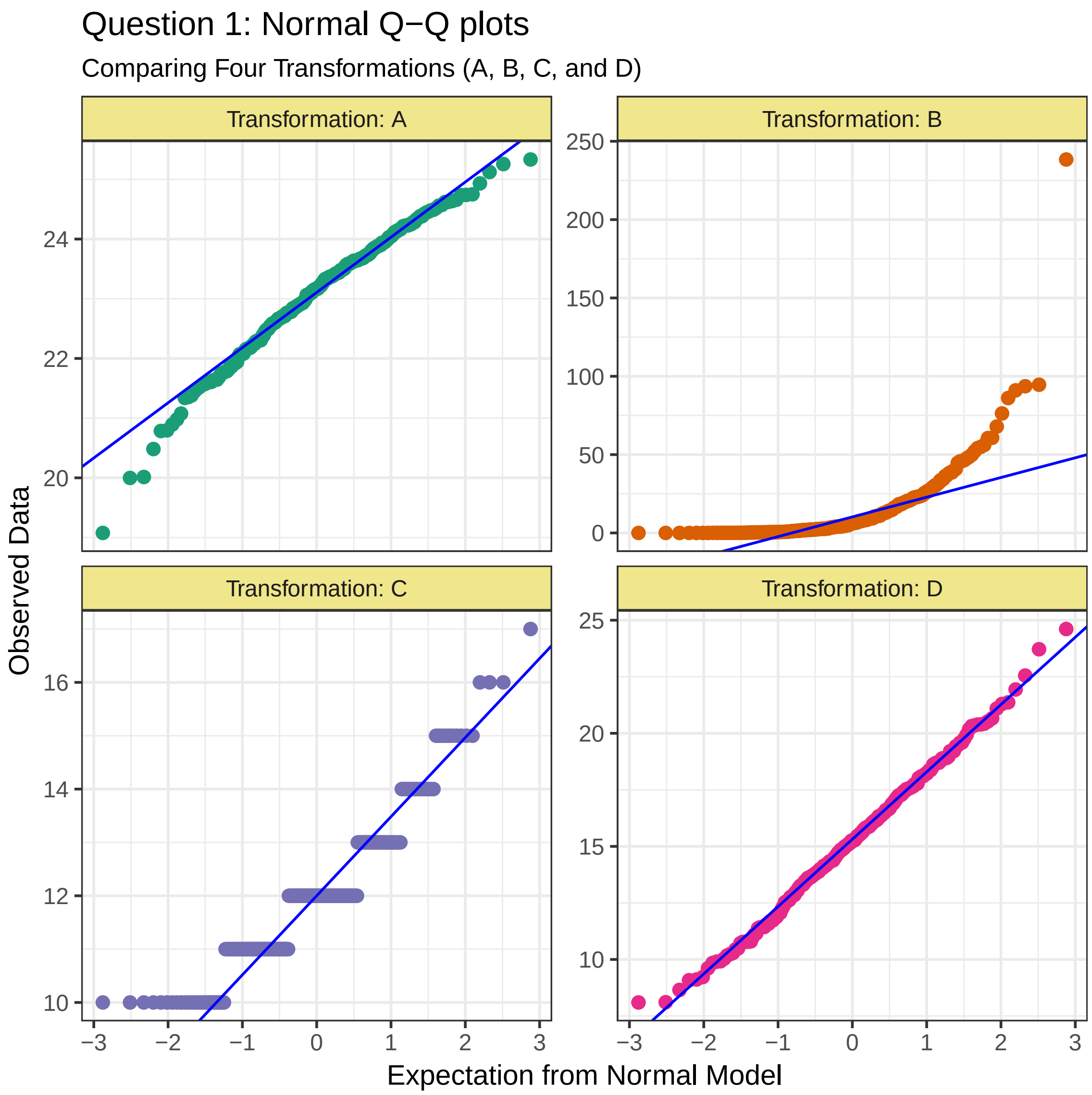
432 Quiz 2

Thomas E. Love

Due 2018-05-01 at Noon. Version: 2018-04-25

1 Question 1

1.1 Display for Question 1

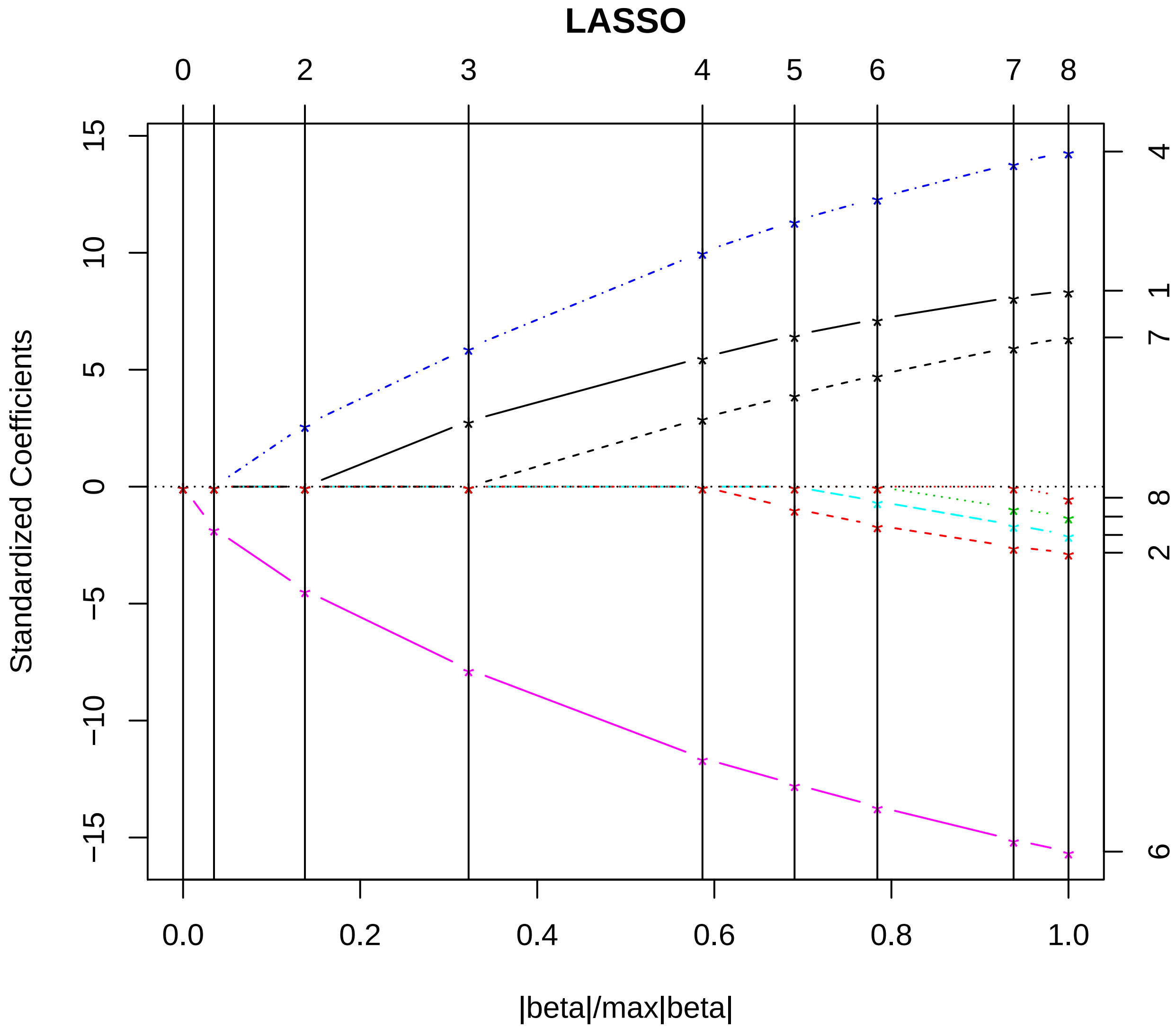


The Display for Question 1 shows normal Q-Q plots of four potential transformations under consideration for modeling a quantitative outcome. Which of the four plots best supports the idea of using a Normal model to fit the data?

```
a. A
b. B
c. C
d. D
e. None of the above.
```

2 Question 2

2.1 Display for Question 2



The Display for Question 2 shows the result of applying the lasso to a data set containing 8 predictors, labeled 1-8 in the plot. If the value of the key fraction to minimize cross-validated mean squared prediction error is 0.42, then how many of the 8 candidate predictors should be included in the model, according to the lasso?

```
a. 1
b. 2
c. 3
d. 4
e. 5
f. 6
g. 7
h. 8
i. It is impossible to tell.
```

3 Question 3

You are part of a study of the effect of a checklist intervention for a surgical procedure on a compliance outcome. Specifically, you have data describing 300 surgical procedures in terms of:

- (a) `compliance` = whether or not the surgical team complied with all guidelines used to formulate the checklist,
- (b) `intervention` = half of the procedures used the checklist and half did not, and
- (c) a quantitative measure of `urgency`, which describes how much of an emergency situation this was (higher values of `urgency` indicate that the surgery was more urgent).

The `urgency` scores ranged from 0 to 100, with median 30. 25% of the surgeries had `urgency` below 20, half were between 20 and 40, and one-quarter were above 40.

