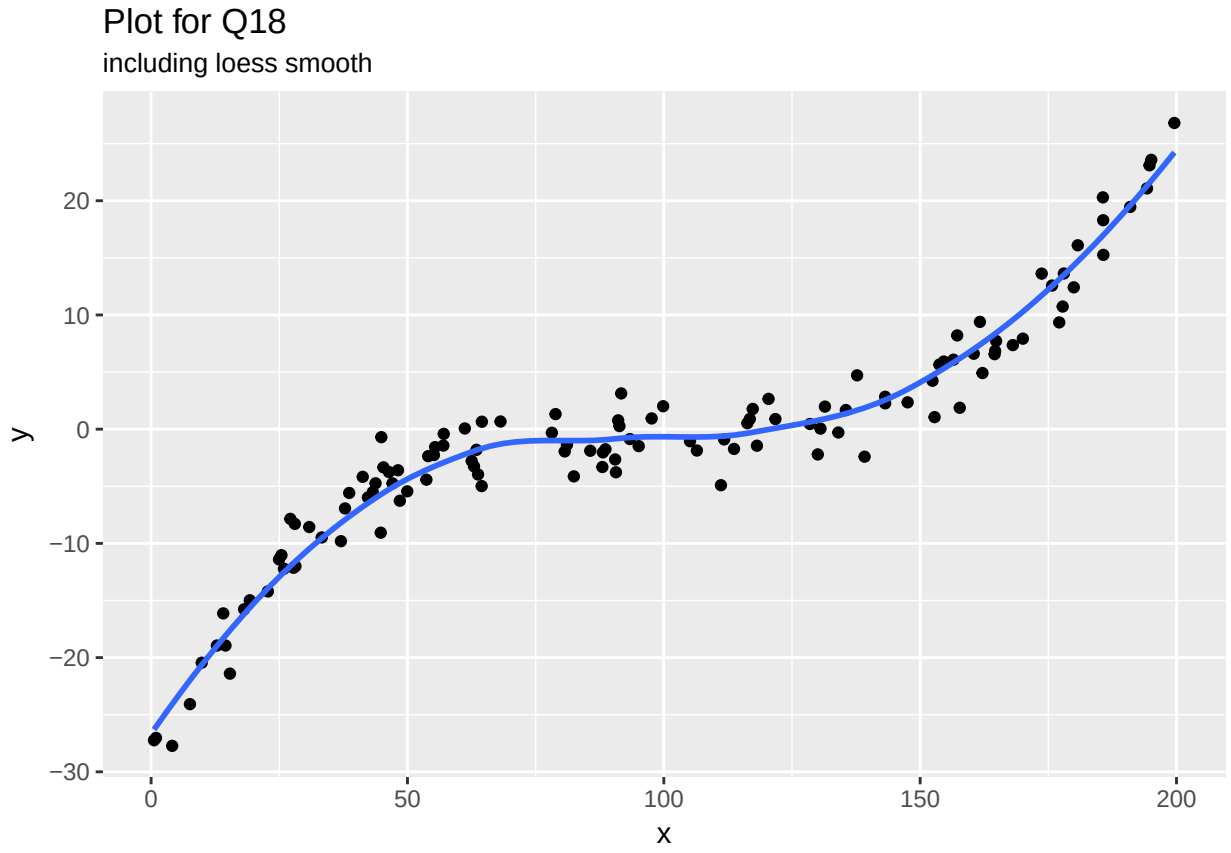
- 32/41 students got full credit.
- 87% of available points were awarded.
- Some partial credit was available. I ran the code for all 41 of you through R, and awarded full credit to several other approaches that also got the job done, so long as they didn't include multiple restatements of the data, used `mutate` correctly, and placed the correct results in the right place.

18 Question 18. (3 points)

18.1 Setup for Question 18

```
set.seed(43218)

x <- runif(120, 0, 200)
y1 <- (-15083955 + ((x - 85) + 60*(x - 100)^2 + 900*(x - 100)^3 ))/32292134
err <- rnorm(120, 0, 2)
y <- y1 + err

data_18 <- data_frame(subject = 1:120, x = x, y = y)

rm(x, y, y1, err)
```

18.2 Plot for Q18




```
Saving 6.5 x 4.5 in image
```

Suppose the relationship between a predictor, x, and an outcome, y, is described by the plot for Q18 shown above, which includes the fit from a loess smooth. What is the minimum number of knots that would be required in a restricted cubic spline on x to fit a model that approximates the general shape of the curve shown in the plot above?

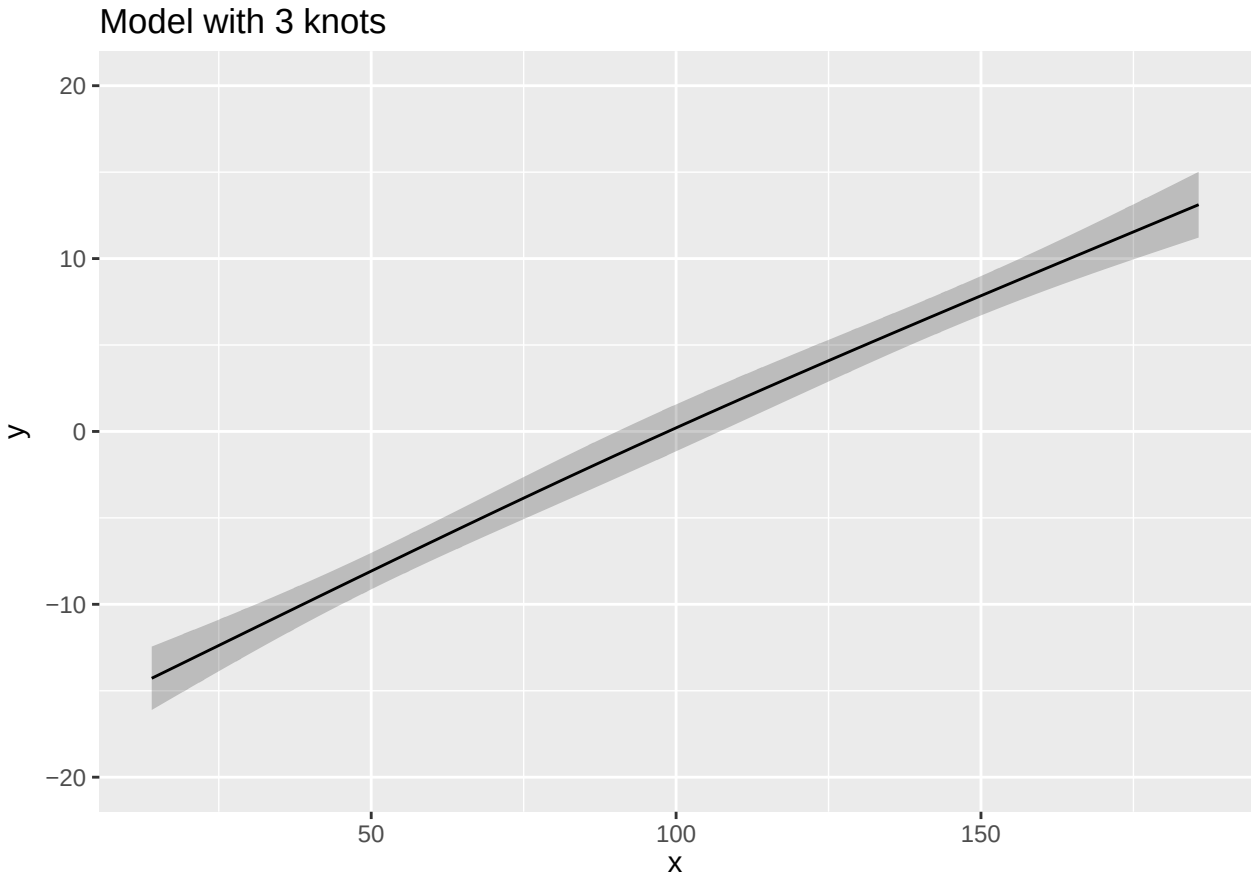
a. Less than three knots would be required.

b. Three knots would be required
c. Four knots would be required.
d. Five or more knots would be required.
e. It is impossible to tell from the information provided.

