432 Quiz 2 for Spring 2019

Thomas E. Love

Due 2019-05-02 at 2 PM. Version: 2019-04-24 11:14:08

Instructions

There are 36 questions (not including the selection of your name, and the final affirmation at the end) on the Quiz, divided into several sections. Please select or type in your best response for each question. You should try to answer all 36 questions, as there is no difference in score between an incorrect answer and a blank answer.

- The more time-consuming questions are gathered toward the front of the Quiz.
- Questions 1-11 involve work with four data sets I have provided to you on the course web site.
- Questions 25-33 make use of the output file I have provided to you on the course web site.
- Most questions are worth 3 points each. The exceptions are Questions 1, 2, 7 and 8, which are each worth 6 points. So the total possible score is 120 points.
- I expect to award some partial credit on the 6-point questions, but do not anticipate awarding meaningful partial credit on most of the 3-point questions, as they are (almost exclusively) multiple-choice items.
- You must complete this Quiz by 2 PM on Tuesday May 2. You will have the opportunity to edit your responses after completing the Quiz, but this must be completed by the deadline. Be sure to save your work. To save your work, you will have to select your name below, and also answer the final question - the attestation that you did your work alone, and then submit it. The software will then provide you (and automatically email you) with a link to return to the Quiz while retaining your answers so far.
- You are welcome to consult the materials provided on the course website, but you are NOT allowed to discuss the questions on this Quiz with anyone other than the teaching assistants, and Professor Love. As a reminder, you can contact us at 431-help at case dot edu. We will try to answer all questions in a timely way, but you should not leave things to the last moment.
- All Quiz 2 Materials are linked at https://github.com/THOMASELOVE/2019-432/tree/master/quizzes/quiz2

Setup for Questions 1 and 2

Questions 1 and 2 involve the following scenario, as well as the `quiz2A.csv` data set. Suppose you wish to model the relationship between a measure of anxiety, on a scale ranging from 0 (very low anxiety) to 100 (very high anxiety) and three predictors. The primary predictor of interest is a measure of childhood trauma, which is available in three categories (describing low, medium and high amounts of trauma as a child.) Also planned for inclusion in the model are each subject's age (which is centered in the models shown below), and their sex.

The `quiz2A.csv` data available on our web site contains a sample of 350 adults ages 35-64. The data include:

- a subject id number (`subject`)
- the subject's age, after centering (`age_c`)
- the subject's sex (`sex` = female or male)
- the subject's trauma category (`trauma` = low, medium or high)
- the subject's measured `anxiety` (on a scale from 0 - 100, where higher scores indicate higher levels of anxiety)

Before you can answer Questions 1 and 2, you will need to:

a. Import the `quiz2A` data into R.
b. Manage the data so that `sex` will be treated as a factor, with female as the baseline category.
c. Manage the data so that `trauma` = low will be the baseline category for that factor.
d. Create Model `m1` for `anxiety`, which includes the main effects of `age_c`, `sex` and `trauma`.
e. Create Model `m2` for `anxiety`, which adds the interaction of `sex` and `trauma` to Model 1.

As a little hint, I have provided some of the output for Model `m1`, below. Yours should look like this.

```

Call:
lm(formula = anxiety ~ trauma + sex + age_c, data = quiz2A)

Coefficients:
 (Intercept)  traumamedium    traumahigh       sexmale         age_c  
     53.6463        9.7117       17.9490       -9.3402        0.1436  
```

1 Question 1 (6 points)

Describe the predicted effect (in terms of both point and interval estimates) of the various levels of trauma on anxiety in Model `m1`, concisely and accurately, without using the term "statistically significant", and using complete English sentences only. You should specify your estimates, but include no other R code or output. This is an observational study, so describe your conclusions appropriately. I expect you will complete this task in fewer than 500 characters (this paragraph uses 499.)

2 Question 2 (6 points)

Describe the impact on the predicted effects (in this question, please discuss only the point estimates) on anxiety that you see in the new Model `m2` incorporating the interaction term. Your goal here is to interpret the effect of the interaction in context, both concisely and accurately, without using the term "statistically significant", and using complete English sentences. Include no R code or output. Remember: this is an observational study. This task also requires about 500 characters.

Setup for Questions 3-5

The `quiz2B.csv` data set (which will be used in Questions 3-5) 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`, as displayed below.

```
quiz2B
```

```
# A tibble: 150 x 6
      x1    x2    x3    x4    x5     y
   <dbl> <int> <dbl> <dbl> <dbl> <dbl>
 1    51     8     1     1  54    -4.2
 2    79    10     1     2  84.4 -20.2
 3    79     8     0     6   2.9 -12.2
 4    73     9     1    11  69.6  19.8
 5    77     9     0     3   1.4  -8.8
 6    69    17     1     2  70.3   4.2
 7    63    14     1     3  62.9 -18.3
 8    74     8     1     1  73.8  -6.2
 9    97    11     1     8 103.   13.7
10    80    12     1    15  80.2  18.5
# ... with 140 more rows
```

3 Question 3

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, x3 and x5
c. x1, x2 and x5
d. x2 and x5
e. x5 alone
f. None of these
```

4 Question 4

Following on from Question 3, fit the model suggested by the stepwise regression (that you identified in Question 3) to the full data set of 150 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
```

5 Question 5

Display for Question 5

```r
set.seed(432)
q5_models <- quiz2B %>%
    modelr::crossv_kfold(k = 10) %>%
```

Use 10-fold cross-validation to evaluate the model you fit in Question 4. Note that the Display for Question 5 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 5 but no larger than 7.
b. Above 7 but no larger than 9.
c. Above 9 but no larger than 11.
d. Above 11 but no larger than 13.
e. None of the above.
```

Setup for Questions 6-8

The quiz2C.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 6-8. 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)

6 Question 6

How many rows in the quiz2C.csv data contain at least one missing value?

7 Question 7 (6 points)

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 quiz2C data other than the subject identifying code, run 20 imputations, and use nk = c(0, 3), tlinear = TRUE, B = 10 and pr= FALSE. Do not show the results here, just the code. Assume all necessary packages have been pre-loaded using the library() function, and that the quiz2C data have been successfully imported into R already.

8 Question 8 (6 points)

Using the model you specified in Question 7, 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 concisely and correctly in complete English sentences only. Do not include any R code or output here.

Setup for Questions 9-11

The `quiz2D.csv` data set (which will be used in Questions 9-11) 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.

9 Question 9

Fit a Poisson regression model to the data in the `quiz2D.csv` file, and compare your result to what you obtain using a negative binomial regression. Treat variable `x2` as a number (1/0) rather than converting it to a factor.

Display for Question 9

- 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.

Which of the statements listed in the Display for Question 9 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.
```

10 Question 10

Display for Question 10

Consider three new subjects, who are named Abigail, Brad and Chen. Here are their values of the predictors.

Name	x1	x2	x3
Abigail	3	0	4
Brad	4	1	0
Chen	2	1	6

Use the Poisson regression model you fit in Question 9 to make a prediction for y for the three new subjects listed in the Display for Question 10. Rank the three new subjects in order of their predicted y, from highest (first) to lowest.

a. Abigail has the highest predicted `y`, then Brad then Chen
b. Abigail is highest, then Chen then Brad
c. Brad is highest, then Abigail then Chen
d. Brad is highest, then Chen then Abigail
e. Chen is highest, then Abigail then Brad
f. Chen is highest, then Brad then Abigail

11 Question 11

Now, instead of treating y in quiz2D 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 (Abigail, Brad and Chen) will fall into, and compare that to the results you found in Question 10.