We want to build a point and interval estimate for how "the odds of successful compliance comparing surgeries using the intervention to surgeries not using the intervention" were different for surgeries depending on whether the urgency level was 40 as opposed to 20. Which of the following R commands would be part of that work?

```
a. lrm(compliance ~ intervention + urgency, data = dat03, x = TRUE, y = TRUE)
b. glm(compliance ~ intervention + urgency, data = dat03, x = TRUE, y = TRUE)
c. lrm(compliance ~ intervention * urgency, data = dat03, x = TRUE, y = TRUE)
d. glm(compliance ~ intervention * urgency, data = dat03, x = TRUE, y = TRUE)
e. None of these commands would be appropriate.
```

4 Question 4

Suppose you are trying to build a regression model to predict a patient's self-reported overall health (where the available responses are Excellent, Very Good, Good, Fair or Poor) where you want to treat the health assessments as categorical. Which of the following models would be most appropriate?

a. An ordinary least squares model.

b. A Cox proportional hazards model.

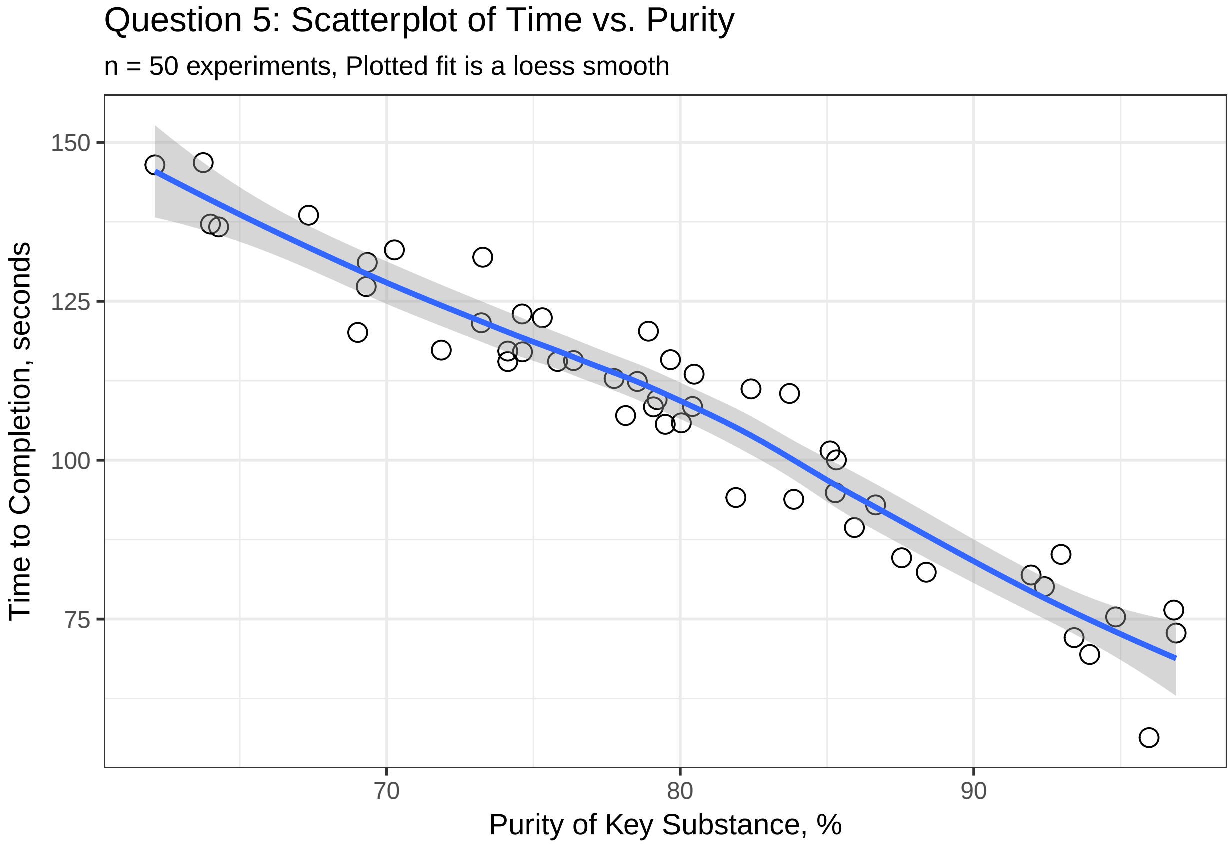
c. A proportional odds logistic regression model.

d. A zero-inflated negative binomial model.

e. None of these models would be appropriate.

5 Question 5

5.1 Display for Question 5

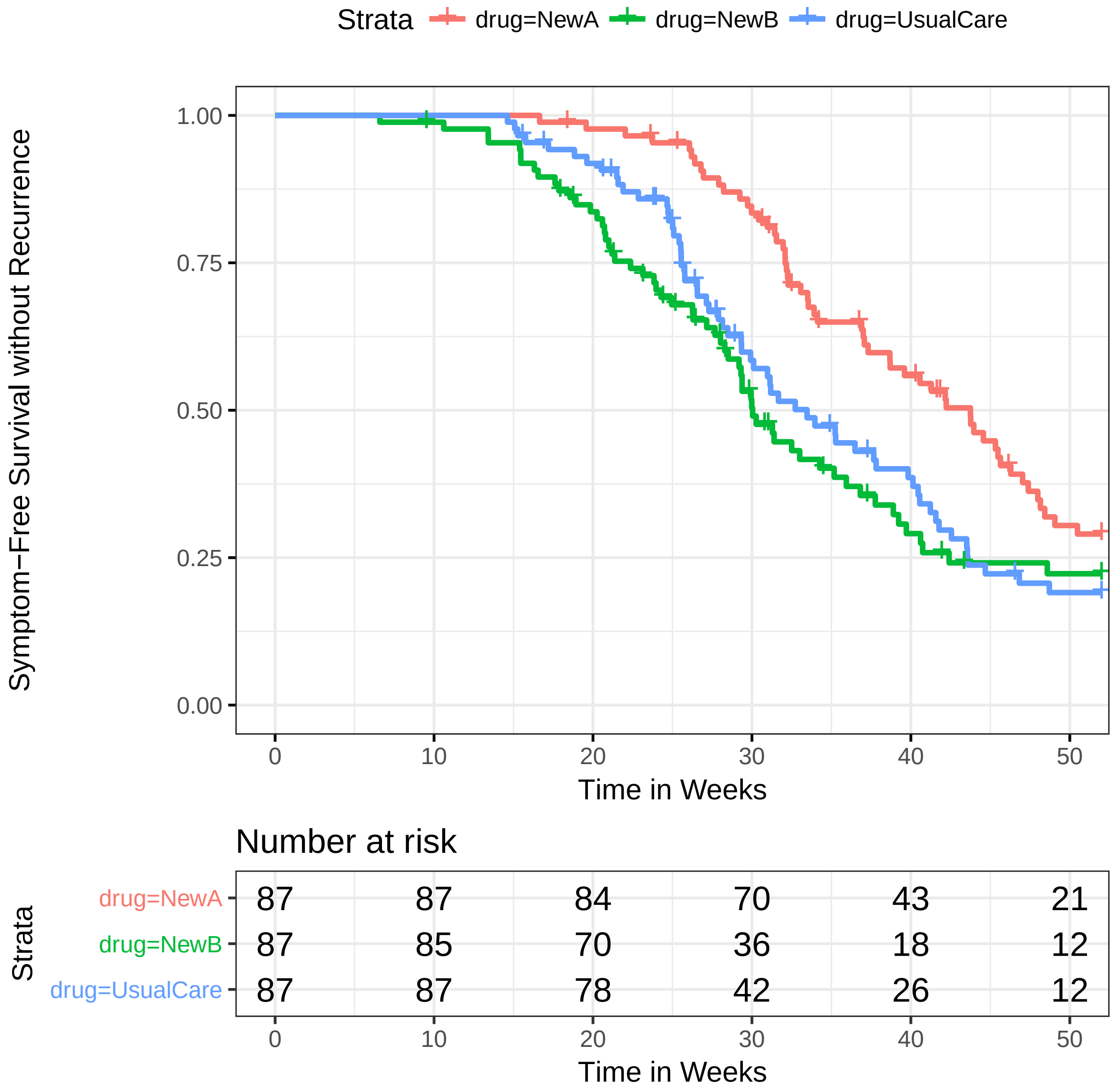


You are fitting a model to describe the time it takes for a chemical reagent to complete a reaction in an experimental setting. You have conducted 50 such experiments, varying the purity level of a key substance. There is variation in the time required, which is associated with the purity, which is measured on a 60-100 scale, since if the substance is not at least 60% pure, the reaction will not happen. The Display for Question 5 shows the scatterplot of time and purity for your 50 experimental runs. Which of the following statements is most true about an simple linear regression model (call it Model 5) fit to represent these data?

```
a. Model 5 will have an R-squared value of about 0.10
b. Model 5 explains between 25% and 50% of the variation in completion time.
c. Model 5 is not helpful, since we should be fitting a Cox model instead.
d. Model 5 explains more than 50% of the variation in completion time.
e. Model 5 fits the data much less well than a model which adds a five-knot
   restricted cubic spline in purity.
```