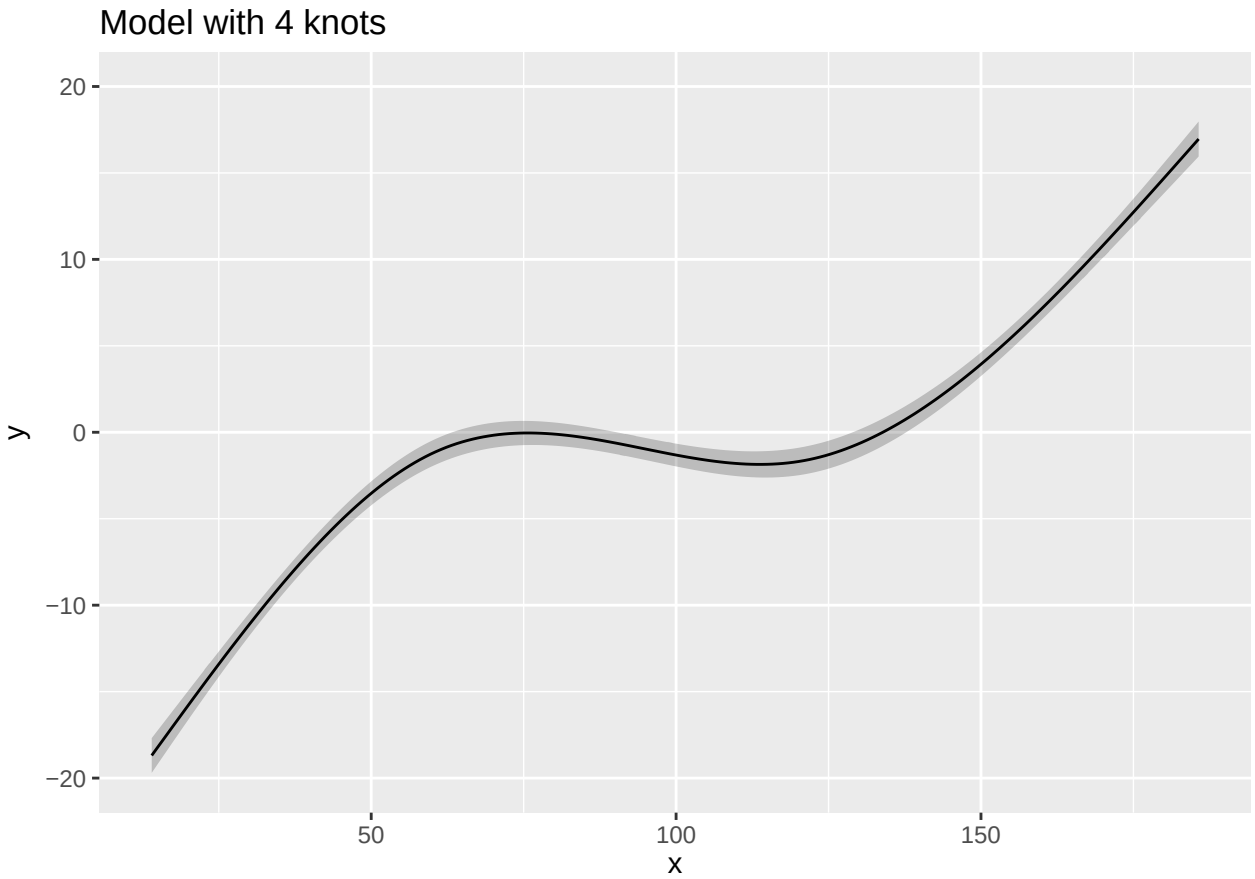
18.3 Answer 18 is c.

We have two bends in the plot, and so would need at least four knots. See Section 9.6 of the Course Notes. We can see from the plots below that the models with 4 or more knots work, and the model with 3 does not. So 4 is the minimum number of knots.

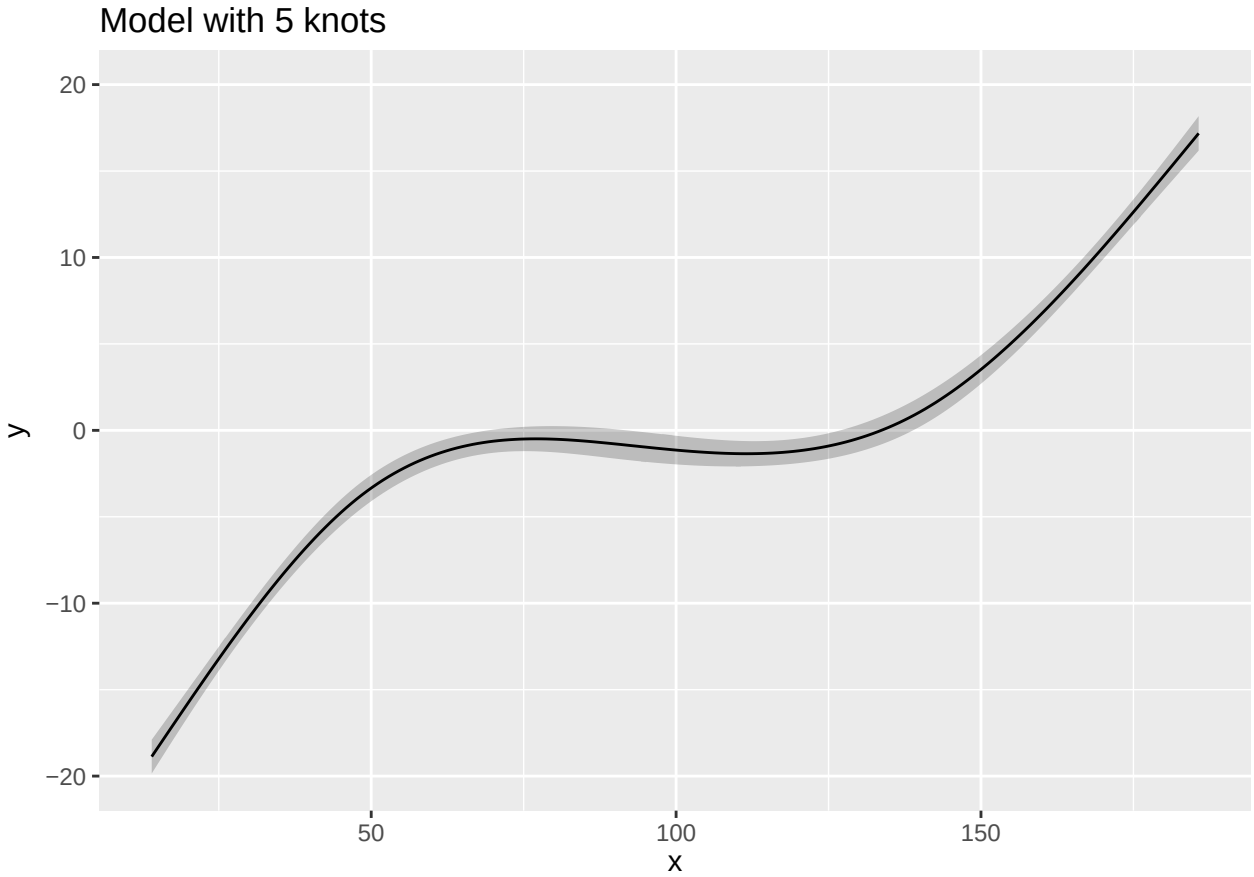
```
d <- datadist(data_18)
options(datadist = "d")

mod_3knots <- ols(y ~ rcs(x, 3), data = data_18, x = T, y = T)
mod_4knots <- ols(y ~ rcs(x, 4), data = data_18, x = T, y = T)
mod_5knots <- ols(y ~ rcs(x, 5), data = data_18, x = T, y = T)

ggplot(Predict(mod_3knots)) + labs(title = "Model with 3 knots")
```




```
ggplot(Predict(mod_4knots)) + labs(title = "Model with 4 knots")
```




```
ggplot(Predict(mod_5knots)) + labs(title = "Model with 5 knots")
```



18.4 Q18 Results

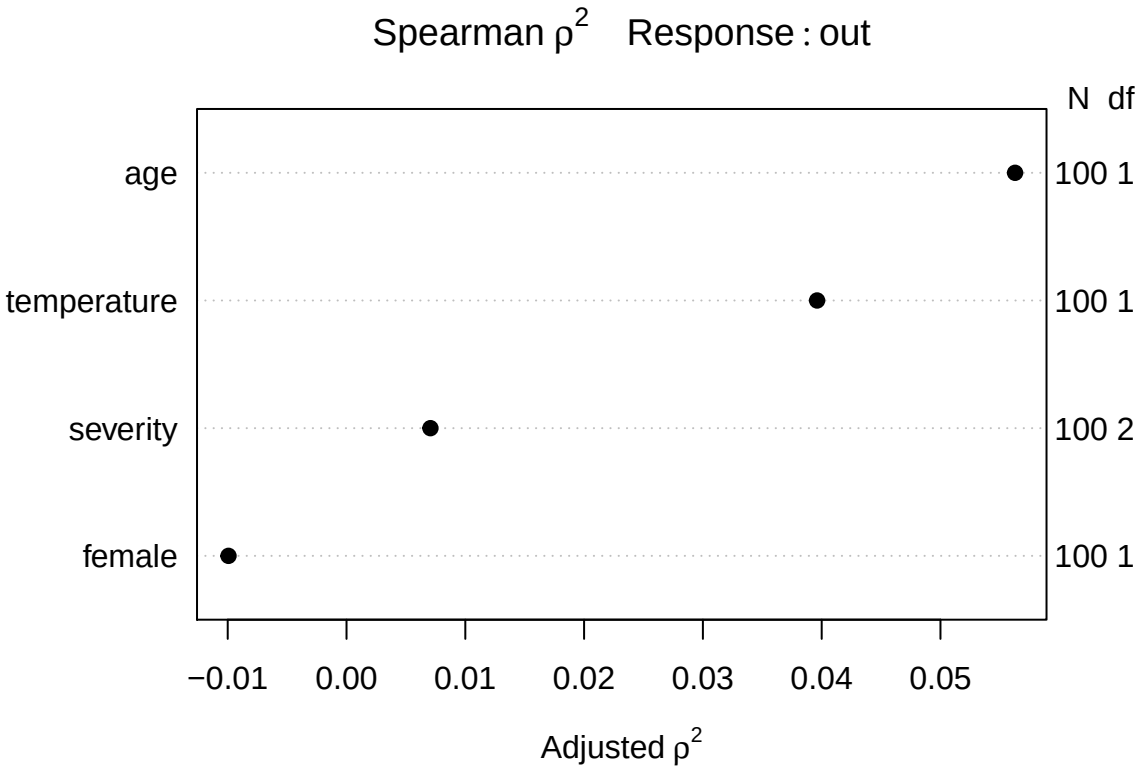
- 35/41 students got full credit.
- 85% of available points were awarded.
- No partial credit was available.