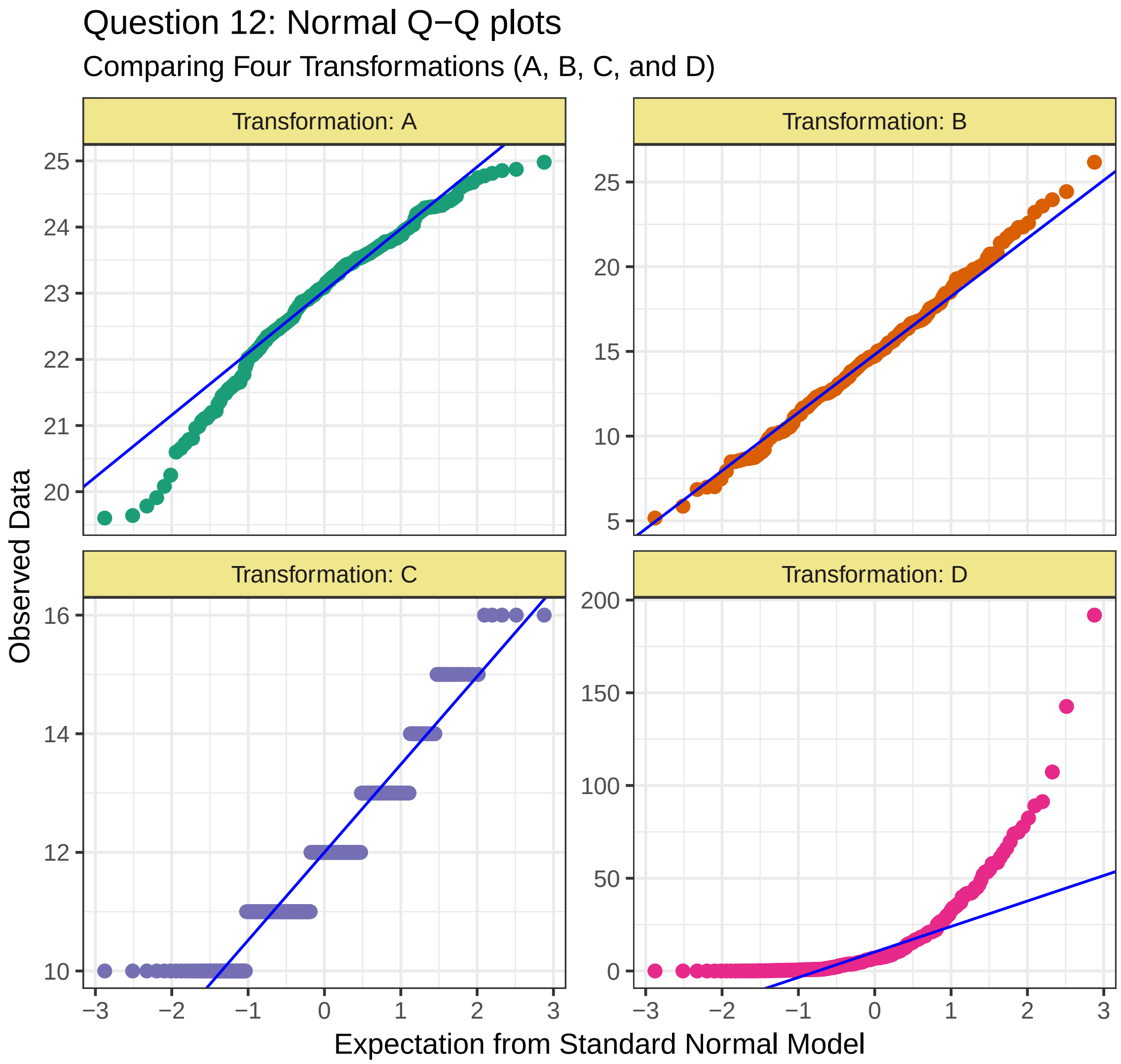
How many of the three new subjects get a different predicted count with this ordinal categorical regression model, than they do when you round the predicted count made with the Poisson model to an integer?

a. None of the three subjects.
b. One subject, specifically Abigail
c. One subject, specifically Brad
d. One subject, specifically Chen
e. Exactly two of the three subjects
f. All three subjects.

12 Question 12

The Display for Question 12 shows normal Q-Q plots of four potential transformations under consideration for modeling a quantitative outcome, based on a sample of 250 observations.

Display for Question 12



Which of the plots in the Display best supports the use of a Normal model for the data?

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

13 Question 13

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

14 Question 14

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:

- `compliance` = whether or not the surgical team complied with all guidelines used to formulate the checklist,
- `intervention` = half of the procedures used the checklist and half did not, and
- 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.

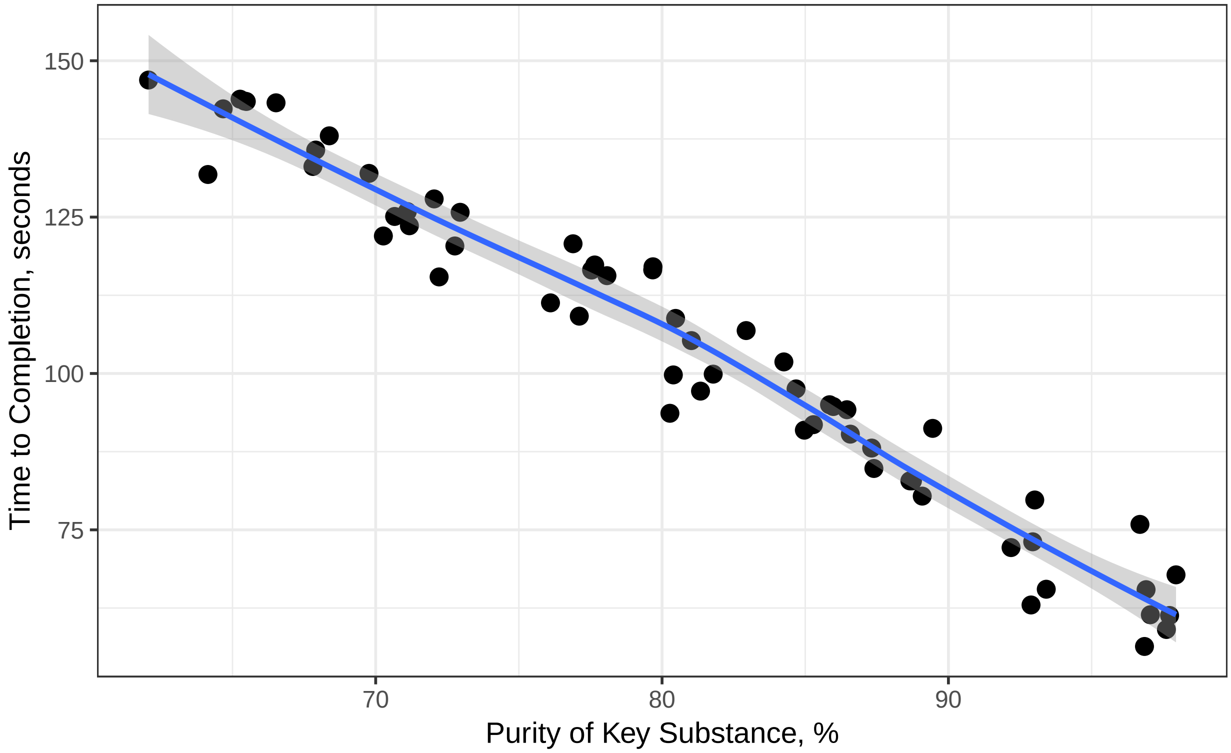
Suppose we want to build a point and interval estimate for how "the odds of successful compliance comparing surgeries using the intervention to surgeries not using the intervention" were different for surgeries depending on whether the urgency level was 40 as opposed to 20.