6 Question 6

6.1 Display 1 for Question 6



Strata	0	10	20	30	40	50
drug=NewA	87	87	84	70	43	21
drug=NewB	87	85	70	36	18	12
drug=UsualCare	87	87	78	42	26	12

6.2 Display 2 for Question 6

```r
print(fit06, print.rmean = TRUE)
```

```
Call: survfit(formula = data06$S ~ data06$drug)

                       n events *rmean *se(rmean) median 0.95LCL 0.95UCL
data06$drug=NewA      87     55   41.0       1.10   43.7    37.3    47.4
data06$drug=NewB      87     58   33.0       1.46   30.0    28.5    35.9
data06$drug=UsualCare 87     60   35.1       1.29   33.5    29.3    40.5
    * restricted mean with upper limit =  52
```

```r
survdiff(data06$S ~ data06$drug)
```

```
Call:
survdiff(formula = data06$S ~ data06$drug)

                       N Observed Expected (O-E)^2/E (O-E)^2/V
data06$drug=NewA      87       55     76.8      6.21     11.31
data06$drug=NewB      87       58     44.5      4.12      5.58
data06$drug=UsualCare 87       60     51.7      1.33      1.91

 Chisq= 11.8  on 2 degrees of freedom, p= 0.00272
```

You are interested in studying the length of time (in weeks) until recurrence of symptoms for adult patients with multiple sclerosis who are treated with new drug A, new drug B or the usual medication. The Kaplan-Meier curve comparing the three drugs is shown in Display 1 for Question 6, and some additional information about the Kaplan-Meier fit is shown in Display 2 for Question 6. Which of the three drugs has the most promising survival curve (longest time to recurrence of symptoms) in these data?

a. Drug A
b. Drug B
c. The Usual Care drug
d. It is impossible to tell from the output provided.

7 Question 7

7.1 Display 1 for Question 7

```
data07
```

```
# A tibble: 100 x 6
      id    x1    x2    x3    x4     y
   <int> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1  104.  102.    0.  103.  20.7
 2     2  110.   NA     1.   NA   21.3
 3     3  104.   97.   NA   123.  30.2
 4     4   94.   97.    0.   NA   21.4
 5     5   94.   NA     1.   97.  28.7
 6     6   99.   93.   NA   105.  23.6
 7     7   NA   100.    0.  119.  22.7
 8     8  106.   NA     0.   NA   21.9
 9     9   80.  102.    1.  102.  19.8
10    10  113.   95.    0.   99.  14.1
# ... with 90 more rows
```

7.2 Display 2 for Question 7

Chunk I

```
set.seed(432)
data07_train1 <- data07 %>%
    sample_frac(size = 0.80, replace = FALSE) %>%
    drop_na

data07_test1 <- data07 %>%
    sample_frac(size = 0.20, replace = TRUE) %>%
    drop_na
```

Chunk II

```r
set.seed(432)
data07_noNA <- data07 %>%
    filter(complete.cases(.))

data07_train2 <- data07_noNA %>%
    sample_frac(size = 0.80, replace = FALSE)

data07_test2 <-
    dplyr::anti_join(data07_noNA, data07_train2, by = "id")
```

Chunk III

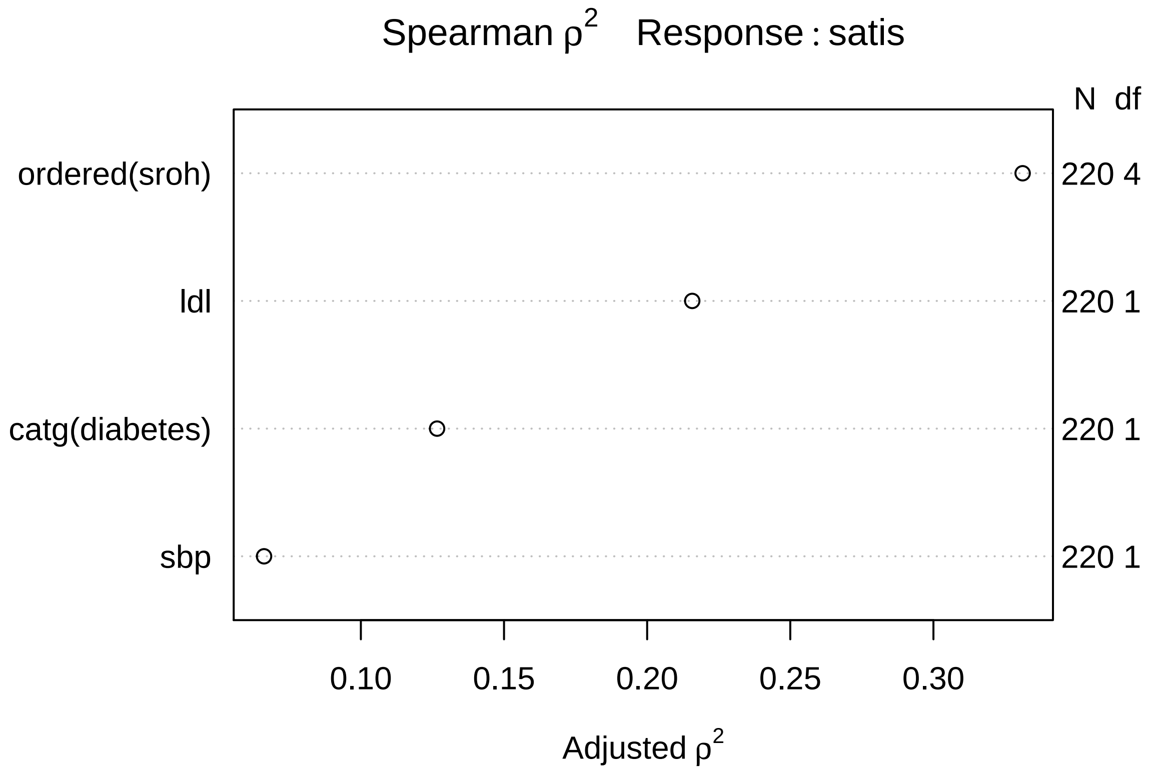
```r
set.seed(432)
data07_noNA3 <- data07 %>%
    drop_na %>%
    mutate(rand = runif(n(), min = 0, max = 1))

data07_train3 <- data07_noNA3 %>%
    slice(rand < quantile(rand, 0.8))

data07_test3 <- data07_noNA3 %>%
    slice(rand >= quantile(rand, 0.8))
```

Given the data set `data07` as shown in Display 1 for Question 7, suppose you want to remove all rows containing missing values, then create a training sample containing 80% of the rows without missing data, and a test sample containing the other 20% of the values after missingness is removed. Which of the chunks of R commands shown in Display 2 for Question 7 will accomplish this?

```
a. Chunk I only.
b. Chunk II only.
c. Chunk III only.
d. Chunks I and II.
e. Chunks I and III.
f. Chunks II and III.
g. All three Chunks.
h. None of these Chunks.
```

8 Question 8

8.1 Display for Question 8



Suppose you plan to fit a model to predict the level of a patient's satisfaction (`satis`, measured on a 0-100 scale, where `satis` = 100 indicates that a patient is extremely satisfied) with their health care, using a sample of 220 subjects, gathered in the `data08` data set.

For each subject, you also have information on

- their systolic blood pressure (`sbp`, in mm Hg),
- their LDL cholesterol (`ldl`, in mg/dl),
- whether or not they have a diabetes diagnosis (`diabetes` = 1 if they do, 0 otherwise) and
- their self-reported overall health (`sroh`) status (Excellent, Very Good, Good, Fair or Poor).