19 Question 19. (3 points)

19.1 Setup for Question 19

```
set.seed(43239)

out <- rnorm(100, 200, 25)
age <- runif(100, 28, 60) + out/20
fem <- rbernoulli(100, 0.48)
sev <- c(rep("High", 20), rep("Middle", 40), rep("Low", 40))
temp <- out/100 + rnorm(100, 98.6, 3)

data_19 <- data_frame(out, age, female = fem, severity = sev, temperature = temp)

rm(out, age, fem, sev, temp)
```

19.2 Plot for Q19



You are building a linear regression model with only a limited number of observations, and need to include four predictors: age (in years), female (1 = female, 0 = male), severity (three categories: High, Medium, Low) and temperature (in degrees Celsius). Examine the Spearman rho-squared plot shown as the Plot for Q19. Suppose you are permitted to spend an additional three degrees of freedom beyond those accounted for by the intercept term and the main effects of these predictors. Based on the plot, which of the models below best does this additional spending?

a. `ols(outcome ~ rcs(age, 3) + temperature + rcs(severity, 3) + female %ia% age, data = data_19)`
b. `ols(outcome ~ rcs(age, 4) + rcs(temperature, 3) + severity + female, data = data_19)`
c. `ols(outcome ~ female*age + temperature*severity, data = data_19)`
d. `ols(outcome ~ rcs(age, 3) + rcs(temperature, 5) + severity + female, data = data_19)`
e. None of these models are appropriate.

19.3 Answer 19 is b.

Model `b` spends exactly three more degrees of freedom, focusing first on `age` and then `temperature`. That's the only option that follows the suggestion of the Spearman ρ^2 plot.

- Model `a` tries to take a spline of a categorical variable: `severity`.
- Model `c` prioritizes interactions including `severity` and `female` which is the opposite of what the plot suggests.
- Model `d` spends more than three additional degrees of freedom.

19.4 Q19 Results

- 16/41 students got full credit.
- 39% of available points were awarded.

This was quite a bloodbath, although the correct response was the most common selection.

Response	a	b	c	d	e
%	10	39	24	0	27

The most common answer wrong answer was e, and I don't know why you wouldn't think b was appropriate, so that was a surprise to me.

20 Question 20. (3 points)

In R for Data Science, Grolemund and Wickham produce a series of models for the number of daily flights that leave New York City for each day in 2013. Their initial models describe the mean effect, but due to some outlier issues, they instead use a model that reduces the impact of the outliers on their estimates. Please answer both parts (a) and (b)... (a) What function do they use to fit such a model? AND (b) What is the name of the package that includes that function?

20.1 Answer 20 is `rlm`. It comes from the `MASS` package.

See, for instance, section 24.3.2 of *R for Data Science*.

20.2 Q20 Results

- At least 38/41 students got full credit.
- More than 95% of available points were awarded.
- No partial credit was available.

21 Question 21 (6 points, 1 each for a-f)

For each row, identify the model-fitting approach that most directly leads to the specified summary.

Rows:

a. C statistic
b. Adjusted R-squared
c. odds ratio estimate showing effect of increasing a quantitative predictor by 1 unit, while holding the others constant
d. odds ratio estimate showing effect of increasing a quantitative predictor from the 25th to the 75th percentile of its distribution, while holding the others constant
e. estimate of effect on a quantitative outcome associated with changing a binary predictor from 0 to 1
f. Nagelkerke R-squared

Columns:

1. ols fit
2. lrm fit
3. glm fit with binomial family
4. None of these

21.1 Answer 21 is `a` = 2, `b` = 1, `c` = 3, `d` = 2, `e` = 1, `f` = 2

- The C statistic and Nagelkerke R^2 are provided in the main output for a `lrm` fit using logistic regression.
- With `summary` of an `lrm` fit, we get the odds ratio estimate showing the effect of increasing a quantitative predictor from the 25th to the 75th percentile of its distribution, while holding the others constant.
- A linear regression with `ols` provides Adjusted R^2 directly, and with `summary`, the estimated effect on a quantitative outcome associated with changing a binary predictor from 0 to 1.
- The `glm` approach, with `exp(coef(modelname))` and `exp(confint(modelname))`, provides the odds ratio estimate showing effect of increasing a quantitative predictor by 1 unit, while holding the others constant

21.2 Q21 Results

Q21 Part	a	b	c	d	e	f
Full credit	37	36	35	26	27	41
% of points	90	88	85	63	66	100

- No partial credit was available.

Taking a closer look at `d`, we see that:

- 26 people said "lrm", correctly
- 10 people said "ols"
- 1 person said "glm"
- and 4 said "none of these"

Regarding 21`e`, we see that:

- 27 said "ols", correctly
- 7 said "glm"
- 2 said "lrm"
- and 5 said "none of these"

22 Question 22. (3 points)

22.1 Output for Q22

Model	Median Predicted Outcome	Root Mean Squared Error	Mean Absolute Error
Model R	55	5.06	4.27
Model S	56	4.97	3.77
Model T	53	7.15	4.14

I investigated three models (which I'll call R, S and T), each of which was suggested by a different summary measure in a training data set. Passing the resulting models through to a test sample of 275 observations, I obtained the summaries shown in the Output for Q22. Which model fits the data in the test sample best?

a. Model R
b. Model S
c. Model T
d. It is impossible to tell from the information provided.

22.2 Answer 22 is `b`

Model S has the smallest RMSE and MAPE of these three models in the test sample.

22.3 Q22 Results

- At least 38/41 students got full credit.
- More than 95% of available points were awarded.
- No partial credit was available.

23 Question 23. (4 points)

23.1 Setup for Question 23

```
data_23 <- UsingR::babies %>%
    mutate(baby.wt = wt,
           mom.age = age,
           mom.ht = ht,
           mom.wt = wt1,
           dad.age = dage,
           dad.ht = dht,
           dad.wt = dwt) %>%
    dplyr::select(baby.wt, gestation,
           mom.age, mom.ht, mom.wt,
           dad.age, dad.ht, dad.wt)

write_csv(data_23, "data/data_23.csv")
```

23.2 Output for Q23

```
model_23 <- lm(baby.wt ~ gestation + mom.age + mom.ht +
                  mom.wt + dad.age + dad.ht + dad.wt,
               data = data_23)

vif(model_23)
```

```
gestation   mom.age    mom.ht    mom.wt   dad.age    dad.ht    dad.wt
 1.004672  1.634678  1.588963  1.580195  1.642931 13.578469 13.582143
```

A child's birth weight depends on many things, among them the parents' genetic makeup, gestation period, and mother's activities during pregnancy. Suppose we are fitting a model using the `data_23` data set (which I have provided to you, so you should take a look at it) to predict a child's weight at birth using seven continuous predictors: gestation period (in days), mother's age (in years), height (in inches) and (pre-pregnancy) weight (in pounds); and father's age, height and weight. We run a full model containing main effects of all predictors, and obtain the variance inflation factors shown in the Output for Q23 provided above. Which of the following statements best describes the most appropriate next step to take in light of this output and your examination of the `data_23` data set?

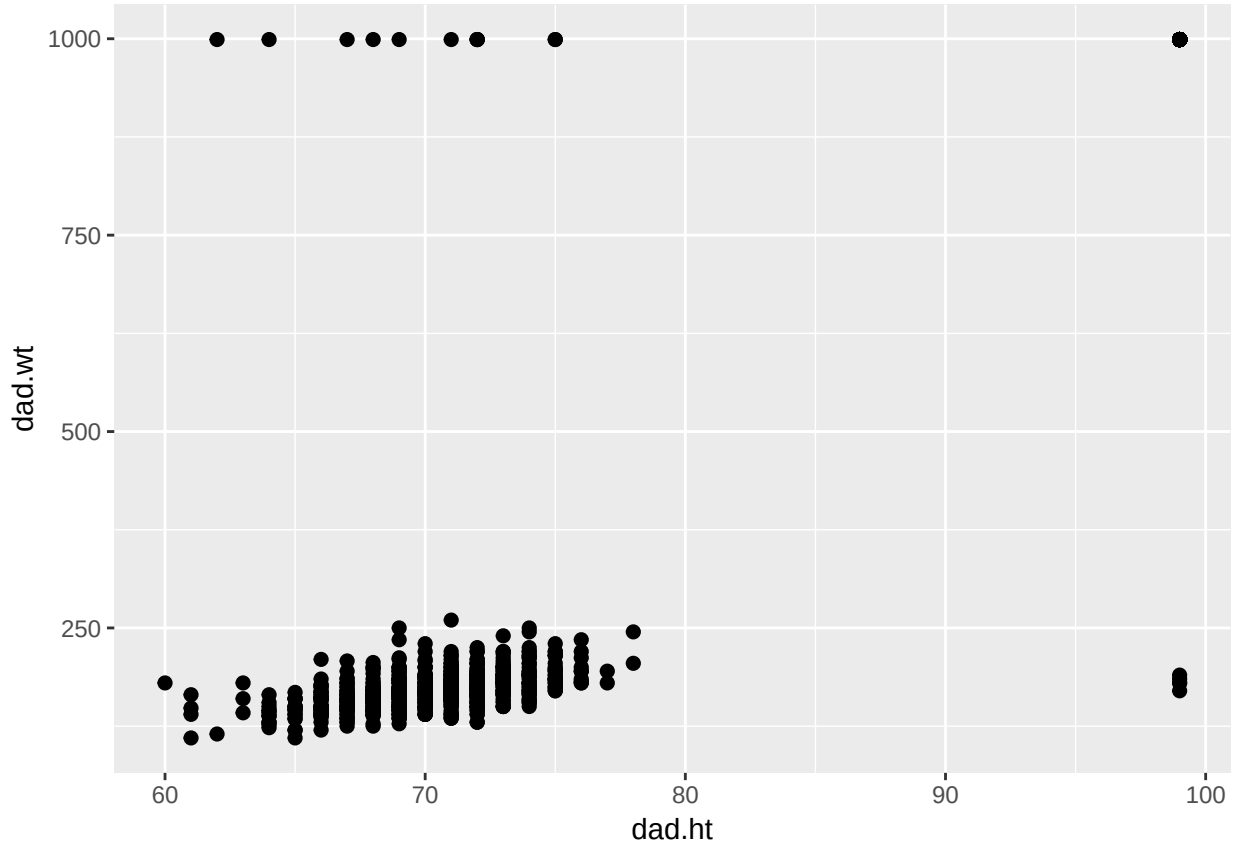
a. We should calculate and examine the leverage values for our observations.
b. We should draw pictures, and study the distributions of the predictors in our model.
c. We should drop the `dad.wt` variable and fit a new model.
d. We should drop the `dad.ht` variable and fit a new model.
e. We definitely need to drop one of the predictors about the father's size, but we're not sure which one.
f. We should calculate and examine the Cook's distance for our observations.
g. There isn't a substantial problem here. We can confidently make predictions using this model.