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.
```

15 Question 15

Display for Question 15



Question 15: Scatterplot of Time vs. Purity

n = 60 experiments, Plotted fit is a loess smooth

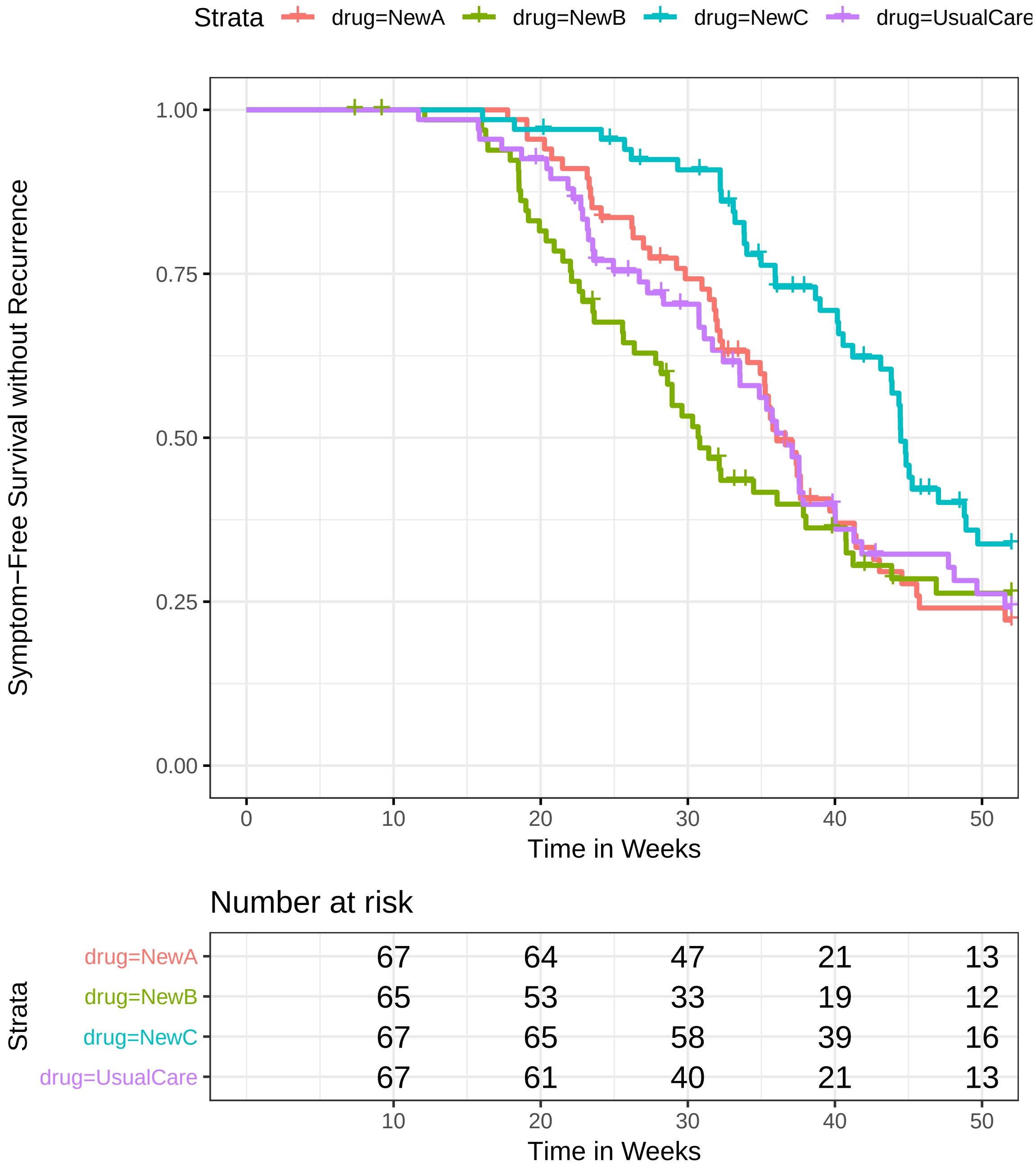
You are fitting a model to describe the time it takes for a chemical reagent to complete a reaction in an experimental setting. You have conducted 60 such experiments, varying the purity level of a key substance. There is variation in the time required, which is associated with the purity, which is measured on a 60-100 scale, since if the substance is not at least 60% pure, the reaction will not happen. The Display for Question 15 shows the scatterplot of time and purity for your 60 experimental runs.

Which of the following statements is most true about an simple linear regression model (call it Model 15) fit to represent these data?

a. Model 15 is not helpful, since we should be fitting a Cox model instead.
b. Model 15 fits the data much less well than a model which adds a five-knot restricted cubic spline in purity.
c. Model 15 explains more than 50% of the variation in completion time.
d. Model 15 explains between 25% and 50% of the variation in completion time.
e. Model 15 will have an R-squared value of about 0.10

16 Question 16

Display 1 for Question 16



Strata	10	20	30	40	50
drug=NewA	67	64	47	21	13
drug=NewB	65	53	33	19	12
drug=NewC	67	65	58	39	16
drug=UsualCare	67	61	40	21	13

Display 2 for Question 16

```
print(fit16, print.rmean = TRUE)
```

```
Call: survfit(formula = data16$S ~ data16$drug)

                       n events *rmean *se(rmean) median 0.95LCL 0.95UCL
data16$drug=NewA      67     47   37.3       1.33   36.0    34.9    41.3
data16$drug=NewB      67     45   34.0       1.65   30.8    28.2    40.7
data16$drug=NewC      67     38   43.0       1.16   44.5    43.1    49.7
data16$drug=UsualCare 67     44   36.7       1.54   36.6    33.5    41.3
    * restricted mean with upper limit =  52
```

```
survdiff(data16$S ~ data16$drug)
```

```
Call:
survdiff(formula = data16$S ~ data16$drug)

                       N Observed Expected (O-E)^2/E (O-E)^2/V
data16$drug=NewA      67       47     42.9     0.393     0.523
data16$drug=NewB      67       45     34.2     3.382     4.222
data16$drug=NewC      67       38     57.1     6.400     9.618
data16$drug=UsualCare 67       44     39.7     0.455     0.590

 Chisq= 10.7  on 3 degrees of freedom, p= 0.01
```

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, new drug C, or the usual medication.

The Kaplan-Meier curve comparing the drugs is shown in Display 1 for Question 16, and some additional information about the Kaplan-Meier fit is shown in Display 2 for Question 16.

Which of the drugs has the most promising survival curve (longest time to recurrence of symptoms) in these data?

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

17 Question 17

Display for Question 17