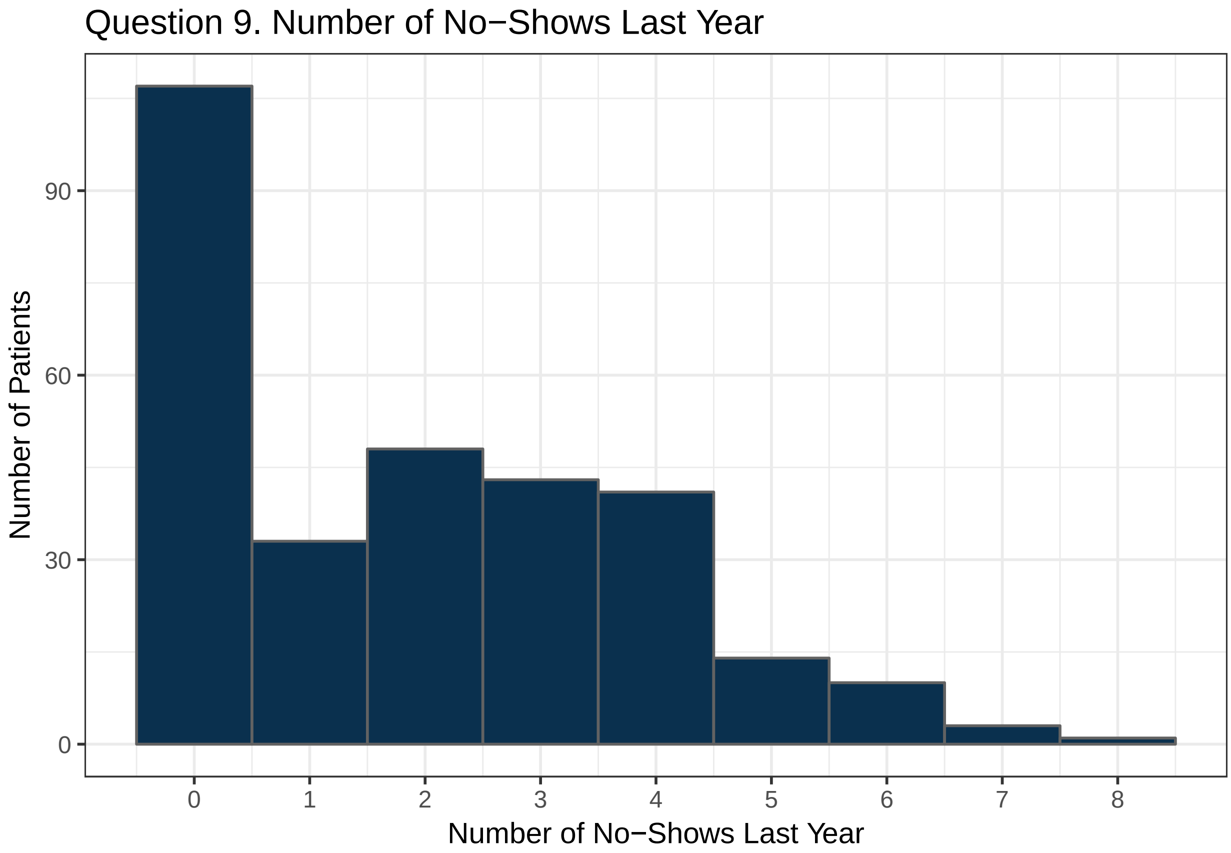
The Display for Question 8 shows a Spearman rho-squared plot for these subjects. Assuming you wish to include all of the main effects for these predictors in your model, and you can afford to add an additional four degrees of freedom to the model, which of the following augmentations to a "main effects" models is the best choice?

```
a. An `ols` model adding a restricted cubic spline in `ldl` with 5 knots.
```

```
b. A `lrm` model adding a restricted cubic spline in `ldl` with 5 knots.
c. An `ols` model including the interactions of `diabetes` with
    both `sbp`, and `ldl`.
d. A `lrm` model including the interactions of `diabetes` with
    both `sbp`, and `ldl`.
e. An `ols` model including the interaction of `ldl` and `sroh`.
f. A `lrm` model including the interaction of `ldl` and `sroh`.
g. It is impossible to tell which of these options is best.
```

9 Question 9

9.1 Display for Question 9



Suppose you are trying to build a regression model to predict `noshow`, the number of times a patient will "no show" an appointment for medical care in the next 12 months, on the basis of several characteristics related to their health, demographics, and satisfaction levels with prior visits. The `noshow` data on 300 patients from last year are visualized in the Display for Question 9. Which of the following models is most likely to be appropriate?

```
a. A binary logistic regression model.
```

b. A zero-inflated Poisson model.
c. A multinomial logistic regression model.
d. A Cox proportional-hazards model.
e. A proportional odds logistic regression.
f. None of these models will be appropriate.

10 Question 10

10.1 Display for Question 10

```
data10
```

```
# A tibble: 120 x 6
      x1    x2    x3    x4      x5      y
   <dbl> <int> <dbl> <dbl>   <dbl>  <dbl>
 1   78.    12    1.   18.   72.9   11.7
 2   94.    11    1.    4.   93.2   11.2
 3  115.     5    0.    4.    5.20 -10.1
 4   43.    14    1.    4.   45.6  -10.1
 5  107.     1    1.    4.  112.     3.70
 6   75.     7    0.    3.   -3.60 -33.8
 7   99.     9    1.    6.   95.1   25.5
 8   66.     8    0.    3.    5.60 -34.6
 9  107.     8    1.    1.  111.    -3.40
10   68.     6    1.    1.   71.4    5.50
# ... with 110 more rows
```

The `data10.csv` data set (which will be used in Questions 10-12) is available to you on the course web site. That data set contains a quantitative outcome, `y`, and five candidate predictors, named `x1` through `x5`. Fit a linear model containing the main effects of all five predictors, and then use stepwise regression (backwards elimination, using AIC as the criterion) to select a new model. Which of the following sets of predictors does the stepwise approach suggest?

a. x1, x2, x3, x4 and x5
b. x1, x2, x4 and x5
c. x2, x4 and x5
d. x2 and x5
e. x5 alone
f. None of these

11 Question 11

Following on from Question 10, fit the model suggested by the stepwise regression (that you identified in Question 10) to the full data set of 120 observations, and study the resulting model diagnostics. Which of the following problems would you regard as substantial and important for this regression model in this sample?

```
a. Non-linearity
b. Collinearity
c. Non-Normality of errors
d. Heteroscedasticity of errors
e. None of the above
```

12 Question 12

12.1 Display for Question 12

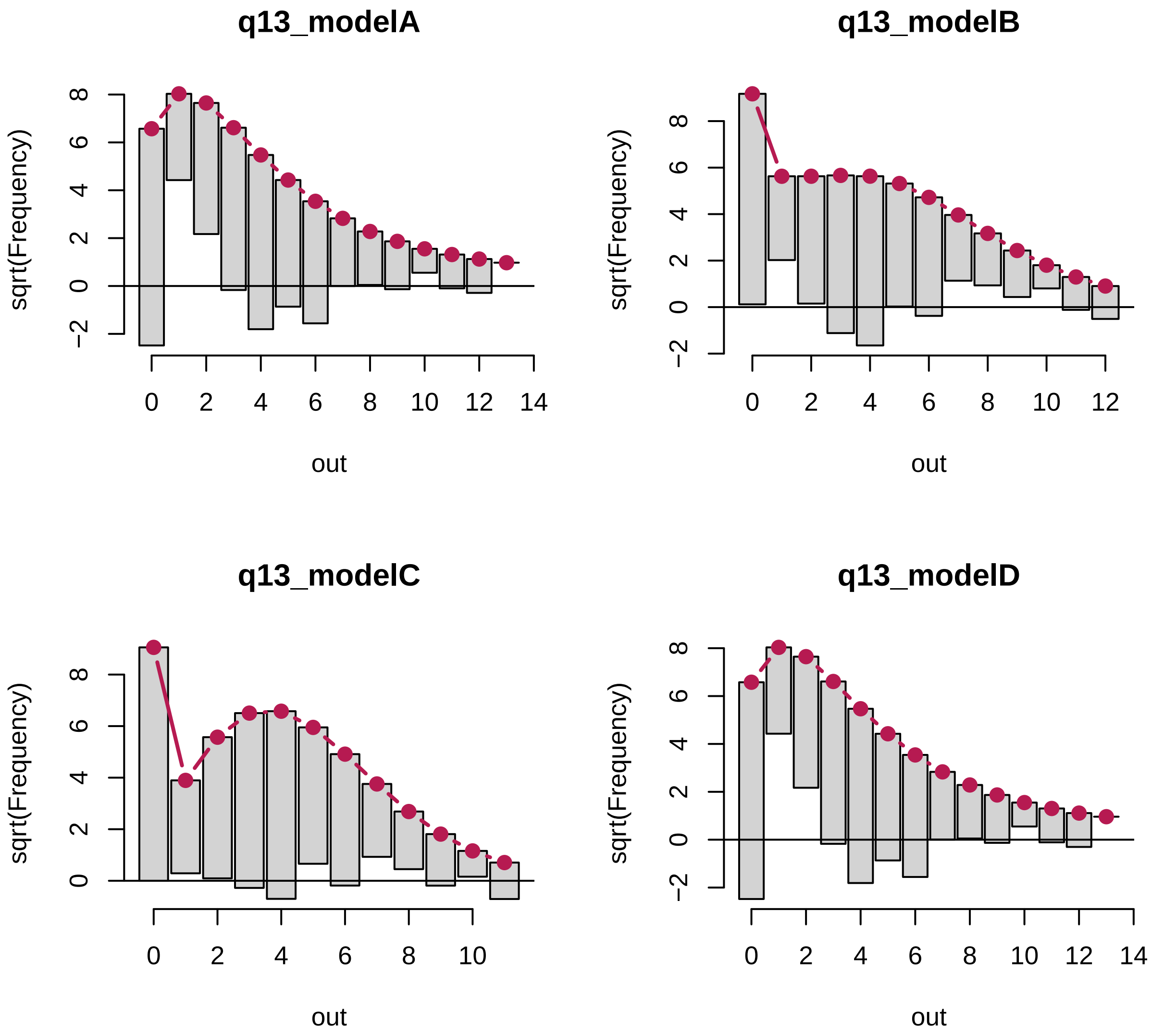
```
set.seed(432)
q12_models <- data10 %>%
    modelr::crossv_kfold(k = 10) %>%
```

Use 10-fold cross-validation to evaluate the model you fit in Question 11. Note that the Display for Question 12 shows the first three lines of my solution, which should be a good way to get started. Set your seed to be `432`, as I have done. What is the root mean squared prediction error for that model, according to this approach?

```
a. Above 7 but less than 8.
b. Above 8 but less than 9.
c. Above 9 but less than 10.
d. Above 10 but less than 11.
e. None of the above.
```

13 Question 13

13.1 Display for Question 13



The Display for Question 13 shows four rootograms, using four different count regression models to fit the same outcome, which is named `out`. Which model (A, B, C, D) shows the best fit to the data?