23.3 Answer 23 is b.

This question is based on the `babies` data frame in the `UsingR` package, and in a related example (10.2) in Verzani J *Using R for Introductory Statistics*, First Edition. The correct answer is that we don't yet know whether we need to drop a predictor in light of this apparent collinearity, or whether something else is wrong,

so we should draw some pictures and look at the distributions of the predictors in our model. When we do this, especially for `dad.ht` and `dad.wt`, we see something interesting.

```
ggplot(data_23, aes(x = dad.ht, y = dad.wt)) +
    geom_point(size = 2)
```



Most of the points make sense, it looks like weight is in pounds, and height is in inches, but we have a lot of values of `dad.wt` that are 999, and a lot of values in `dad.ht` that are 99. It turns out that in entering these data into the computer, someone chose 99 (for age and height) and 999 (for gestation and weight) as indicators of missing values. You could have caught this with a single `skim`.

```
skim(data_23)
```

```
Skim summary statistics
 n obs: 1236
 n variables: 8

Variable type: numeric
  variable missing complete    n   mean     sd  p0    p25 median p75 p100
   baby.wt       0     1236 1236 119.58  18.24  55 108.75    120 131  176
   dad.age       0     1236 1236  30.74   8.52  18  25        29  35   99
    dad.ht       0     1236 1236  81.67  14.28  60  70        73  99   99
    dad.wt       0     1236 1236 505.4  406.69 110 165       190 999  999
 gestation       0     1236 1236 286.91  75.16 148 272       280 288  999
   mom.age       0     1236 1236  27.37   6.46  15  23        26  31   99
    mom.ht       0     1236 1236  64.67   5.26  53  62        64  66   99
    mom.wt       0     1236 1236 153.98 147.87  87 115       126 140  999
     hist
```

```
 <U+2581><U+2581><U+2582><U+2586><U+2587><U+2585><U+2581><U+2581>
 <U+2587><U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2581><U+2585><U+2587><U+2581><U+2581><U+2581><U+2581><U+2587>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2586>
 <U+2581><U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2587><U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2581><U+2587><U+2586><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
```

If we drop these observations with missing `dad.wt` or `dad.ht` from our model, then the collinearity problem vanishes in the remaining 701 cases.

```
babies_nomiss <- data_23 %>%
    filter(gestation < 999 & mom.age < 99 & mom.ht < 99 &
                mom.wt < 999 & dad.age < 99 & dad.ht < 99 &
                dad.wt < 999)

nrow(babies_nomiss)
```

```
[1] 701
```

```
model_23_nomiss <- lm(baby.wt ~ gestation + mom.age + mom.ht +
                      mom.wt + dad.age + dad.ht + dad.wt,
                data = babies_nomiss)

vif(model_23_nomiss)
```

```
gestation   mom.age    mom.ht    mom.wt   dad.age    dad.ht    dad.wt
 1.010953  3.230649  1.349065  1.310139  3.321813  1.556530  1.457728
```

23.4 Q23 Results

Before the test, I thought this was going to be the question with the poorest results. I assumed that people would not look at the data on their own, and just look at the output provided on the test. Since the output on the test suggests a problem with the VIF values, I thought people would leap to the conclusion that we have a collinearity problem, rather than a "poor coding of missing values" problem. That was, deliberately, tricky.

- 9/41 students got full credit.
- 22% of available points were awarded.
- No partial credit was available.

By far, the most common answer was `e`.

Option	%
e	61
b	**22**
g	10
a, c, f	2 each

24 Question 24. (4 points, 1 each for a-d)

24.1 Setup for Question 24

```
d <- datadist(babies_nomiss)
options(datadist = "d")

model_24 <- ols(baby.wt ~ gestation + mom.age + mom.ht +
                  mom.wt + dad.age + dad.ht + dad.wt,
                data = babies_nomiss, x = TRUE, y = TRUE)
```

24.2 Output for Q24

```
model_24
```

```
Linear Regression Model

 ols(formula = baby.wt ~ gestation + mom.age + mom.ht + mom.wt +
     dad.age + dad.ht + dad.wt, data = babies_nomiss, x = TRUE,
     y = TRUE)

                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
 Obs      701    LR chi2    166.77    R2       0.212
 sigma16.4288    d.f.            7    R2 adj   0.204
 d.f.     693    Pr(> chi2) 0.0000    g        9.172

 Residuals

      Min       1Q   Median       3Q      Max
 -48.6748 -10.5277   0.4026  10.1235  54.9603


           Coef      S.E.    t     Pr(>|t|)
 Intercept -101.9075 23.2918 -4.38 <0.0001
 gestation    0.4503  0.0391 11.52 <0.0001
 mom.age      0.1350  0.1881  0.72 0.4733
 mom.ht       1.2230  0.2852  4.29 <0.0001
 mom.wt       0.0308  0.0343  0.90 0.3693
 dad.age      0.0603  0.1655  0.36 0.7157
 dad.ht      -0.0783  0.2706 -0.29 0.7723
 dad.wt       0.0783  0.0331  2.37 0.0182
```

```
set.seed(43224)
validate(model_24)
```

```
          index.orig training     test optimism index.corrected  n
R-square      0.2117   0.2294   0.1987   0.0308          0.1809 40
MSE         266.8258 258.4144 271.2482 -12.8337        279.6596 40
g             9.1715   9.5738   8.9931   0.5806          8.5909 40
Intercept     0.0000   0.0000   7.5952  -7.5952          7.5952 40
```

```
Slope          1.0000   1.0000   0.9364   0.0636          0.9364 40
```

Consider the model summaries shown above, built from a subset of the data studied in Question 23. Identify the value in the columns that matches the description for each row.

Multiple Choice Grid Columns are 0.224, 0.212, 0.204, 0.199, 0.189

Rows are

a. % of variation explained using this model for the data used to fit this model
b. estimate of % of variation that will be explained by this model in a new data set
c. value that is plotted in a best subsets analysis
d. square of correlation between observed and predicted baby weights

24.3 Answer 24 is `a` is 0.212, `b` is 0.189, `c` is 0.204, `d` is 0.212

- `a` and `d` are definitions of R-squared
- `b` is the index.corrected value for R-squared obtained from the bootstrap validation process
- `c` is adjusted R-squared which we use in the top left plot when we do best subsets analyses

24.4 Q24a Results

- 31/41 students got full credit.
- 76% of available points were awarded.
- The most common incorrect response was 0.224
- No partial credit was available for any part of Q24.

24.5 Q24b Results

- 28/41 students got full credit.
- 68% of available points were awarded.
- The most common incorrect response was 0.204, then 0.224

24.6 Q24c Results

- 30/41 students got full credit.
- 73% of available points were awarded.
- The most common incorrect responses were 0.189 and 0.224

24.7 Q24d Results

- 27/41 students got full credit.
- 66% of available points were awarded.
- The most common incorrect responses were 0.189 and 0.224

25 Question 25. (3 points)

25.1 Setting Up Question 25

25.2 Output for Q25

```
summary(m25)
```

```
             Effects              Response : out

 Factor     Low   High Diff. Effect   S.E.   Lower 0.95 Upper 0.95
 var1        44.0  57  13.0    72.446 5.6005   61.415     83.478
 var2       121.5 183  61.5    22.096 3.5451   15.113     29.079
 var3         0.0   1   1.0   219.130 4.1710  210.920    227.350
 var4 - 1:3   3.0   1    NA    99.475 8.0338   83.651    115.300
 var4 - 2:3   3.0   2    NA    85.137 6.0311   73.258     97.017
 var4 - 4:3   3.0   4    NA  -130.400 6.0528 -142.320   -118.470
 var4 - 5:3   3.0   5    NA  -377.750 5.7160 -389.010   -366.490

Adjusted to: var2=156 var3=0
```

The model `m25` described in the summary above includes four predictors of a continuous outcome (which is measured in days). `var4` takes five possible values, and was included as a factor with levels 1, 2, 3, 4 and 5. If we were to compare two subjects (Jacob, who has var4 = 1, and Olivia, who has var4 = 4) who are the same on all other variables in the model, then which subject would be predicted to have a larger outcome, and by how much?

a. Jacob, by about 100 days
b. Olivia, by about 100 days
c. Jacob, by about 230 days
d. Olivia, by about 230 days
e. It is impossible to tell from the information provided.