```
set.seed(9432171); validate(m17)
```

```
          index.orig training    test optimism index.corrected  n
Dxy           0.6260   0.6570  0.6092   0.0478          0.5782 40
R2            0.3884   0.4213  0.3584   0.0629          0.3255 40
Intercept     0.0000   0.0000 -0.0438   0.0438         -0.0438 40
Slope         1.0000   1.0000  0.8973   0.1027          0.8973 40
Emax          0.0000   0.0000  0.0315   0.0315          0.0315 40
D             0.3343   0.3745  0.3033   0.0712          0.2631 40
U            -0.0200  -0.0200  0.0068  -0.0268          0.0068 40
Q             0.3543   0.3945  0.2965   0.0980          0.2564 40
B             0.1745   0.1660  0.1841  -0.0181          0.1926 40
g             1.6954   1.8766  1.5915   0.2851          1.4103 40
gp            0.3201   0.3289  0.3065   0.0224          0.2977 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 0.83, Nagelkerke R-square about 0.42
b. Area under the ROC curve will be about 0.81, Nagelkerke R-square about 0.39
c. Area under the ROC curve will be about 0.80, Nagelkerke R-square about 0.36
d. Area under the ROC curve will be about 0.79, Nagelkerke R-square about 0.33
e. Area under the ROC curve will be about 0.66, Nagelkerke R-square about 0.42
f. Area under the ROC curve will be about 0.63, Nagelkerke R-square about 0.39
g. Area under the ROC curve will be about 0.61, Nagelkerke R-square about 0.36
h. Area under the ROC curve will be about 0.58, Nagelkerke R-square about 0.33
i. None of the above.
```

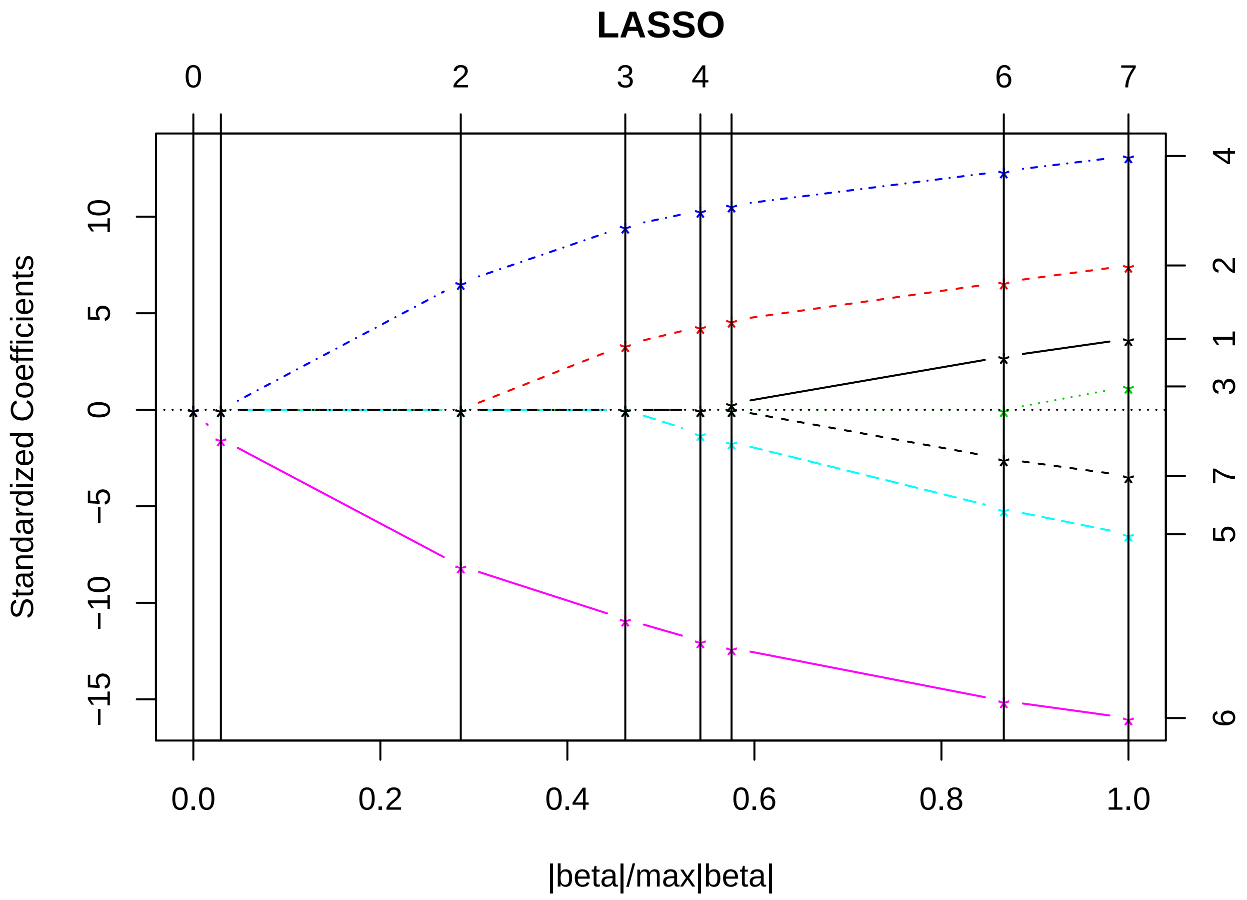
18 Question 18

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.
```

19 Question 19

Display for Question 19



The Display for Question 19 shows the result of applying the lasso to a data set containing seven predictors, labeled 1-7 in the plot. If the value of the key fraction to minimize cross-validated mean squared prediction error is 0.38, then how many of the 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. It is impossible to tell.
```

Setup for Question 20

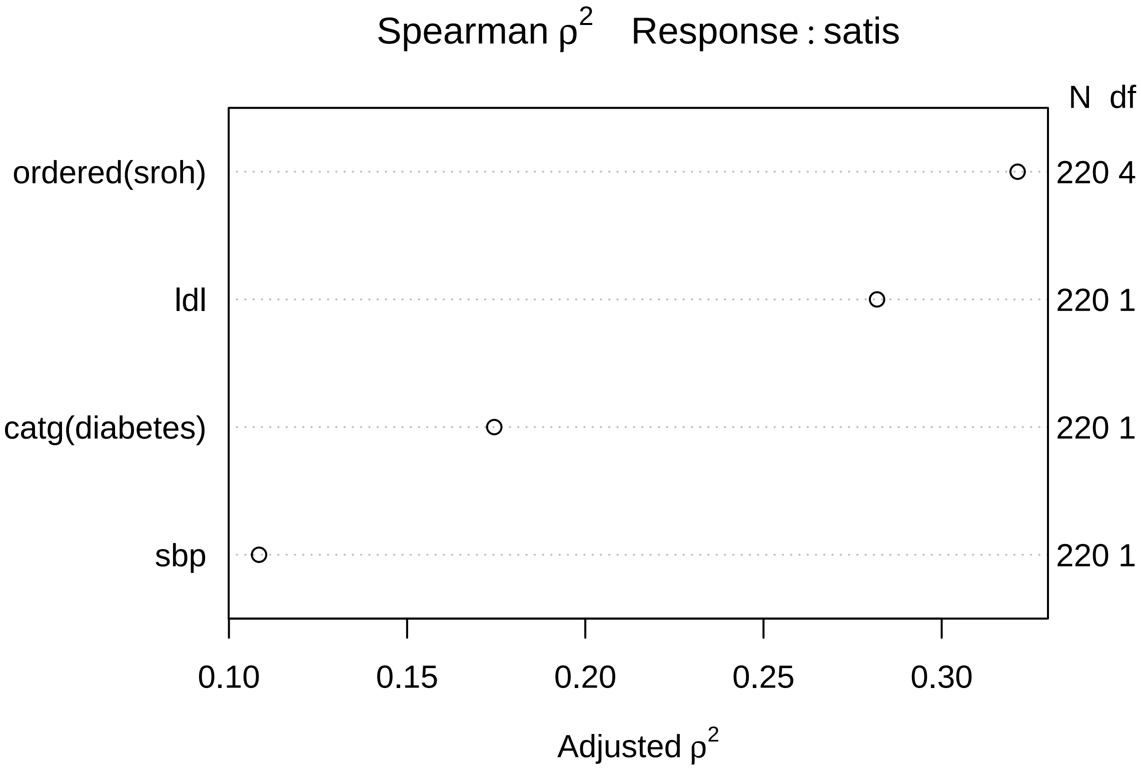
Suppose you plan to fit a model to predict the level of a patient's satisfaction (`satis`, measured on a 0-100 scale, where `satis` = 100 indicates that a patient is extremely satisfied) with their health care, using a sample of 220 subjects. For each subject, you have information on:

- 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 20 shows a Spearman rho-squared plot for these subjects.

- Note that the use of the `catg` function tells the Spearman plot to consider the `diabetes` information as an unordered factor.
- The use of the `ordered` function tells the plot to consider the `sroh` information as an ordered factor.

Display for Question 20



20 Question 20

Consider the output and details provided above. Assuming you wish to include all of the main effects for the specified four 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. A `lrm` model including the interactions of `diabetes` with
    both `sbp`, and `ldl`.
b. An `ols` model including the interaction of `ldl` and `sroh`.
c. A `lrm` model including the interaction of `ldl` and `sroh`.
d. An `ols` model adding a restricted cubic spline in `ldl` with 5 knots.
e. A `lrm` model adding a restricted cubic spline in `ldl` with 5 knots.
f. An `ols` model including the interactions of `diabetes` with
    both `sbp`, and `ldl`.
g. It is impossible to tell which of these options is best.
```

21 Question 21

Suppose you have a data set 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_drop`
b. `fct_recode` and `fct_lump`
c. `fct_count` and `fct_relabel`
d. `fct_reorder`
e. `fct_collapse` and `fct_relevel`
```

22 Question 22

Display 1 for Question 22