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

14 Question 14

Suppose you are trying to build a regression model to predict whether or not a patient hospitalized with heart failure will need to return to the hospital in the 30 days after they are released. You gather a series of predictors that should be useful. Which of the following models would be most appropriate?

```
a. An ordinary least squares model.
b. A Cox proportional hazards model.
c. A multinomial logit model.
d. A binary logistic regression model.
e. None of these models would be appropriate.
```

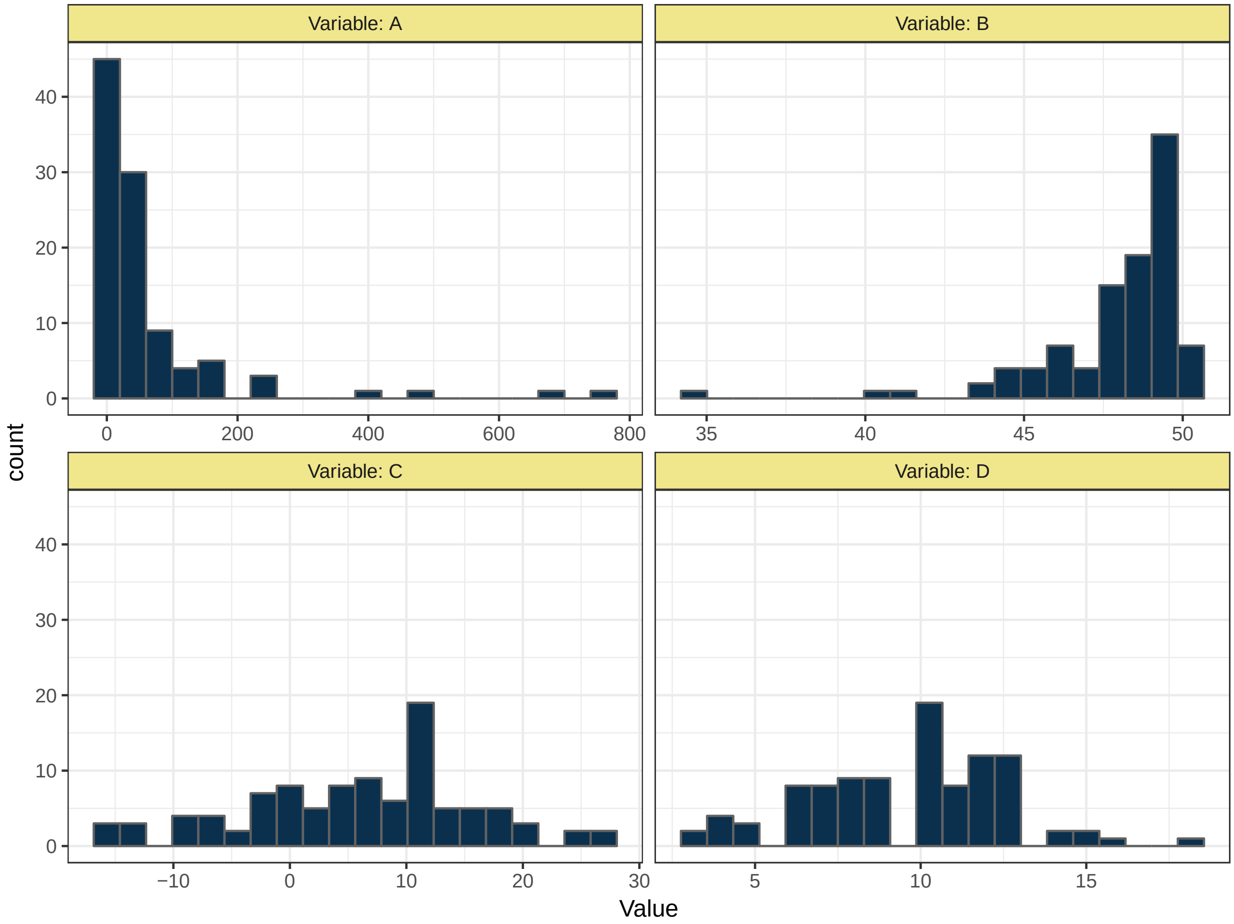
15 Question 15

Suppose you want to build a plot to describe the relationship between a child's score on a measure of depression, and the child's favorite color. The depression scores emerge from a questionnaire, whose final score is standardized to have a mean of 50 and standard deviation of 10 across a prior large and representative sample of children. In your sample, the childrens' favorite colors are easily collapsed into four main categories. Which of the following `geoms` in `ggplot2` is most likely to be helpful?

```
a. `geom_violin`
b. `geom_rug`
c. `geom_point`
d. `geom_histogram`
e. `geom_qq`
```

16 Question 16

16.1 Display for Question 16



Which of the four variables plotted in the Display for Question 16 can be most effectively modeled by applying a Normal model to its logarithm?

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

17 Question 17

17.1 Display for Question 17

```
set.seed(432171); validate(m17)
```

```
          index.orig training   test optimism index.corrected  n
Dxy           0.3588   0.3884 0.3144   0.0740          0.2848 40
R2            0.1359   0.1662 0.1080   0.0581          0.0778 40
Intercept     0.0000   0.0000 0.0138  -0.0138          0.0138 40
Slope         1.0000   1.0000 0.8044   0.1956          0.8044 40
Emax          0.0000   0.0000 0.0508   0.0508          0.0508 40
D             0.0975   0.1244 0.0747   0.0498          0.0477 40
U            -0.0200  -0.0200 0.0044  -0.0244          0.0044 40
Q             0.1175   0.1444 0.0703   0.0741          0.0434 40
B             0.2245   0.2170 0.2345  -0.0176          0.2420 40
g             0.8078   0.9077 0.6989   0.2088          0.5990 40
gp            0.1865   0.1999 0.1636   0.0363          0.1502 40
```

Based on the Display for Question 17, which of the following descriptions is the best choice for specifying the likely effectiveness of this logistic regression model in a new data set?

```
a. Area under the ROC curve will be about 0.28, Nagelkerke R-square about 0.08
b. Area under the ROC curve will be about 0.31, Nagelkerke R-square about 0.11
c. Area under the ROC curve will be about 0.36, Nagelkerke R-square about 0.14
d. Area under the ROC curve will be about 0.39, Nagelkerke R-square about 0.17
e. Area under the ROC curve will be about 0.64, Nagelkerke R-square about 0.08
f. Area under the ROC curve will be about 0.66, Nagelkerke R-square about 0.11
g. Area under the ROC curve will be about 0.68, Nagelkerke R-square about 0.14
h. Area under the ROC curve will be about 0.69, Nagelkerke R-square about 0.17
```

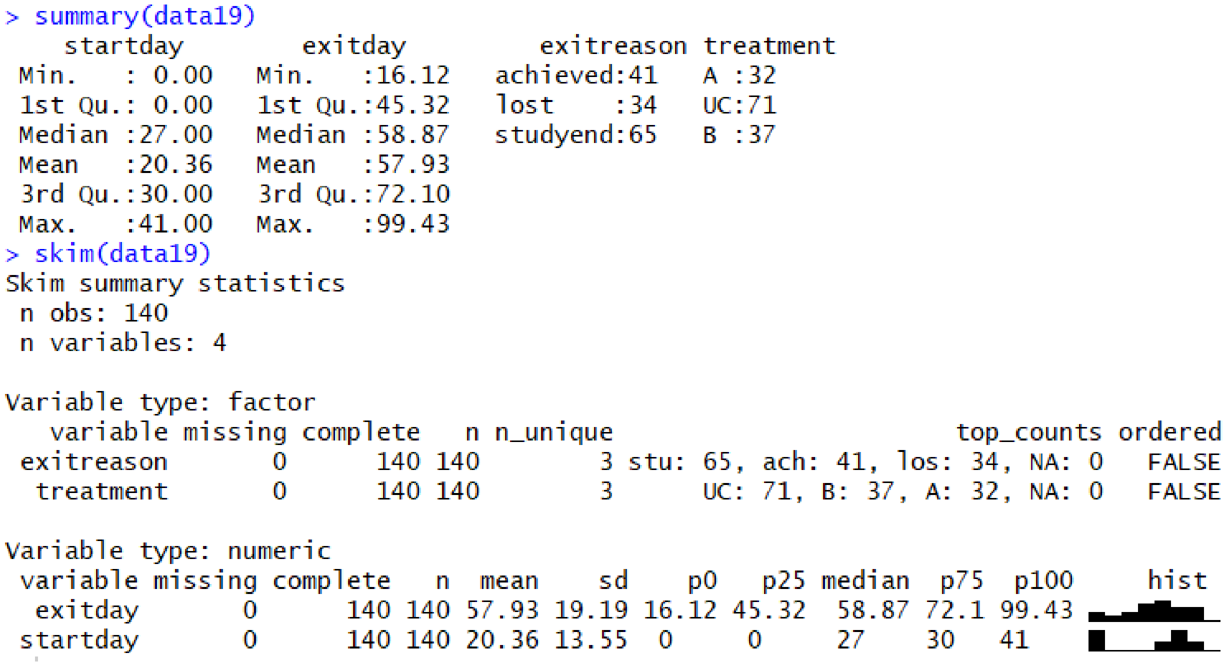
18 Question 18

Suppose you have a data set called `data18` which contains a variable called `preference` which specifies whether the subject preferred option A, B, C, D, or E. Suppose option C is most expensive, followed by options A and then B, and that options D and E are of about the same cost, which is much lower than the other options. Further, suppose that option E was rarely chosen, and you have decided to collapse it together with option D. If you want to develop a plot that will show the `preferences` after collapsing D and E, in order of their costs, on your x axis, then which of the following functions from the `forcats` package would be helpful in doing so?

```
a. `fct_reorder`
b. `fct_collapse` and `fct_relevel`
c. `fct_recode` and `fct_lump`
d. `fct_count` and `fct_relabel`
e. `fct_drop`
```