25.3 Answer 25 is `c`

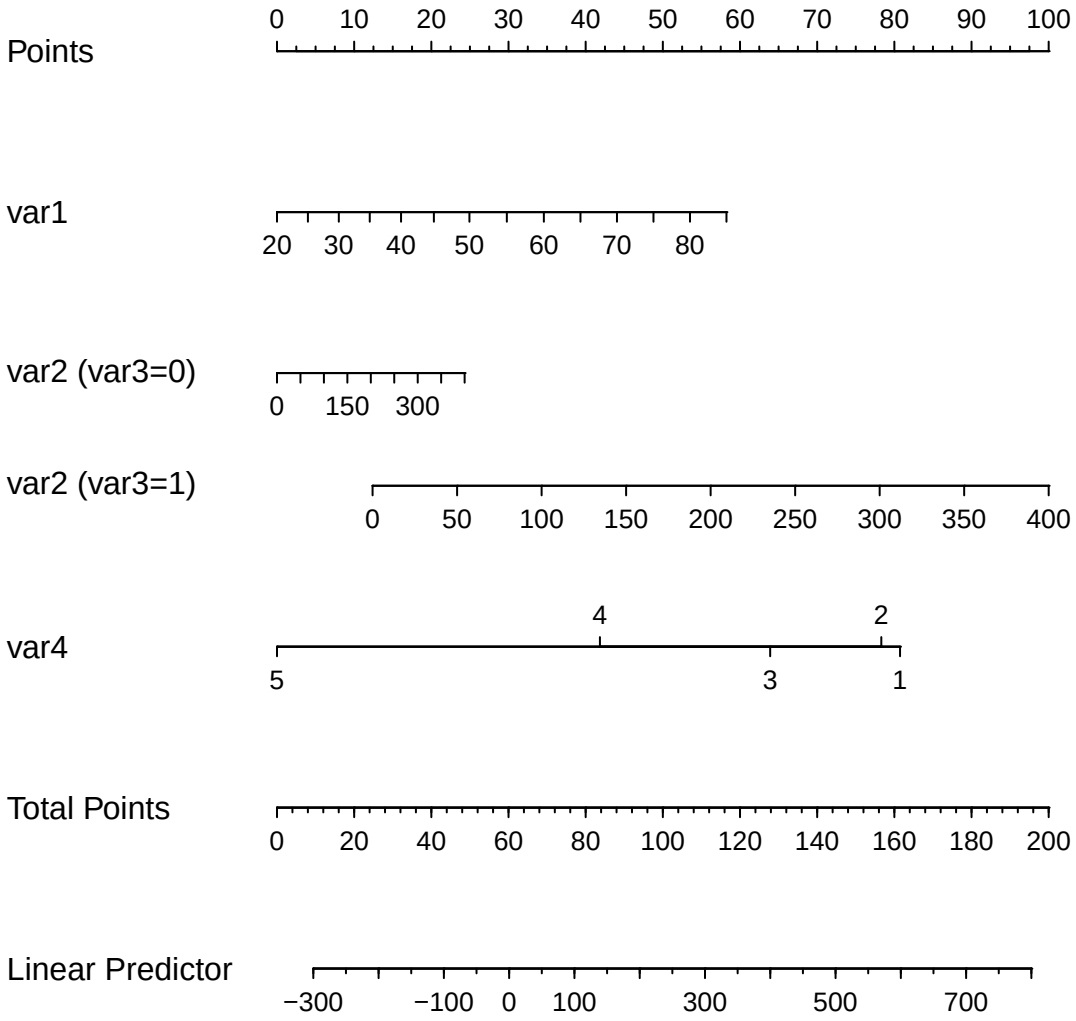
Jacob has `var4` = 1, which means, according to the first row related to `var4` in the summary, that his predicted outcome will be 99.475 days longer than a subject with `var4` = 3 (and otherwise the same value of the predictors). On the other hand, Olivia has `var4` = 4, so her predicted outcome will be 130.4 days shorter than a subject with `var4` = 3 (and the same predictor values otherwise.) So if Jacob and Olivia have the same values of var1, var2 and var3, then Jacob's prediction will be about 230 days (really `130.4 + 99.475` days) longer than Olivia's.

25.4 Q25 Results

- 36/41 students got full credit.
- 88% of available points were awarded.
- No partial credit was available.

26 Question 26. (3 points)

26.1 Nomogram for Q26



Use the nomogram shown above to make a prediction about the outcome variable (which is measured in days) for two subjects. Noah has `var1` = 45, `var2` = 150, `var3` = 0 and `var4` = 4. Sophia has `var3` = 1, but otherwise has the same values of each variable. Which of the following descriptions is most appropriate?

a. Noah and Sophia will have the same predicted outcome.
b. Noah's predicted outcome is longer than Sophia's, but by 50 days or fewer.
c. Noah's predicted outcome is longer than Sophia's, and by more than 50 days.

d. Noah's predicted outcome is shorter than Sophia's, but by 50 days or fewer
e. Noah's predicted outcome is shorter than Sophia's, and by more than 50 days.
f. It is impossible to tell from the information provided.

26.2 Answer 26 is `e`

From the nomogram,

- Noah receives:
  - 20 points for his `var1` of 45
  - since his `var3` = 0, 10 points for his `var2` = 150
  - and 40 points for his `var4` of 4
  - for a total of 70 points, which corresponds to an outcome of a little less than 100 days.
- Sophia receives:
  - 20 points for her `var1` of 45
  - since her `var1` = 0, 40 points for her `var2` = 150
  - and 40 points for her `var4` of 4
  - for a total of 100 points, which corresponds to an outcome of a little more than 200 days.
- So the difference between them must be at least 100 days (and thus, certainly more than 50 days), with Noah having a shorter predicted outcome.

26.3 Q26 Results

- 36/41 students got full credit. This was a different group of 36 than the people who got Q25 right.
- 88% of available points were awarded.
- No partial credit was available.

27 Question 27. (3 points)

In addition to the raw data, name the other three things that should be part of the "data package" that you share, according to Jeff Leek, when you are trying to maximize speed in the analysis of the data.

27.1 Answer 27 is below.

- A tidy data set.
- A code book describing each variable and its values.
- An explicit recipe describing how you went from the raw data to the tidy data set and code book.

27.2 Q27 Results

- 36/41 students got full credit.
- 92% of available points were awarded.
- I gave some partial credit to people focused on README files rather than explicit recipes.

28 Question 28. (2 points)

28.1 Setup for Question 28

The `data_28` set includes 500 observations on 7 potential predictors (labeled `a`, `b`, `c`, `d`, `e`, `f` and `g`) of a continuous `outcome`. Summary statistics follow. Variable `a` falls between 0 and 1, `c` takes on integer values between 1 and 10, and `d` and `f` are binary categorical variables. You have the data available to you on our web site.

```
data_28 %>% skim()
```

```
Skim summary statistics
 n obs: 500
 n variables: 8

Variable type: integer
 variable missing complete   n mean   sd p0 p25 median p75 p100     hist
        c       0      500 500 5.51 2.87  1   3      6   8   10 <U+2587><U+2583><U+2583><U+2583><U+2583>
        d       0      500 500 0.48 0.5   0   0      0   1    1 <U+2587><U+2581><U+2581><U+2581><U+2581>
        f       0      500 500 0.49 0.5   0   0      0   1    1 <U+2587><U+2581><U+2581><U+2581><U+2581>

Variable type: numeric
 variable missing complete   n   mean    sd     p0   p25 median    p75
        a       0      500 500   0.59  0.15  0.072  0.5    0.6    0.69
        b       0      500 500 109.63 15.25 59.61  99.48 110.73 119.79
        e       0      500 500  10.41  1.15  6.54   9.65  10.4   11.19
        g       0      500 500 102.51  5.46 87.85  98.72 102.54 105.91
  outcome       0      500 500  59     7.18 39.54  53.74  58.57  64.85
   p100     hist
   1    <U+2581><U+2581><U+2582><U+2586><U+2587><U+2587><U+2582><U+2581>
 151.22 <U+2581><U+2581><U+2583><U+2587><U+2587><U+2587><U+2582><U+2581>
  13.78 <U+2581><U+2581><U+2583><U+2586><U+2587><U+2586><U+2582><U+2581>
 118.23 <U+2581><U+2582><U+2586><U+2587><U+2587><U+2583><U+2582><U+2581>
  76.6  <U+2581><U+2582><U+2586><U+2587><U+2587><U+2587><U+2583><U+2581>
```