```
data22
```

```
# A tibble: 100 x 6
      id    x1    x2    x3    x4     y
   <int> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1    96    99     0   100  23.2
 2     2   107    NA     0    NA  23.5
 3     3    97    84    NA   105  25.8
 4     4    97    96     1    NA  26  
 5     5    93    NA     0   108  15.6
 6     6   106    98    NA   110  23.1
 7     7    NA   112     0   109  11.1
 8     8   109    NA     1    NA  26.1
 9     9   109   110     1   103  26.8
10    10   109   119     1   107  29.7
# ... with 90 more rows
```

Given the data set data22 shown in Display 1 for Question 22, suppose you want to remove all rows containing missing values, then create a training sample containing 70% of the rows without missing data, and a test sample containing the other 30% of the values after missingness is removed.

Which of the chunks of R commands shown (on the next page) in Display 2 for Question 22 will accomplish the desired result?

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.

Display 2 for Question 22

Chunk I

```
set.seed(432)
data22_noNA <- data22 %>%
    filter(complete.cases(.))

data22_train2 <- data22_noNA %>%
    sample_frac(size = 0.70, replace = FALSE)

data22_test2 <-
    dplyr::anti_join(data22_noNA, data22_train2, by = "id")
```

Chunk II

```
set.seed(432)
data22_train1 <- data22 %>%
    sample_frac(size = 0.70, replace = FALSE) %>%
    drop_na

data22_test1 <- data22 %>%
    sample_frac(size = 0.30, replace = TRUE) %>%
    drop_na
```

Chunk III

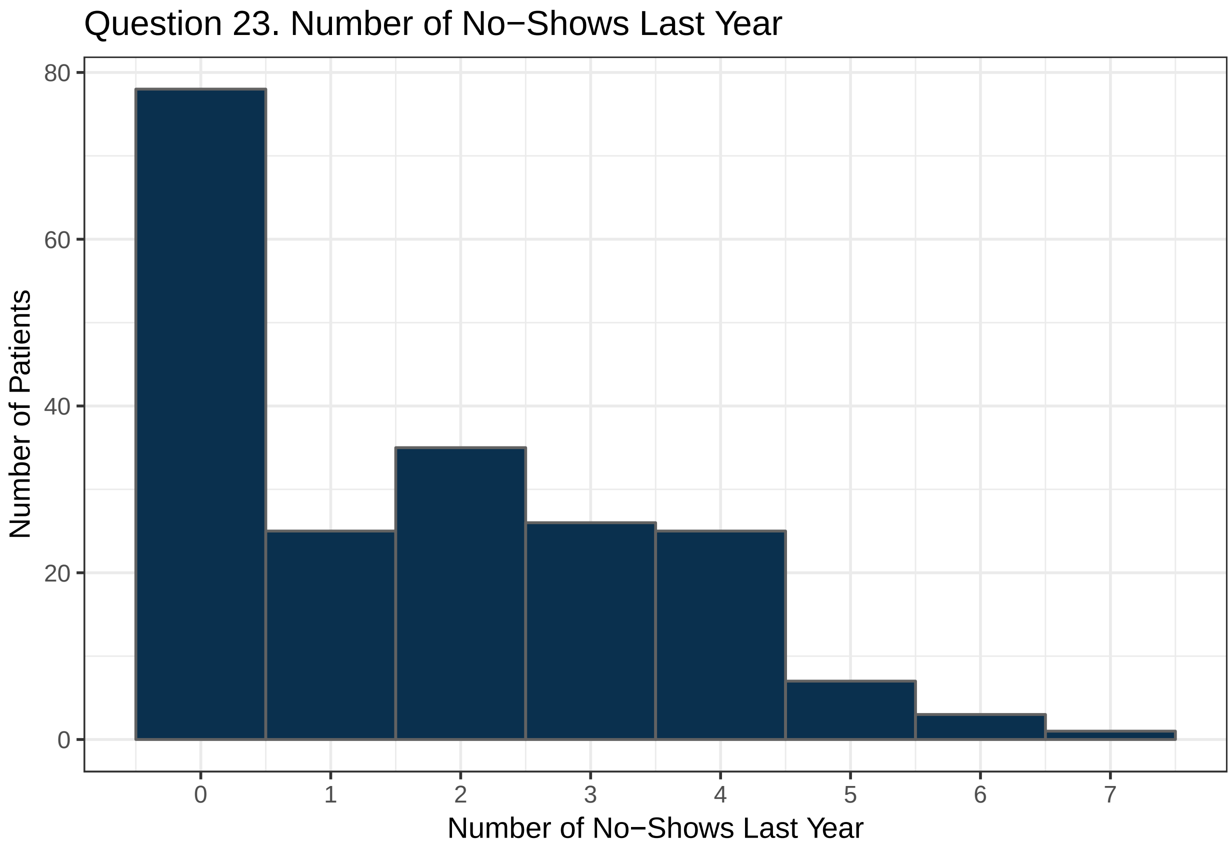
```
data22_noNA3 <- data22 %>%
    drop_na %>%
    mutate(rand = runif(n(), min = 0, max = 1))

data22_train3 <- data22_noNA3 %>%
    slice(which(rand < quantile(rand, 0.7)))

data22_test3 <- data22_noNA3 %>%
    slice(which(rand >= quantile(rand, 0.7)))
```

23 Question 23

Display for Question 23

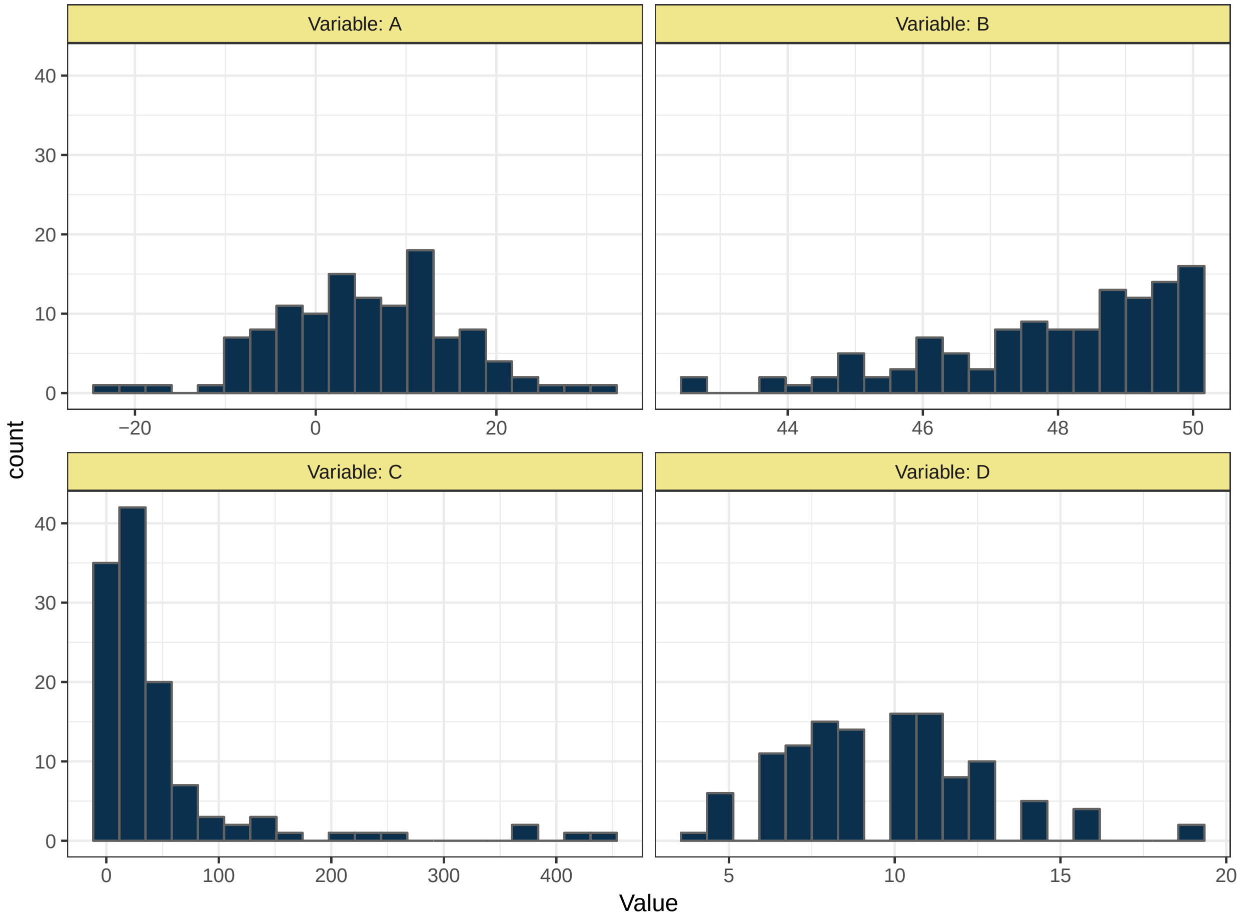


Suppose you are trying to build a regression model to predict `noshow`, the number of times a patient will "no show" an appointment for medical care in the next 12 months, on the basis of several characteristics related to their health, demographics, and satisfaction levels with prior visits. The `noshow` data on 200 patients from last year are visualized in the Display for Question 23. Which of the following models is most likely to be appropriate?

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

24 Question 24

Display for Question 24



Which of the four variables plotted in the Display for Question 24 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.
```