19 Question 19

19.1 Display 1 for Question 19



```
> summary(data19)
    startday         exitday            exitreason treatment
 Min.   : 0.00   Min.   :16.12   achieved:41   A :32    
 1st Qu.: 0.00   1st Qu.:45.32   lost    :34   UC:71    
 Median :27.00   Median :58.87   studyend:65   B :37    
 Mean   :20.36   Mean   :57.93                          
 3rd Qu.:30.00   3rd Qu.:72.10                          
 Max.   :41.00   Max.   :99.43                          
> skim(data19)
Skim summary statistics
 n obs: 140 
 n variables: 4 

Variable type: factor 
   variable missing complete   n n_unique                       top_counts ordered
 exitreason       0      140 140        3 stu: 65, ach: 41, los: 34, NA: 0   FALSE
  treatment       0      140 140        3      UC: 71, B: 37, A: 32, NA: 0   FALSE

Variable type: numeric 
 variable missing complete   n  mean    sd    p0   p25 median  p75  p100     hist
  exitday       0      140 140 57.93 19.19 16.12 45.32  58.87 72.1 99.43 
 startday       0      140 140 20.36 13.55  0     0     27    30   41    
```

19.2 Display 2 for Question 19

Chunk I for Question 19

```r
survdiff(Surv(time = data19$exitday, event = data19$exitreason) ~ treatment)
```

Chunk II

```r
data19$S = Surv(time = data19$exitday - data19$startday,
                event = data19$exitreason %in% c("lost", "studyend"))
survdiff(S ~ treatment, data = data19)
```

Chunk III

```r
data19$S = Surv(time = data19$exitday - data19$startday,
                event = data19$exitreason == "achieved")
survdiff(S ~ treatment, data = data19)
```

Display 1 for Question 19 shows a summary of the `data19` data. The study was arranged to begin on day 0, and we have available the `startday` and `exitday` for each subject in a tobacco cessation study, comparing three `treatments` (called A, B and usual care). The `exitreason` variable shows the reason why each subject exited the study, either because they achieved the outcome (`achieved`), they stopped coming to appointments and were thus lost to follow up (`lost`), or because the study ended (`studyend`). Suppose you want to add a survival object called `S` to the `data19` data, and want to treat the subjects who did not achieve the outcome as being right-censored, then fit a log rank test to compare the three `treatment` groups in terms of that survival object. Which of the chunks of R code shown in Display 2 for Question 19 will accomplish this?

```
a. Chunk I only.
b. Chunk II only.
c. Chunk III only.
d. Chunks I and II.
e. Chunks I and III.
f. Chunks II and III.
g. All three Chunks.
h. None of these Chunks.
```

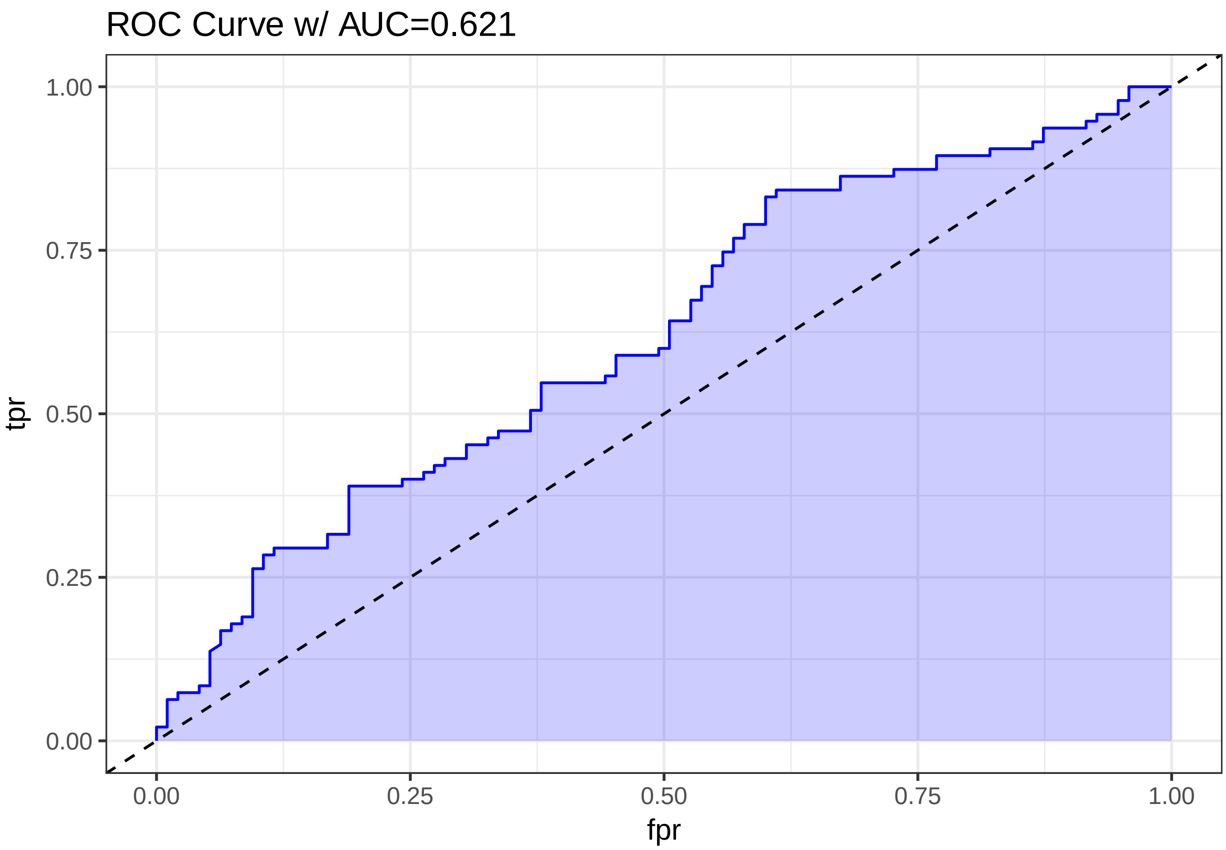
20 Question 20

20.1 Display 1 for Question 20

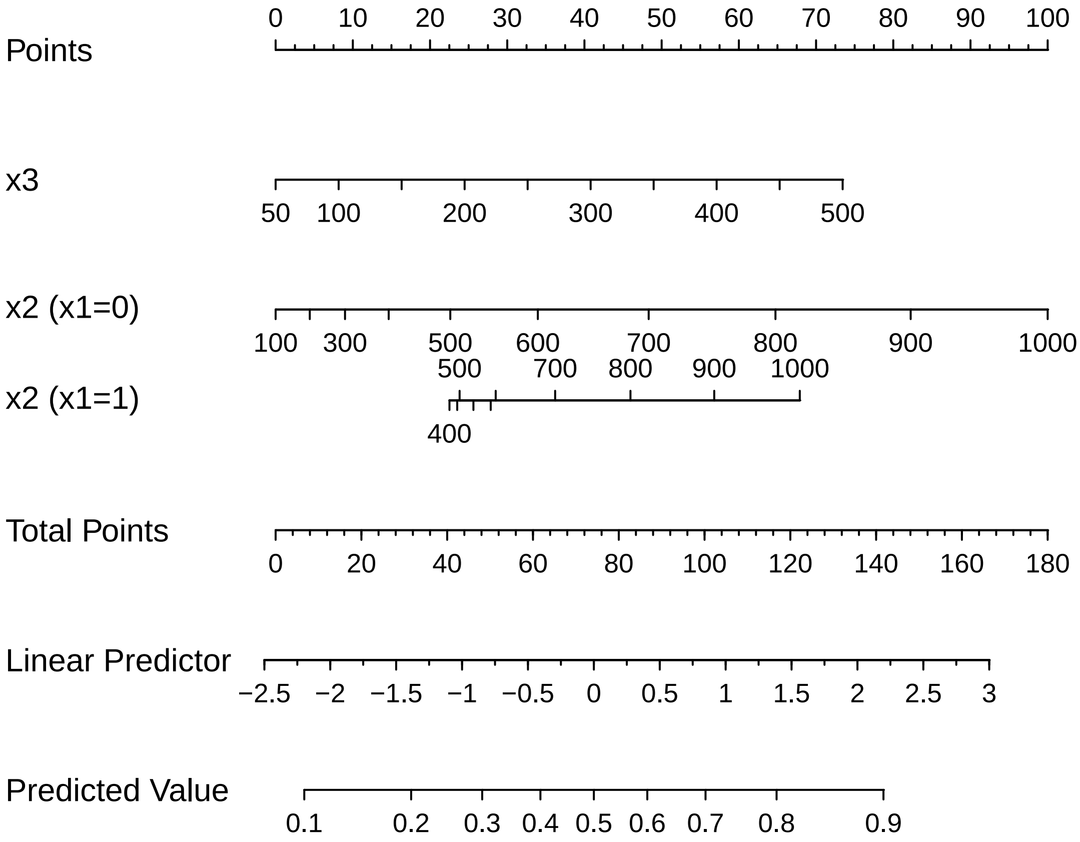
```
Logistic Regression Model

 lrm(formula = outcome ~ catg(x1) + x3 + rcs(x2, 3) + x1 %ia%
     x2, data = data20, x = TRUE, y = TRUE)
```

```
                            Model Likelihood      Discrimination      Rank Discrim.
                               Ratio Test             Indexes            Indexes
Obs              190      LR chi2      37.35      R2         0.238    C        0.741
 0                95      d.f.             5      g          1.153    Dxy      0.481
 1                95      Pr(> chi2) <0.0001      gr         3.169    gamma    0.481
max |deriv| 6e-05                                 gp         0.245    tau-a    0.242
                                                  Brier      0.206

          Coef    S.E.   Wald Z Pr(>|Z|)
Intercept -2.8241 0.8965 -3.15  0.0016
x1=1       1.1235 0.9323  1.21  0.2282
x3         0.0053 0.0018  2.94  0.0033
x2         0.0014 0.0018  0.81  0.4199
x2'        0.0033 0.0025  1.29  0.1964
x1 * x2   -0.0022 0.0016 -1.34  0.1791
```