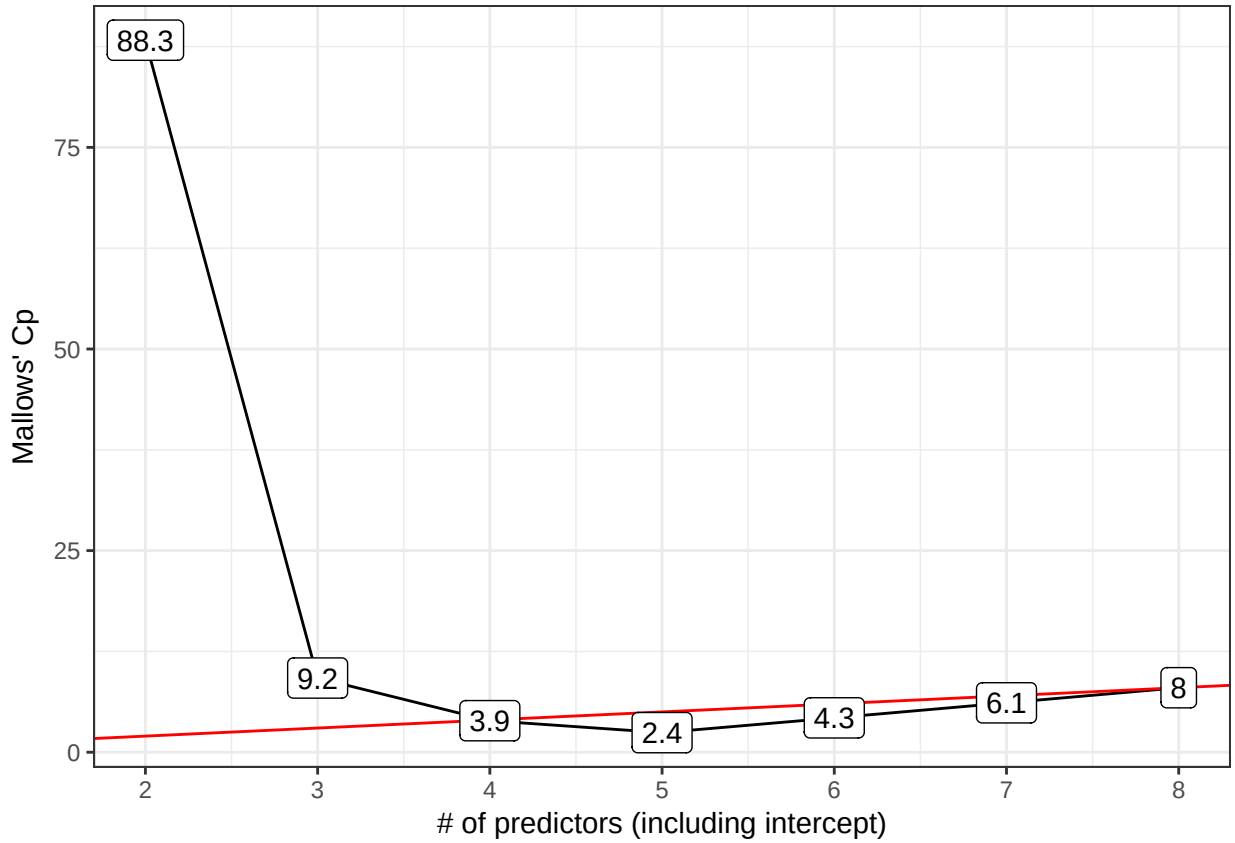
28.2 Output for Q28

```
library(leaps)

preds <- with(data_28, cbind(a, b, c, d, e, f, g))
x1 <- regsubsets(preds, data_28$outcome)
rs <- summary(x1)

rs$which
```

```
  (Intercept)     a     b     c    d     e     f     g
1        TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE FALSE  TRUE TRUE FALSE FALSE FALSE
3        TRUE FALSE FALSE  TRUE TRUE  TRUE FALSE FALSE
4        TRUE FALSE FALSE  TRUE TRUE  TRUE  TRUE FALSE
5        TRUE FALSE FALSE  TRUE TRUE  TRUE  TRUE  TRUE
6        TRUE FALSE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE
7        TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE
```

28.3 Plot for Q28

```
p2
```




```
ggsave("figures/fig28.png")
```

```
Saving 6.5 x 4.5 in image
```

Three different pieces of output for this question were provided above. Variable `a` falls between 0 and 1, `c` takes on integer values between 1 and 10, and `d` and `f` are binary categorical variables. You have the data available to you on our web site. Which variables are included in the model suggested by the Cp plot in the Output for Q28 (part 3 of 3)? Note that the red line is drawn with slope 1 and intercept 0.

a. a
b. b
c. c
d. d
e. e
f. f
g. g
h. None of these.

28.4 Answer 28 is `c`, `d`, and `e`

The model with four predictors including the intercept is suggested by the Cp plot. From the `which` output, this is the model with c, d, and e.

28.5 Q28 Results

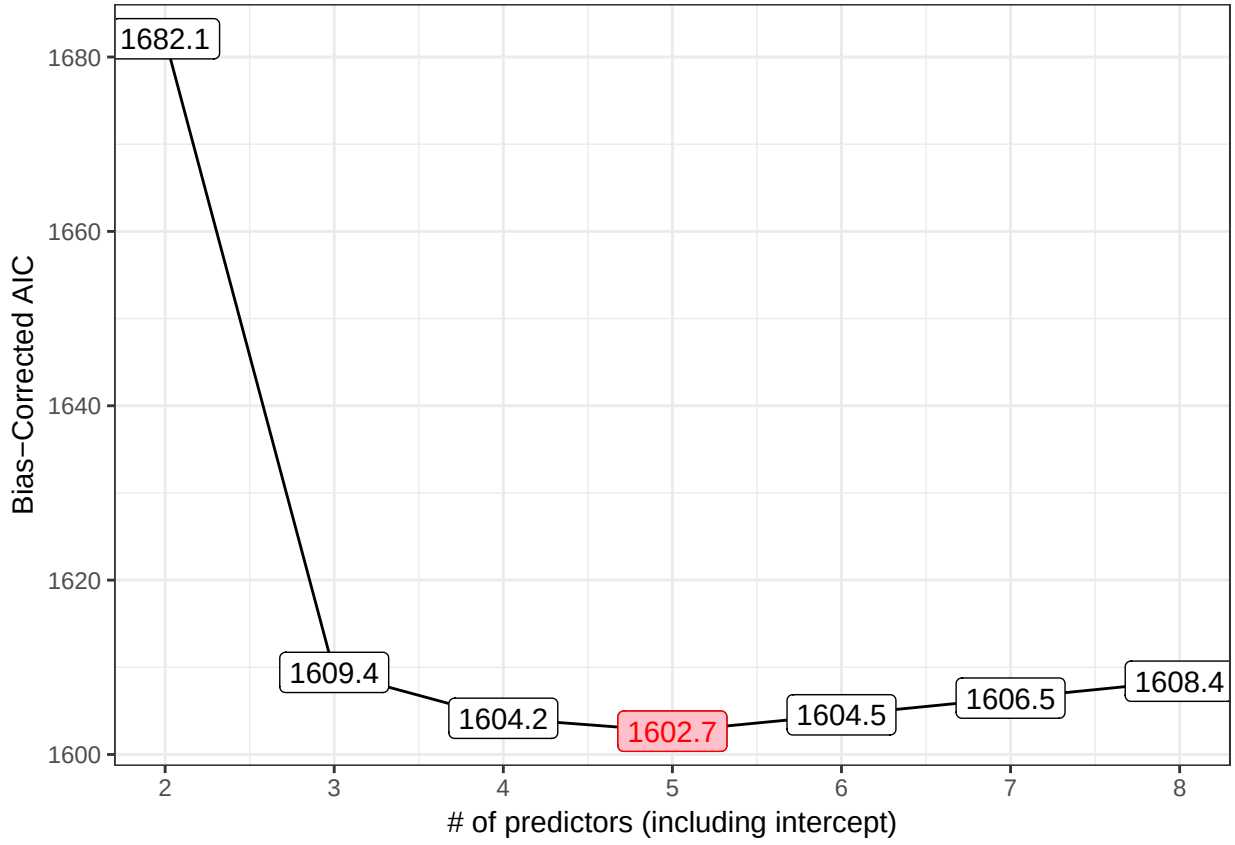
- 22/41 students got full credit.
- 54% of available points were awarded.
- No partial credit was available.

I assumed that people who got this wrong would instead list the model with four predictors NOT including the intercept, in other words, they would list the model with c, d, e and f. Of those who got this wrong, 9/12 made that mistake.

29 Question 29. (2 points)

29.1 Plot for Q29

```
p3
```




```
ggsave("figures/fig29.png")
```

```
Saving 6.5 x 4.5 in image
```

Which variables are included in the model suggested by the bias-corrected AIC plot above?

a. a
b. b
c. c
d. d
e. e
f. f
g. g
h. None of these.

29.2 Answer 29 is `c`, `d`, `e` and `f`

The model with five predictors including the intercept is suggested by the bias-corrected AIC plot. From the `which` output, this is the model with predictors c, d, e and f.

29.3 Q29 Results

- 26/41 students got full credit.
- 63% of available points were awarded.
- No partial credit was available.

I assumed that people who got this wrong would instead list the model with five predictors NOT including the intercept, in other words, they would list the model with c, d, e, f and g. Of those who got Q29 wrong, nearly all made that mistake.