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 28 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 file (entitled `quiz2_output_for_students`) 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 after all of the other predictors have been accounted for. How many of the six predictors 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
```

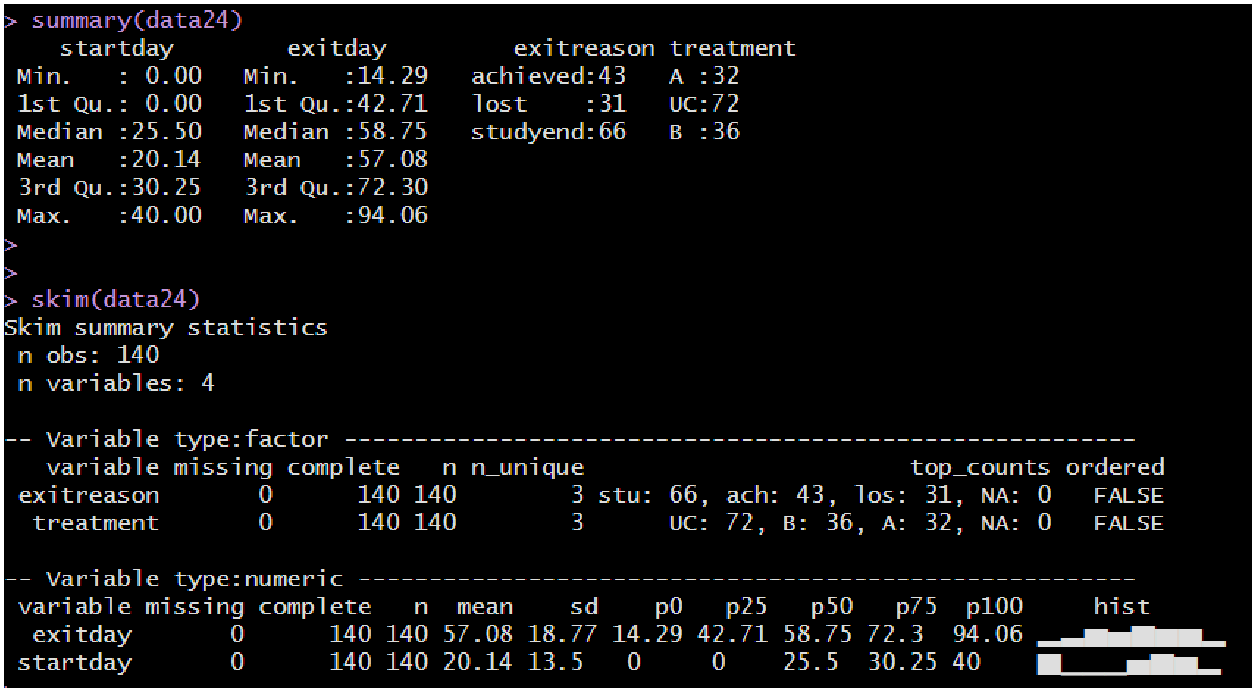
33 Question 33

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.
```

34 Question 34

Display 1 for Question 34



```
> summary(data24)
    startday        exitday           exitreason treatment
 Min.   : 0.00   Min.   :14.29   achieved:43   A :32
 1st Qu.: 0.00   1st Qu.:42.71   lost    :31   UC:72
 Median :25.50   Median :58.75   studyend:66   B :36
 Mean   :20.14   Mean   :57.08
 3rd Qu.:30.25   3rd Qu.:72.30
 Max.   :40.00   Max.   :94.06
>
>
> skim(data24)
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: 66, ach: 43, los: 31, NA: 0   FALSE
  treatment       0      140 140        3      UC: 72, B: 36, A: 32, NA: 0   FALSE

-- Variable type:numeric -------------------------------------------------------
 variable missing complete   n  mean    sd    p0   p25   p50   p75  p100     hist
  exitday       0      140 140 57.08 18.77 14.29 42.71 58.75 72.3   94.06
 startday       0      140 140 20.14 13.5   0     0     25.5  30.25 40
```

Display 1 for Question 34 shows a summary of the `data34` 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 `data24` 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 (on the next page) in Display 2 for Question 34 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.
```

34.1 Display 2 for Question 34

Chunk I

```
data34$S = Surv(time = data34$exitday - data34$startday,
                event = data34$exitreason %in% c("lost", "studyend"))
survdiff(S ~ treatment, data = data34)
```

Chunk II

```
survdiff(Surv(time = data34$exitday, event = data34$exitreason) ~ treatment)
```

Chunk III

```
data24$S = Surv(time = data34$exitday - data34$startday,
                event = data34$exitreason == "achieved")
survdiff(S ~ treatment, data = data34)
```

Setup for Question 35

Display 1 for Question 35

```
Logistic Regression Model

 lrm(formula = outcome ~ a + c + rcs(b, 3) + a %ia% b, data = data35, 
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.    
                          Ratio Test           Indexes           Indexes    
 Obs           190    LR chi2      71.85    R2       0.446    C       0.858    
  0             57    d.f.             5    g        2.740    Dxy     0.717    
  1            133    Pr(> chi2) <0.0001    gr      15.488    gamma   0.717    
 max |deriv| 0.002                          gp       0.298    tau-a   0.303    
                                            Brier    0.141                     