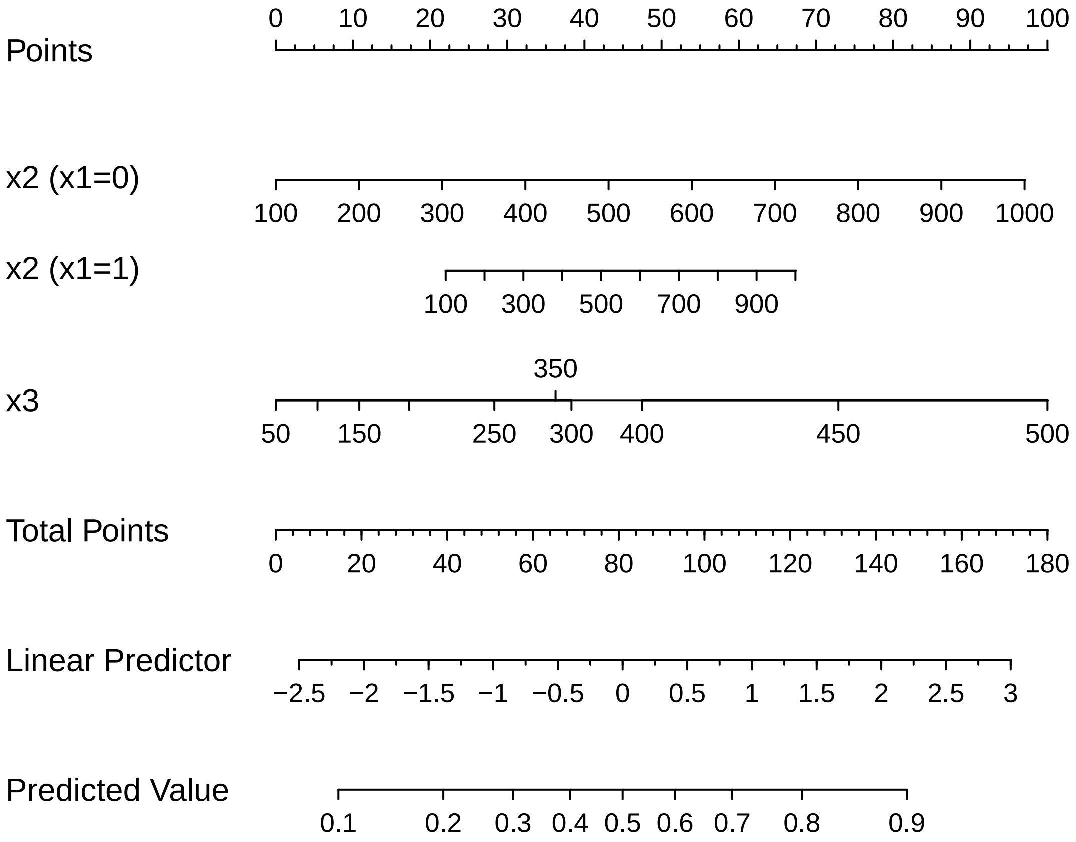
20.2 Plot A for Question 20



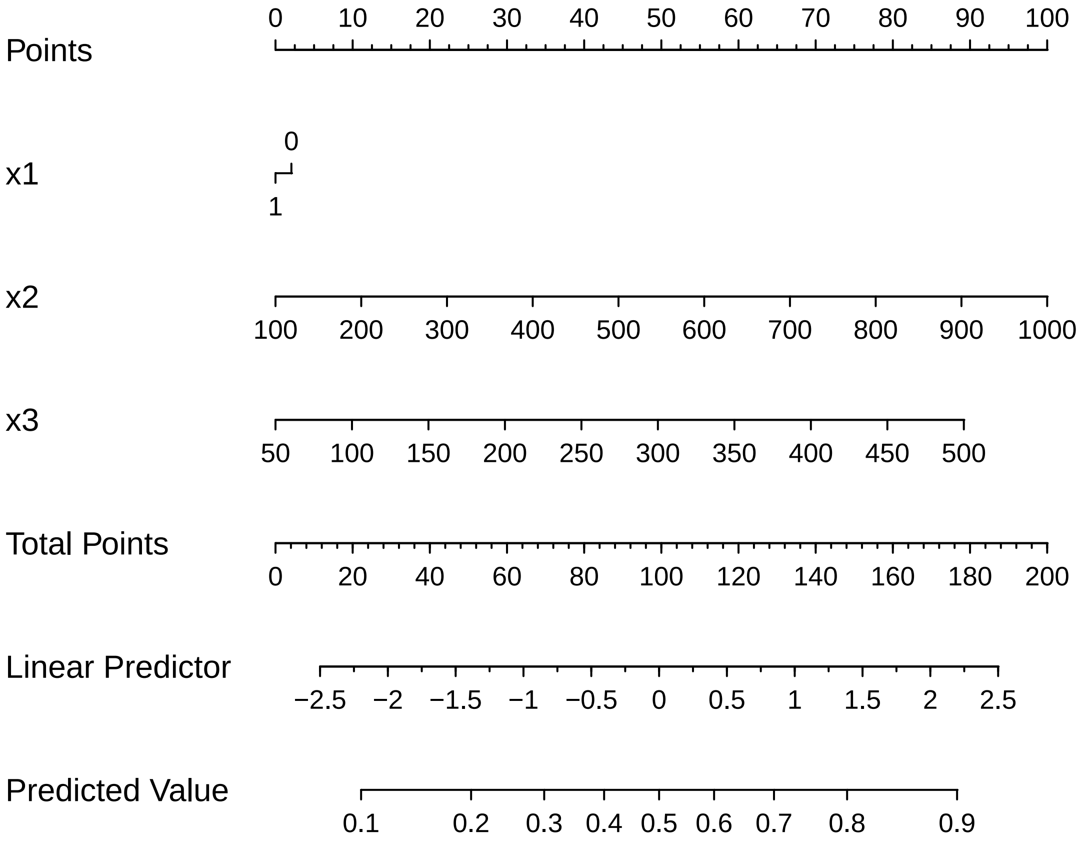
20.3 Plot B for Question 20



20.4 Plot C for Question 20



20.5 Plot D for Question 20



Display 1 for Question 20 describes the results of a logistic regression model fit. One of the four Plots for Question 20 describes that same model. Which one? (Hint: the nomograms in Plots B, C, and D all show the probability of the outcome being 1 as the "Predicted Value".)

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

21 Question 21

21.1 Display for Question 21

- Statement I. A main effects model fit with Poisson regression provides a statistically significantly worse fit (at the 95% confidence level) than a model fit with Negative Binomial regression.
- Statement II. The rootogram for the Poisson model indicates a substantially better fit than the rootogram for the Negative Binomial model.
- Statement III. The rootogram for the Poisson model indicates a substantially worse fit than the rootogram for the Negative Binomial model.

The `data21.csv` data set (which will be used in Questions 21-23) is available to you on the course web site. The outcome of interest in that data set, labeled `y`, is the number of standards (out of 6) met by subjects involved in an alcoholism treatment program. Subjects are released from the program when they meet all six standards. The data in `y` describe the number of standards met after one week of treatment for 200 recent subjects. Measures `x1`, `x2` and `x3` are predictors of `y`, whose main effects (only) are of interest to us. `x1` and `x3` are quantitative measures, and `x2` indicates whether or not the subject has completed a specific group of tasks. Fit a Poisson regression model to these data, and compare it to a negative binomial regression. Which of the statements listed in the Display for Question 21 are true?

```
a. I only.
b. II only.
c. III only.
d. I and II
e. I and III
f. II and III
g. All three statements.
h. None of these three statements.
```

22 Question 22

22.1 Display for Question 22

The three new subjects are Amy, Bart and Chris.