30 Question 30. (3 points)

30.1 Setup for Q30

```
d <- datadist(data_28)
options(datadist = "d")

m28 <- ols(outcome ~ c + d + e,
           data = data_28, x = TRUE, y = TRUE)
m29 <- ols(outcome ~ c + d + e + f,
           data = data_28, x = TRUE, y = TRUE)
```

30.2 Output for Q30

```
set.seed(4320301)
validate(m28)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5264   0.5267  0.5226   0.0041          0.5222 40
MSE          24.3405  23.8618 24.5352  -0.6734         25.0139 40
g             5.9750   5.9175  5.9634  -0.0460          6.0210 40
Intercept     0.0000   0.0000 -0.3345   0.3345         -0.3345 40
Slope         1.0000   1.0000  1.0050  -0.0050          1.0050 40
```

```
set.seed(4320302)
validate(m29)
```

```
          index.orig training    test optimism index.corrected  n
R-square      0.5297   0.5378  0.5260   0.0118          0.5180 40
MSE          24.1697  23.8040 24.3588  -0.5548         24.7245 40
g             5.9960   6.0426  5.9818   0.0607          5.9352 40
Intercept     0.0000   0.0000  0.3539  -0.3539          0.3539 40
Slope         1.0000   1.0000  0.9929   0.0071          0.9929 40
```

Consider the validation summaries provided for the models identified in Question 28 (through the Cp plot) and Question 29 (through the bias-corrected AIC plot.) Compare these two models in terms of validated R-square statistic and validated mean squared error statistics.

a. Model 28 has the better R-square and better MSE, after validation
b. Model 28 has the better R-square and weaker MSE, after validation
c. Model 28 has the weaker R-square and better MSE, after validation
d. Model 28 has the weaker R-square and weaker MSE, after validation
e. It is impossible to tell from the information provided

30.3 Answer 30 is `b`

- Model 28 has the better (larger) validated (index-corrected) R-square, at 0.5222, as compared to 0.5180 for Model 29.
- But Model 28 has the weaker (larger) validated (index-corrected) MSE at 25.0139, as compared to 24.7245 for Model 29.

30.4 Q30 Results

- 33/41 students got full credit.
- 80% of available points were awarded.
- No partial credit was available.

The people who didn't get Q30 right gave a wide range of responses. It's tough to see much of a pattern. `a` and `d` were selected three times, each.

31 Question 31. (3 points)

31.1 Setup for Question 31

```
set.seed(43231)
data_31.raw <- data_frame(
                      x1 = rnorm(1000, 10, 2),
                      x2 = rnorm(1000, 10, 2),
                      x3 = rnorm(1000, 100, 20),
                      x4 = rchisq(1000, 1),
                      x5 = as.integer(rbernoulli(1000, 0.4)),
                      x6 = as.integer(rbernoulli(1000, 0.3)),
                      x7 = rpois(1000, 20),
                      x8 = rpois(1000, 10),
                      x9 = runif(1000, 200, 800),
                      y = rpois(1000, 100))

data_31.na <- map_df(data_31.raw, function(x) {x[sample(c(TRUE, NA),
                                                        prob = c(0.98, 0.02),
                                                        size = length(x),
                                                        replace = TRUE)]})

data_31 <- data_31.na %>%
    mutate(subject = 1:1000)
```