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -2.8107 1.2360 -2.27  0.0230  
 a         -1.9587 1.8961 -1.03  0.3016  
 c          0.0066 0.0022  2.95  0.0031  
 b          0.0013 0.0035  0.36  0.7196  
 b'         0.0155 0.0062  2.51  0.0120  
 a * b      0.0125 0.0075  1.66  0.0968  
```

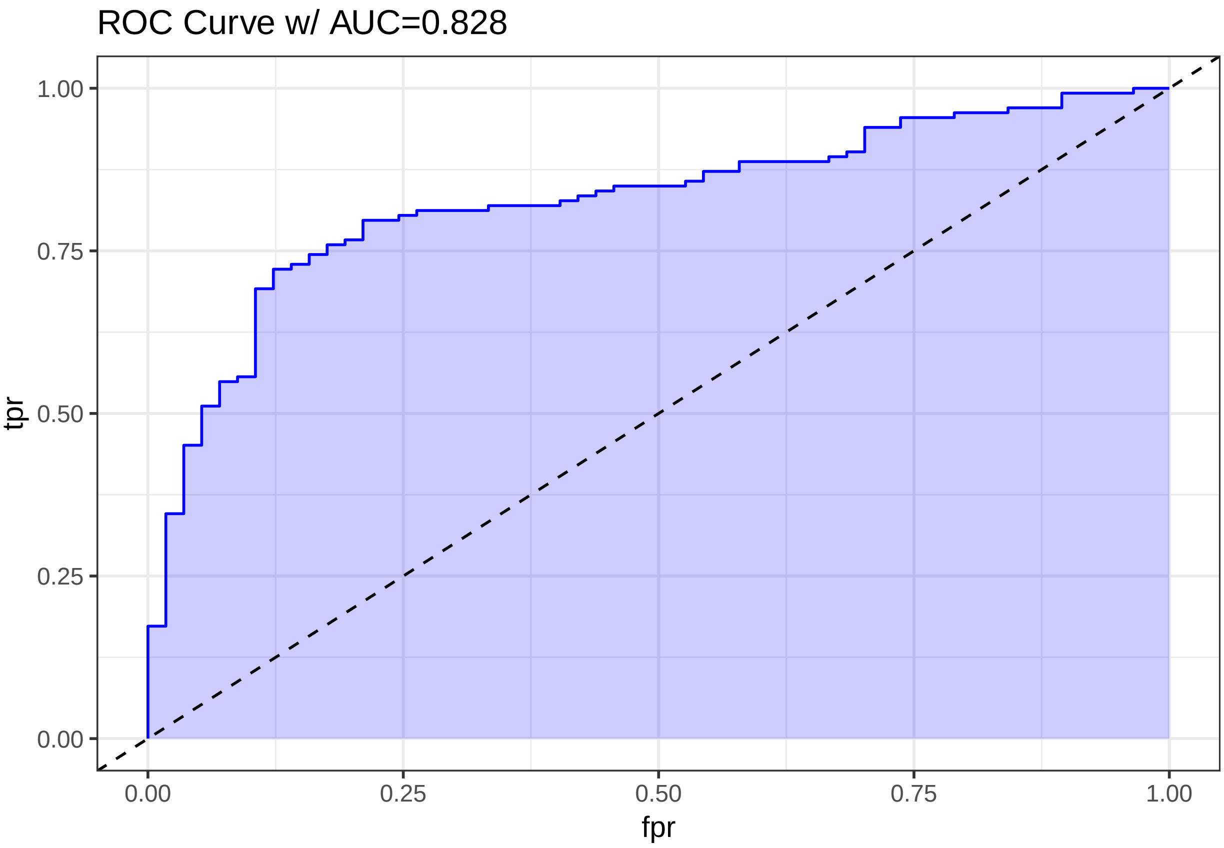
35 Question 35

Display 1 for Question 35 (shown on the previous page) describes the results of a logistic regression model fit. Exactly one of the four Plots for Question 35 (shown below and on the next few pages) describes that same model. Which one?

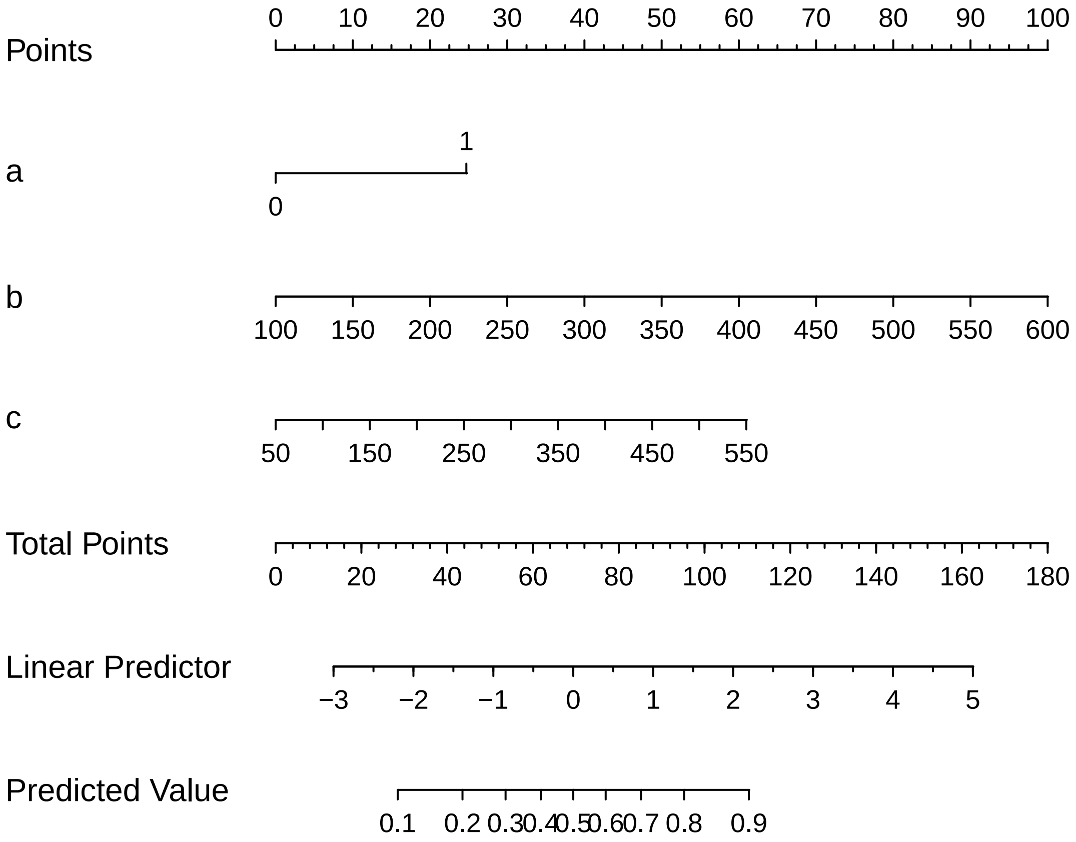