Name	`x1`	`x2`	`x3`
Amy	3	1	4
Bart	2	0	0
Chris	4	1	6

Use the Poisson regression model you fit in Question 21 to make a prediction for `y` for the three new subjects listed in the Display for Question 22. Rank the three new subjects in order of their predicted `y`, from highest (first) to lowest.

```
a. Amy has the highest predicted `y`, then Bart then Chris
b. Amy is highest, then Chris then Bart
c. Bart is highest, then Amy then Chris
d. Bart is highest, then Chris then Amy
e. Chris is highest, then Amy then Bart
f. Chris is highest, then Bart then Amy
```

23 Question 23

Now, instead of treating `y` in `data21` as a count variable, treat it as an ordinal category, and fit a new model that is appropriate for such an outcome using again the main effects of x1, x2 and x3 as predictors. Use that model to predict the actual category that our three new subjects (Amy, Bart and Chris) will fall into, and compare that to the results you found in Question 22. How many of the three new subjects get a different predicted count with this ordinal categorical regression model, than they do when you round the predicted count made with the Poisson model to an integer?

```
a. None of the three subjects.
b. One subject, specifically Amy.
c. One subject, specifically Bart.
d. One subject, specifically Chris.
e. Exactly two of the three subjects.
f. All three subjects.
```

24 Question 24

24.1 Display for Question 24

```
res24_BIC <- bestglm(Xy = data.frame(data24), family = binomial,
                  IC = "BIC", method = "exhaustive", TopModels = 3)
```

```
Morgan-Tatar search since family is non-gaussian.
```

```
res24_BIC$BestModels
```

```
   cov1  cov2  cov3  cov4  cov5  cov6 Criterion
1 FALSE  TRUE FALSE  TRUE FALSE FALSE  106.5410
2 FALSE FALSE FALSE  TRUE FALSE FALSE  106.7934
3 FALSE  TRUE FALSE FALSE FALSE FALSE  110.0057
```

```
res24_CV <- bestglm(Xy = data.frame(data24), family = binomial,
                 IC = "CV")
```

```
Morgan-Tatar search since family is non-gaussian.
```

```
res24_CV
```

```
CVd(d = 82, REP = 1000)
BICq equivalent for q in (0.133590203067706, 0.468490321674287)
Best Model:
               Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -4.62670316 1.26666101 -3.652677 0.0002595209
cov4         0.03065343 0.01131937  2.708051 0.0067679590
```

An “all subsets” approach as implemented in the `bestglm` package was used to fit a logistic regression model to describe the relationship between a binary outcome, y, and six predictors labeled x1 through x6, using two different approaches to variable selection, as shown in the Display for Question 24. The first approach shown in the Display was an exhaustive search for the best possible BIC result. The second approach shown in the Display involved cross-validation. Which of the following statements is true?

```
a. The model selected by the cross-validation procedure has the best BIC
available in a subset of these predictors.
b. The model selected by the cross-validation procedure has the second
best BIC available in a subset of these predictors.
c. The model selected by the cross-validation procedure has the third
best BIC available in a subset of these predictors.
d. None of these statements are true.
e. It is impossible to identify the true statement from the information
provided.
```

Setup for Questions 25-33

Questions 25-33 on your exam relate to data which describe the mass (our outcome of interest) and six additional physical measurements of 24 randomly chosen male subjects of ages 16-30 in good health. The outcome, `mass`, is in kilograms. All other measurements are in centimeters. Subjects slightly tensed each muscle being measured, and each measure was taken in a standard way, in an effort to ensure measurement consistency.

You have been provided, in a separate HTML file (entitled quiz02_output_for_students.html) with 30 different pieces of R output that may be useful in responding to Questions 25-33. Please consult that material carefully in answering these questions.

25 Question 25

Which of the following predictors has the weakest correlation with the outcome variable, mass?

a. bicep
b. chest
c. forearm
d. height
e. neck
f. waist

26 Question 26

26.1 Display for Question 26

- R. The model that uses all six predictors
- S. The model that uses four predictors, leaving out bicep and neck.
- T. The model that uses three predictors, specifically forearm, height and waist.

Several models are studied in this output, including the three listed in the Display for Question 26. In which of those three regression models do we see a substantial problem with collinearity?

```
a. Model R, only
b. Model S, only
c. Model T, only
d. Exactly two of Models R, S and T
e. Models R, S and T
f. None of the above.
```

27 Question 27

How many predictors are included in the most attractive model based on the bias-corrected Akaike Information Criterion, according to the best subsets output? Please count the intercept as a predictor here.

```
a. 2
b. 3
c. 4
d. 5
e. 6
f. 7
```

28 Question 28

Which predictors are contained in the model identified as having the maximum adjusted R-squared value (0.921) by the best subsets procedure?

```
a. `forearm` only
b. `forearm` and `waist`
c. `forearm`, `waist`, and `height`
d. `forearm`, `waist`, `height`, and `chest`
e. the five predictors other than `bicep`
f. all six predictors
```

29 Question 29

Consider the 95% confidence interval estimate for each of the predictors below after all of the other listed predictors has been accounted for? How many of these six predictors will have confidence intervals including zero?

```
a. 1
b. 2
c. 3
d. 4
e. 5
f. None of them.
g. All of them.
```

30 Question 30

Which of these predictors are identified as important on the basis of a backwards elimination procedure starting with the full model and using AIC to determine steps?

```
a. `forearm` only
b. `forearm` and `waist`
c. `forearm`, `waist`, and `height`
d. `forearm`, `waist`, `height`, and `chest`
e. the five predictors other than `bicep`
f. all six predictors
```

31 Question 31

According to the output provided regarding the Cp statistic, which of the following models is worthy of further consideration?

```
a. The simple regression model on the predictor most highly correlated
    with mass.
b. The model that uses all of the predictors except height.
c. The model that uses three predictors, specifically forearm, height
    and waist.
d. The model that uses two predictors, specifically forearm and waist.
e. None of these.
```

32 Question 32

Of the predictors `bicep`, `chest` and `waist`, how many add statistically significant (at the 10% level) predictive value to a model which already accounts for forearm size?

```
a. 0
b. 1
c. 2
d. 3
```

33 Question 33

Using the model suggested by the adjusted R-squared plot, what is the effect on mass of moving from the 25th percentile to the 75th percentile of forearm measurement, while holding all other predictors constant?

```
a. Mass increases by fewer than 6 kilograms.
b. Mass increases by 6 or more kilograms.
c. Mass decreases by fewer than 6 kilograms.
d. Mass decreases by 6 or more kilograms.
```

Setup for Questions 34-36

The `data34.csv` file available on the course web site contains information on 1000 animal subjects who took part in an observational study. You will use this data set for Questions 34-36. The data includes information on:

- `alive` = the subject's vital status at the end of the study (`alive` = 1 if alive at the end of the study, 0 otherwise)
- `treated` = 1 if the subject received the treatment of interest and 0 if the subject received usual care
- `age`, in years, at the start of the study
- `female` = 1 if the subject is female, biologically
- `comor` = a count of comorbid conditions (maximum = 7)

34 Question 34

How many rows in the `data34.csv` contain at least one missing value?

Questions 35 and 36 are BONUS questions

Questions 1-34 are required, and are worth 3 points each. Despite this, we treat the maximum score on the Quiz as 100, rather than 102. Questions 35 and 36 are BONUS questions, each worth 5 points for a correct response, and with no partial credit awarded. So if you do questions 35 and 36 correctly, your total score on the Quiz could be as high as 112, but, again, we will treat your score as if it were out of 100 points. You are welcome to skip Questions 35 and 36 if you like. You must answer Question 35 correctly in order for us to grade your Question 36. Questions 35 and 36 use the data setup for Question 34, which we repeat below.

Setup for Questions 34-36

The `data34.csv` file available on the course web site contains information on 1000 animal subjects who took part in an observational study. You will use this data set for Questions 34-36. The data includes information on:

- `alive` = the subject's vital status at the end of the study (`alive` = 1 if alive at the end of the study, 0 otherwise)
- `treated` = 1 if the subject received the treatment of interest and 0 if the subject received usual care
- `age`, in years, at the start of the study
- `female` = 1 if the subject is female, biologically
- `comor` = a count of comorbid conditions (maximum = 7)

35 OPTIONAL BONUS Question 35

Specify the R code you would use to fit a logistic regression model to predict `alive` on the basis of main effects of `treated`, `age`, `female` and `comor`, using multiple imputation to deal with missing values, and setting a seed of `43237` for the imputation work. In your imputation process, you should include all variables in the `data37` data, run 20 imputations, and use `nk = c(0, 3), tlinear = TRUE, B = 10` and `pr= FALSE`.

36 OPTIONAL BONUS Question 36

Using your model specified in Question 35, estimate the effect of treatment (vs. control) on the odds of being alive at the end of the study. Your odds ratio estimate should compare `treated` to `control`, while adjusting for the effects of `age`, `female` and `comor`. Provide both a point estimate and a 95% confidence interval. Interpret your result in a single sentence.