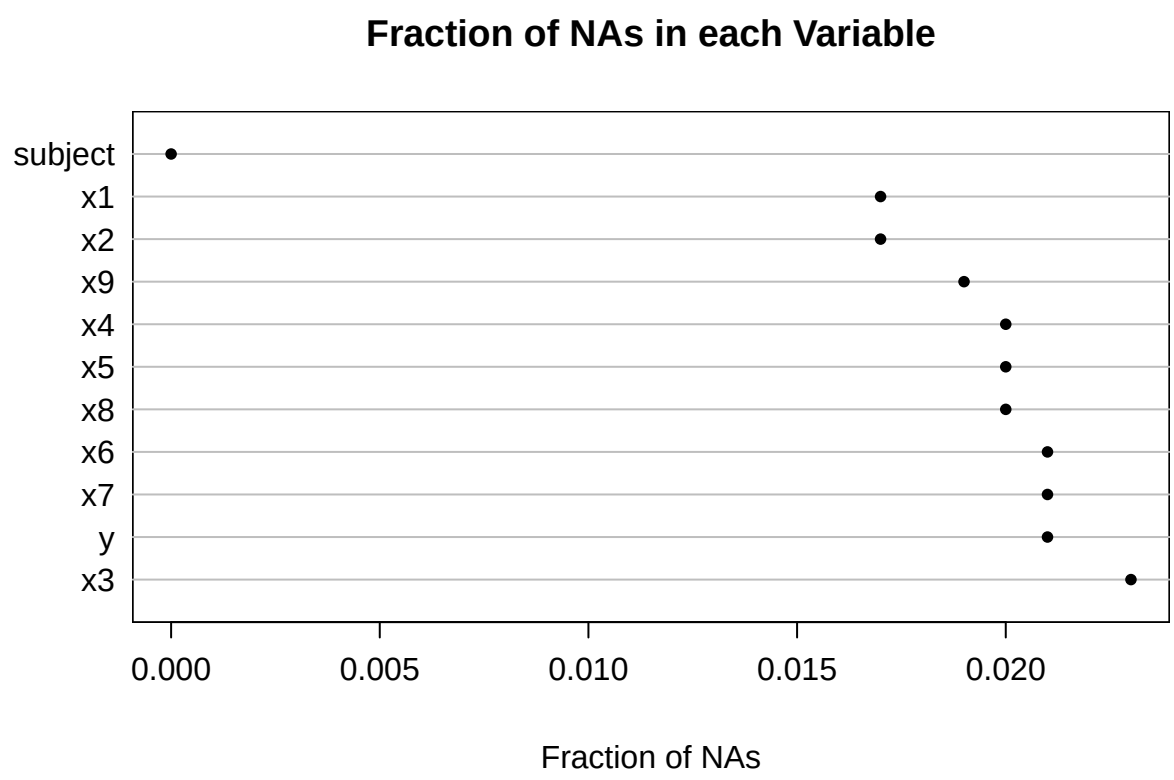
31.2 Output for Q31

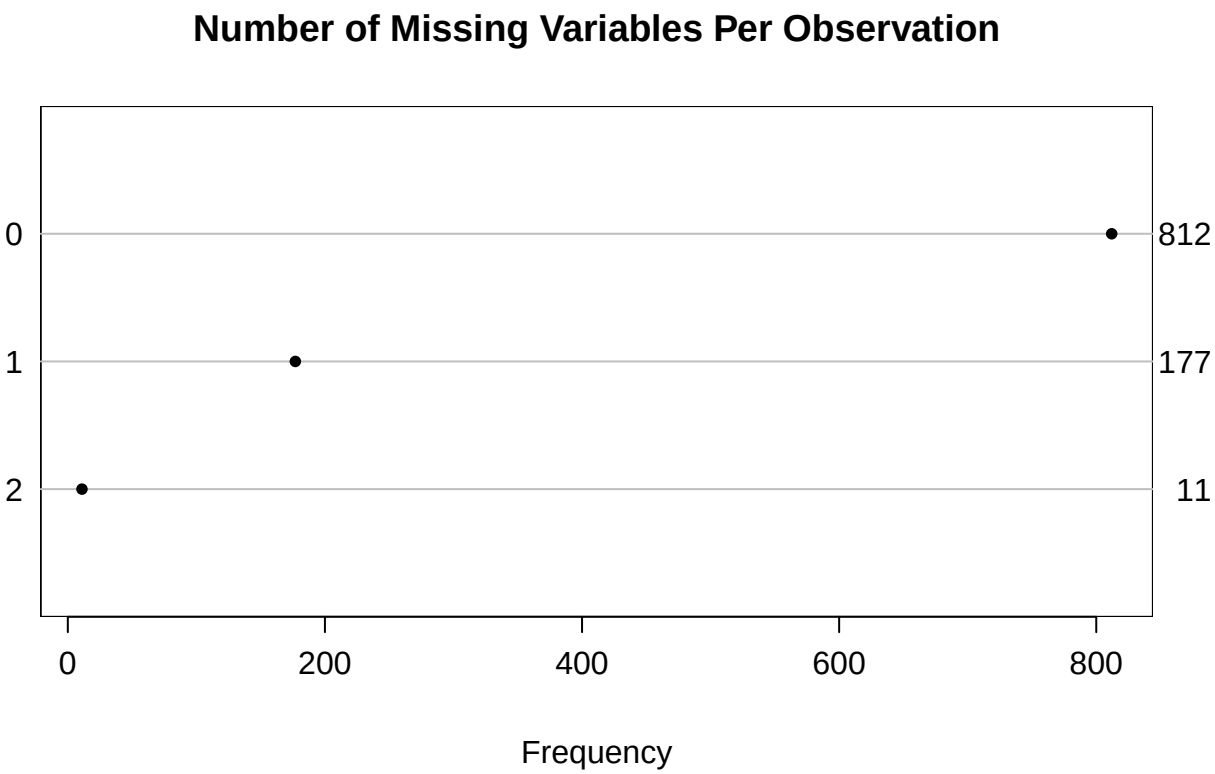
```
skim(data_31)
```

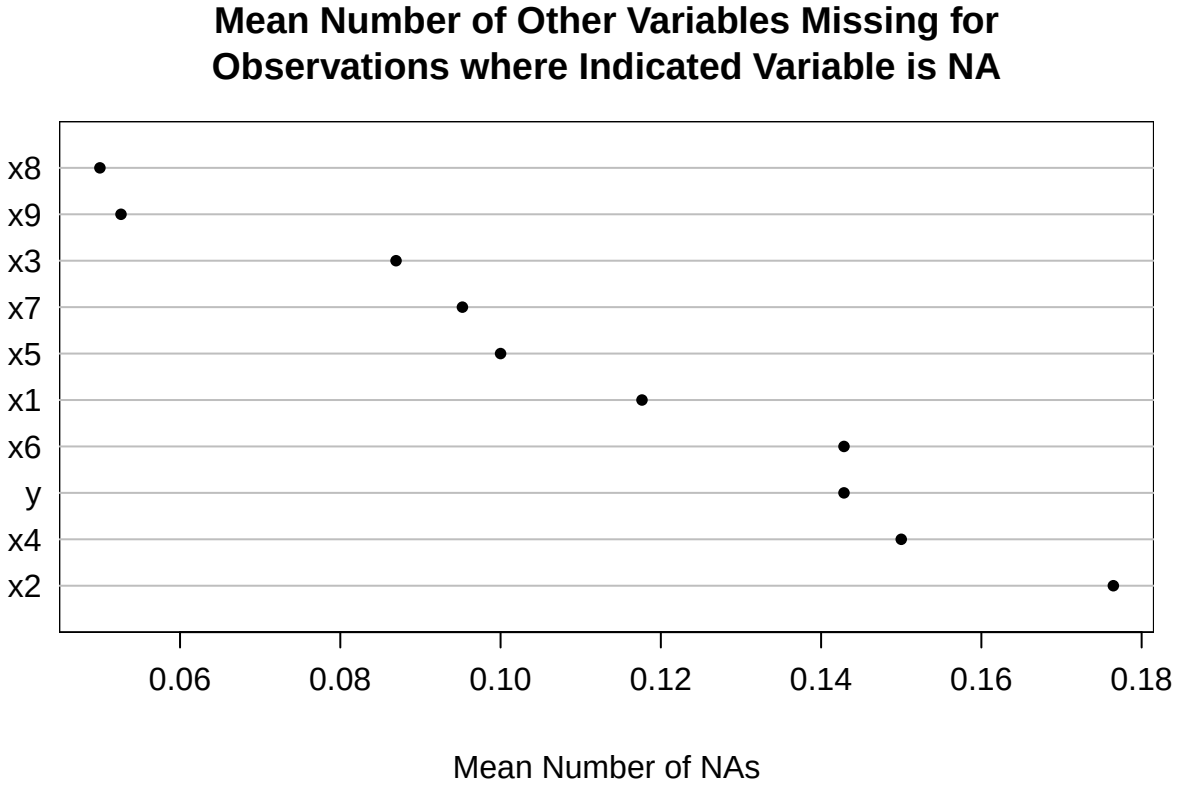
```
Skim summary statistics
 n obs: 1000
 n variables: 11

Variable type: integer
 variable missing complete    n   mean     sd p0    p25 median    p75 p100
  subject       0     1000 1000 500.5  288.82  1 250.75  500.5 750.25 1000
       x5      20      980 1000   0.41   0.49  0   0       0     1       1
       x6      21      979 1000   0.3    0.46  0   0       0     1       1
       x7      21      979 1000  19.85   4.47  7  17      20    23      32
       x8      20      980 1000   9.92   3.15  2   8      10    12      21
        y      21      979 1000  99.88  10.02 70  93     100   106     143
     hist
 <U+2587><U+2587><U+2587><U+2587><U+2587><U+2587><U+2587><U+2587>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2586>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2583>
 <U+2581><U+2582><U+2585><U+2586><U+2587><U+2585><U+2582><U+2581>
 <U+2581><U+2583><U+2587><U+2586><U+2585><U+2582><U+2581><U+2581>
 <U+2581><U+2582><U+2586><U+2587><U+2583><U+2581><U+2581><U+2581>

Variable type: numeric
 variable missing complete    n   mean     sd        p0     p25 median
```

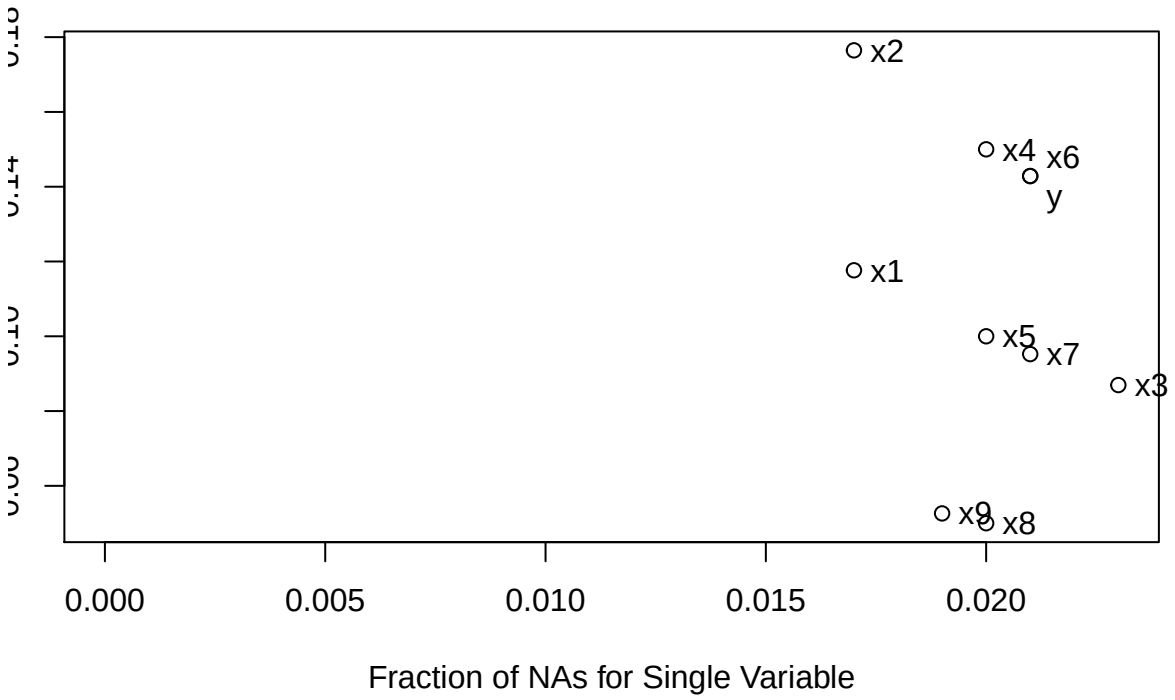
```
      x1      17      983 1000  10.03   1.99   1.89      8.71   10.09
      x2      17      983 1000   9.96   2.05   3.39      8.61    9.99
      x3      23      977 1000 100.61  19.57  30.07     87.33  101.05
      x4      20      980 1000   0.98   1.47   4.3e-10   0.087   0.41
      x9      19      981 1000 497.38 174.54 200.83    345.46  498.43
    p75   p100     hist
  11.38  17.39 <U+2581><U+2581><U+2582><U+2586><U+2587><U+2583><U+2581><U+2581>
  11.36  17.29 <U+2581><U+2581><U+2585><U+2587><U+2587><U+2583><U+2581><U+2581>
 113.47 159.93 <U+2581><U+2581><U+2583><U+2586><U+2587><U+2585><U+2582><U+2581>
   1.23  14.02 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 645.63 799.36 <U+2587><U+2587><U+2587><U+2587><U+2587><U+2587><U+2586><U+2587>
```

```
naplot(naclus(data_31))
```


Fraction of NAs in each Variable
subject
x1
x2
x9
x4
x5
x8
x6
x7
y
x3
0.000
0.005
0.010
0.015
0.020
Fraction of NAs


Number of Missing Variables Per Observation
0
1
2
812
177
11
0
200
400
600
800
Frequency


Mean Number of Other Variables Missing for
Observations where Indicated Variable is NA
x8
x9
x3
x7
x5
x1
x6
y
x4
x2
0.06
0.08
0.10
0.12
0.14
0.16
0.18
Mean Number of NAs



The `data_31` data set contains information on a subject ID code and ten meaningful variables, labeled x1 through x9 and y. Use the output provided to identify the number of observations (out of the total of 1000 subjects) which are missing data on at least one of the ten meaningful variables.

31.3 Answer 31 is 188.

From the second plot in the `naplot` set, 812 of the observations have no missing values. We see that subject is missing in no cases, so the missing values for the other 188 observations (of the total of 1000) must be from the x variables and the y.

31.4 Q31 Results

- 36/41 students got full credit.
- 88% of available points were awarded.
- No partial credit was available.

3 of the students who missed Q31 selected 177, which I think means that they picked only those with exactly 1 NA.

32 Summary

32.1 The Nine "Hardest" Questions

Question	Maximum	# correct	% awarded
23	4	9	22
19	3	16	39
28	2	22	54
15	3	24	59
21d	1	26	63
29	2	26	63
21e	1	27	66
24d	1	27	66
24b	1	28	68

32.2 Results by Respondent

41 people took the quiz. The high score was 99/100, and the median was 85.

1. Tier 1 consists of the 9 people who scored in the 90s, which is definitely worth an A. Congratulations!
2. Tier 2 includes the 8 people who scored 87-89, so that's an A-/B+ grade.
3. Tier 3 includes the 8 people who scored between 84 and 86, which is in the B+/B range.
4. Tier 4 includes the 8 people who scored between 78 and 83, which is still a reasonably strong performance that I'd call a low B.
5. This leaves 8 people with scores below 78, which is below what I'd hoped for.