(Hint: the nomograms in Plots B, C, and D all show the estimated probability of the outcome being 1 as the "Predicted Value".)

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

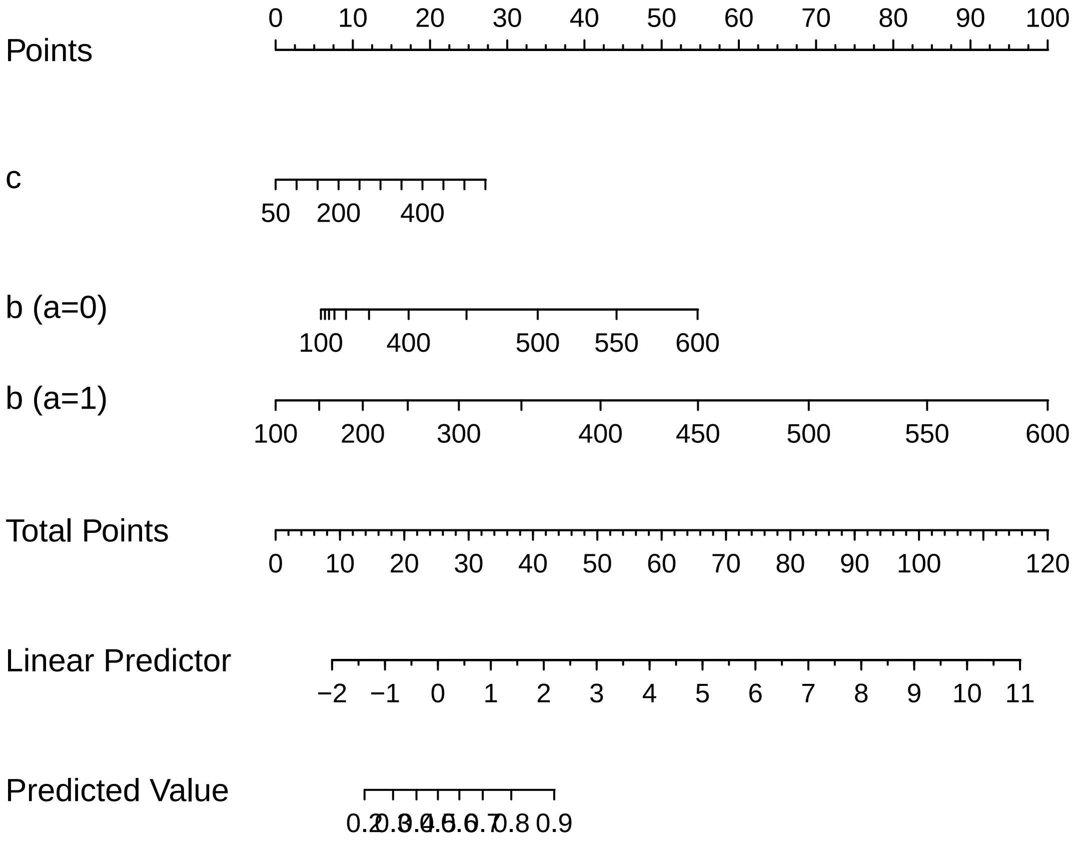
Plot A for Question 35



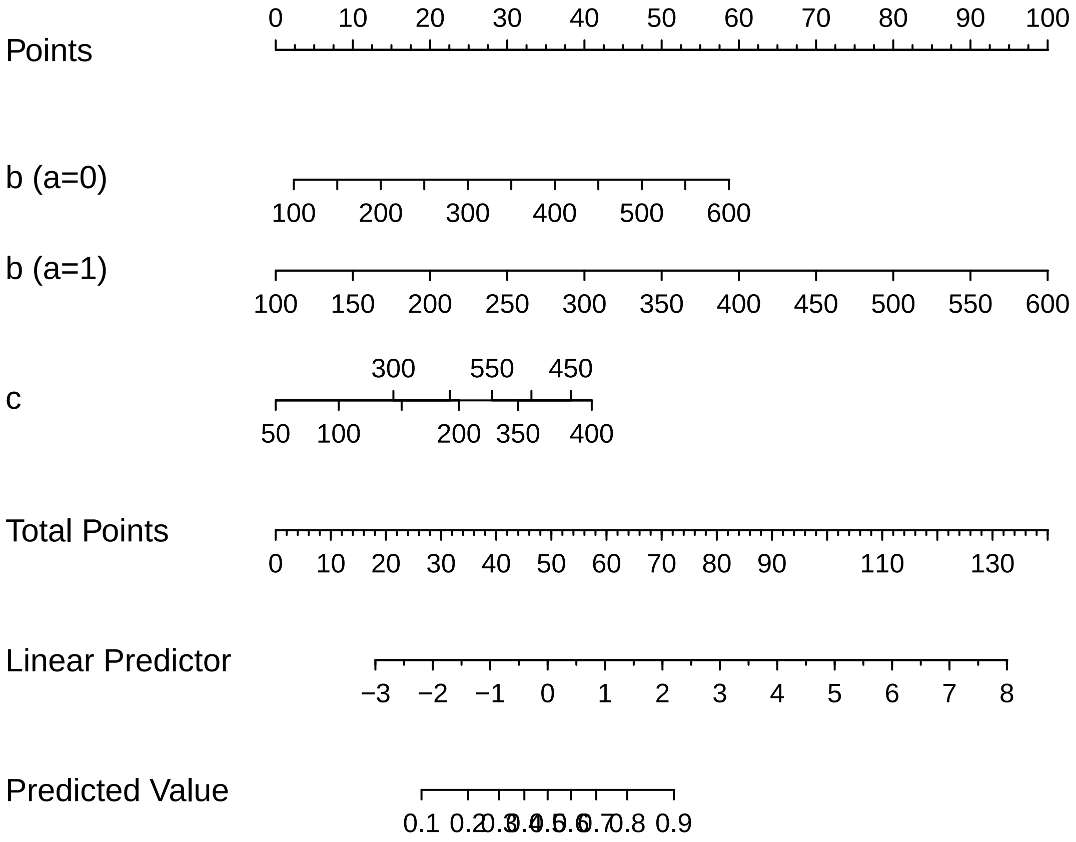
Plot B for Question 35



Plot C for Question 35

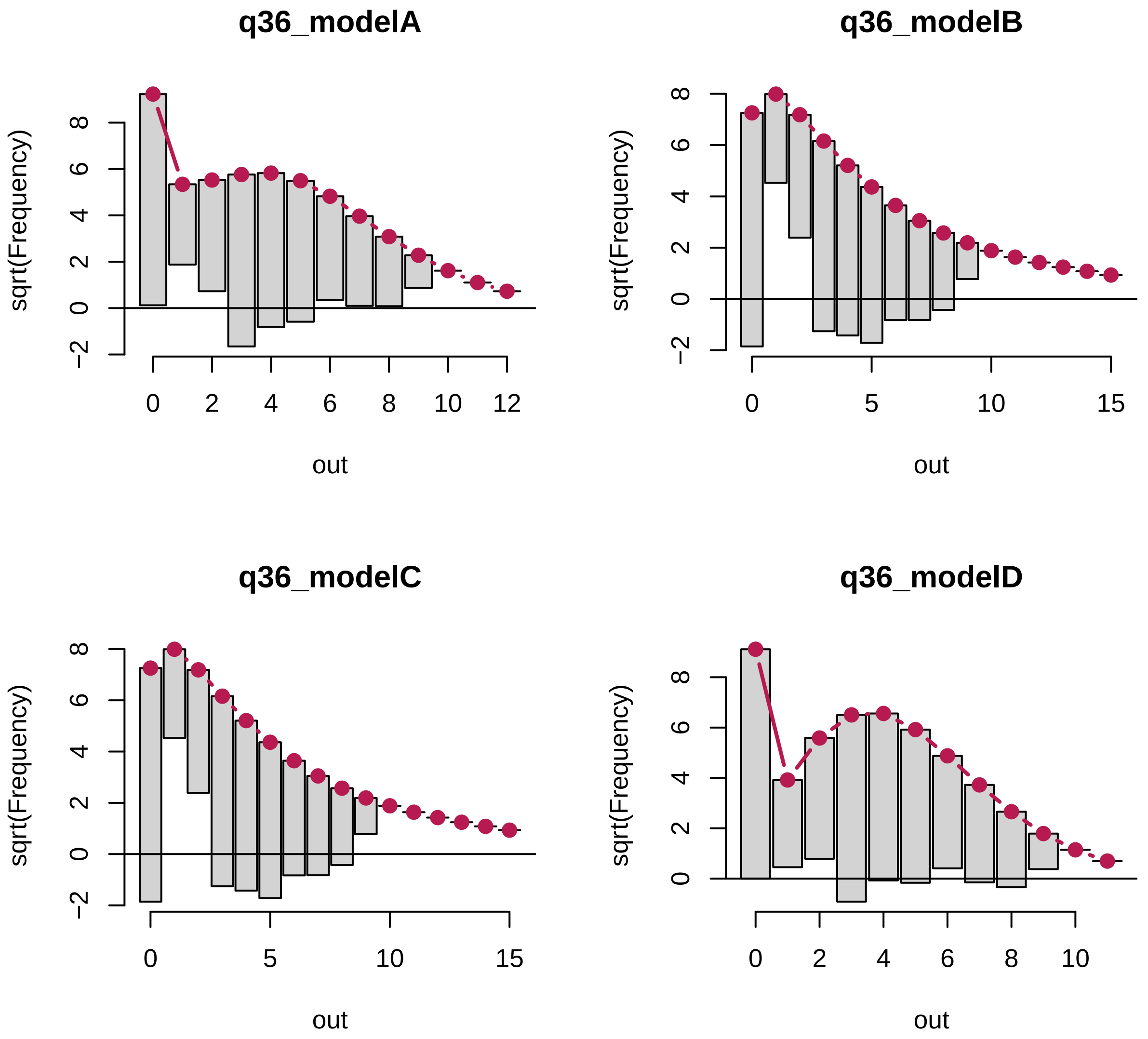


Plot D for Question 35



36 Question 36

36.1 Display for Question 36



The Display for Question 36 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. Model A
b. Model B
c. Model C
d. Model D
e. It is impossible to tell from the information provided.
```