Answer Sketch for 432 Quiz 2 - Spring 2019

Thomas E. Love

Due 2019-05-02 at 2 PM. Version: 2019-05-05

Instructions

There are 36 questions.

- 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.

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 my Model `m1`, below.

```
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.)

1.1 Answer 1

The model I would fit here is

```
# A tibble: 5 x 7
  term         estimate std.error statistic   p.value conf.low conf.high
  <chr>           <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)    53.6      1.49       36.1  3.24e-119  50.7       56.6
2 traumamedium    9.71     1.73        5.61 4.18e-  8   6.31      13.1
3 traumahigh     17.9      1.90        9.46 4.96e- 19  14.2       21.7
4 sexmale        -9.34     1.43       -6.52 2.56e- 10 -12.2       -6.52
5 age_c           0.144    0.0889      1.62 1.07e-  1  -0.0312     0.318
```

A complete response will describe the effect on anxiety associated with a change from low to medium to high in trauma, holding sex and age_c constant (briefly, the effect of trauma being medium instead of low is associated with a 9.7 point increase in the predicted anxiety score, with 95% CI 6.3, 13.1), and comparing high to low we see a 17.9 point increase (95% CI 14.2, 21.7) in predicted anxiety.

I didn't ask you to do this, but you might have been interested in:

- describing the effect of a change from female to male (holding the other terms constant), as being associated with a drop of 9.3 points (95% CI 6.5, 12.2) in predicted anxiety, and
- describing the effect of adding a year to age, holding trauma and sex constant, as being associated with an increase of 0.14 point (95% CI -0.03, 0.32) in predicted anxiety.

1.2 Grading 1 (out of 6 points)

Points Awarded	0	1	2	3	4	5	**6**
Students	0	0	5	3	8	8	**14**

Points Awarded: 76.8% of those available.

Ways to lose points included:

- failing to demonstrate that you calculated the point estimates and/or the measures of uncertainty correctly
- failing to compare the levels appropriately (the direct estimates are high vs. low and medium vs. low, not, for example, high vs. medium.)
  - you needed to interpret both comparisons of trauma levels, not just one
  - the levels of the trauma variable are low, medium and high, and not anything else, like "traumalow", or "middle" or "lesser".
  - the units of the outcome are in points, but definitely not percentages of anything.
- using language like "caused" instead of a statement about predictions or associations
  - a related problem was writing that the anxiety scale score increased without specifying something about the predicted value increasing.
- failing to specify the direction of the effect, as well as its magnitude
- failing to specify a confidence interval, and/or failing to specify a confidence level for that confidence interval
  - those who specified a standard error got away with it, but only so long as they labeled the SE appropriately (just +/- was certainly insufficient), in general, though, I would suggest that as a sub-optimal strategy compared to calculating a proper interval estimate
- failing to specify the other variables `sex` and `age` / `age_c` that must be held constant for the interpretation to hold
  - writing "otherwise identical" without specifying the variables was a problem
  - I let people get away with a description of three mythical people of the same sex and age, but you really should have been clearer that the same would hold for any three people who had the same age and sex, regardless of what those age and sex

values were.

- Some folks referred to both “baseline” and “reference” groups. Stick to a single name. Also, “sex” and “gender” are different things. Don’t interchange the terms.
- Some folks tried to interpret the intercept, which could work, but only if you specified the correct values of `sex` and `age_c`.
- Lots of people didn’t round off the estimates very much, if at all. Round, a lot.

In general, I didn’t deduct multiple points for repetitiveness or run-on sentences, so long as you didn’t add anything that was incorrect. - Many people tried to sneak in a thought about statistical significance by comparing the confidence interval to zero, or in some way noting that the two confidence intervals (for medium vs. low and for high vs. low) did not overlap. This (especially the latter of these) was not a productive strategy.

Two of the answers that I gave full credit to follow…

> For two subjects A and B, who are the same sex and age, but subject A has a trauma category “low” and subject B has a trauma category “medium”, this model predicts that subject B’s anxiety score will be 9.712 points higher than subject A’s, with a 95% confidence interval of 6.306, 13.117. Similarly, for two subjects A and C, who are also the same sex and age, but subject A has a trauma category of “low” and subject B has a trauma category of “high”, this model predicts that subject C’s anxiety score will be 17.949 points higher than subject A’s, with a 95% confidence interval of 14.217, 21.681.
>
> An individual moving from low to medium trauma would result in a predicted increase of anxiety score by 9.71 points with a 95% CI (6.31, 13.12) compared to an individual remaining in the low trauma group, with all other covariates held equal. To go from low to high trauma, the predicted anxiety score would increase by 17.95 points with a 95% CI (14.22, 21.68) over an individual with identical sex and age_c but in the low trauma group, so long as everything else is kept constant.

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.

2.1 Answer 2

Here’s the model:

```
# A tibble: 7 x 7
  term             estimate std.error statistic  p.value conf.low conf.high
  <chr>               <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)        50.1      1.75      28.6   9.21e-93  46.6      53.5
2 traumamedium       12.6      2.35       5.37  1.48e- 7   7.99     17.2
3 traumahigh         26.1      2.50      10.5   2.15e-22  21.2      31.1
4 sexmale            -1.37     2.61      -0.526 5.99e- 1  -6.50      3.76
5 age_c               0.146    0.0865     1.69  9.12e- 2  -0.0236    0.316
6 traumamedium:se~   -6.70     3.37      -1.99  4.75e- 2 -13.3      -0.0754
7 traumahigh:sexm~  -17.8      3.68      -4.83  2.02e- 6 -25.0     -10.5
```

The key thing I was looking for here was an appropriate interpretation of the interaction term. It appears that in these data, the impact of higher levels of trauma is much more strongly associated with increases in predicted anxiety among females, than among males. If you got that, with an appropriate indication as to the point estimates in this situation, then that was the goal.

A technical point: Several people wanted to use `ols` in this setting, but there is a reason I asked you to do it the way I did (in particular, requiring `low` as the baseline level of the `trauma` factor.) I did that in order to be sure that you can either (a) interpret `lm` output appropriately, without the "ease" of an `ols` approach, or (b) figure out on your own how to get `ols` to do what you want. `ols` does not recognize the specification of something like `fct_relevel` in deciding which level of a multi-categorical variable to use as a category, so that was meant to push you towards `lm`, which avoids that problem.

2.2 Grading 2 (out of 6 points)

Points Awarded	0	1	2	3	4	5	**6**
Students	0	0	4	8	18	8	**0**

Points Awarded: 63.2% of those available.

This didn't go well, and I'd feared it wouldn't. Interpreting interactions can be very tricky, and it's clear that I should have spent more time on this, earlier in the course. For full credit, you needed to:

- state explicitly that the effect of increasing levels of anxiety was higher in females than it was in males, holding age constant (failing to do this lost 2 points, and most people failed to do this) *and*
- get the numbers right for the two assessments of medium trauma's impact (1 point)
- get the numbers right for the two assessments of high trauma's impact (1 point)
- not write anything down that was incorrect (this could lose 1 or 2 points *per incident*)

Most of the people who scored 4/6 failed to make an explicit statement about the impact of trauma differing by the two sexes, but instead left it to the reader to figure this out from a restatement of all of the coefficients in the model. Some of the 4/6 responses actually did the statement reasonably well, but had other problems.

Some failed to specify the direction of the effect. Many people restricted their analysis to a single sex, which completely missed the point of looking at the interaction effect4. Quite a few people made smaller errors in interpretation, similar to those described as potential pitfalls in assessing question 1. Using `ols` and thus having the wrong "baseline" or "adjusted to" values cost you points, too.

A relatively promising approach (scoring 5/6) was this:

> If we had two subjects that were the same age, and both medium trauma but one was male and one was female we would predict the male to have an anxiety score that is 6.7 points lower than the female. Similarly, if we had two subject of high trauma, and the same age, but one was female and one was male we would predict the male to have an anxiety score 17.78 points lower than that of the female.

which only leaves out an explanation of what happens between males and females when the trauma level is low. Several people had explanations of this type, but most also made some sort of additional mistake, or perhaps only interpreted one or two of the comparisons (trauma levels).

Another "nearly there" response receiving 5/6 was:

> This interaction shows the difference in how trauma levels associate with anxiety score depending on sex. The increase in anxiety score associated with medium trauma is 6.70 less for men than it is for women and the increase in anxiety score associated with high trauma is 17.78 less for men that it is for women, both while holding age at the mean. Thus, being a male lessens the associated increase in anxiety due to trauma level.

This was great, except, again, we don't describe what happens at low trauma levels (the main effect of sex).

A fairly clean response (with just a few small edits from me to fix syntax and grammatical errors) was:

> Model m2 predicts the anxiety measure of a new female subject compared to a male subject of the same age would be: 1.4 higher if both subjects have low trauma, 8.1 higher if both subjects have medium trauma and 19.2 higher if both subjects have high trauma. This indicates there is an interaction between the effect of sex and trauma level on anxiety since the average (predicted) difference in anxiety score between the female and male subjects is not constant but rather increasing as the trauma level increases.

I think that if that answer had not had some (now corrected) grammatical errors, it would have received 6/6.

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
```

3.1 Answer 3 is C

- The stepwise model suggests x1, x2, and x5.

```
Start:  AIC=693.41
y ~ x1 + x2 + x3 + x4 + x5


       Df Sum of Sq   RSS    AIC
```

```
- x4    1      13.43 14105 691.55
- x3    1      83.39 14175 692.29
<none>               14092 693.41
- x1    1     279.68 14372 694.36
- x2    1     364.53 14456 695.24
- x5    1    1638.48 15730 707.91

Step:  AIC=691.55
y ~ x1 + x2 + x3 + x5

       Df Sum of Sq   RSS    AIC
- x3    1      79.29 14185 690.39
<none>               14105 691.55
- x1    1     269.71 14375 692.39
- x2    1     382.50 14488 693.57
- x5    1    1627.80 15733 705.93

Step:  AIC=690.39
y ~ x1 + x2 + x5

       Df Sum of Sq   RSS    AIC
<none>               14185 690.39
- x1    1      281.2 14466 691.34
- x2    1      389.0 14574 692.45
- x5    1    12103.8 26289 780.94


Call:
lm(formula = y ~ x1 + x2 + x5, data = quiz2B)

Coefficients:
(Intercept)           x1           x2           x5
   -12.5278      -0.0745      -0.4912       0.2960
```

3.2 Grading 3: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

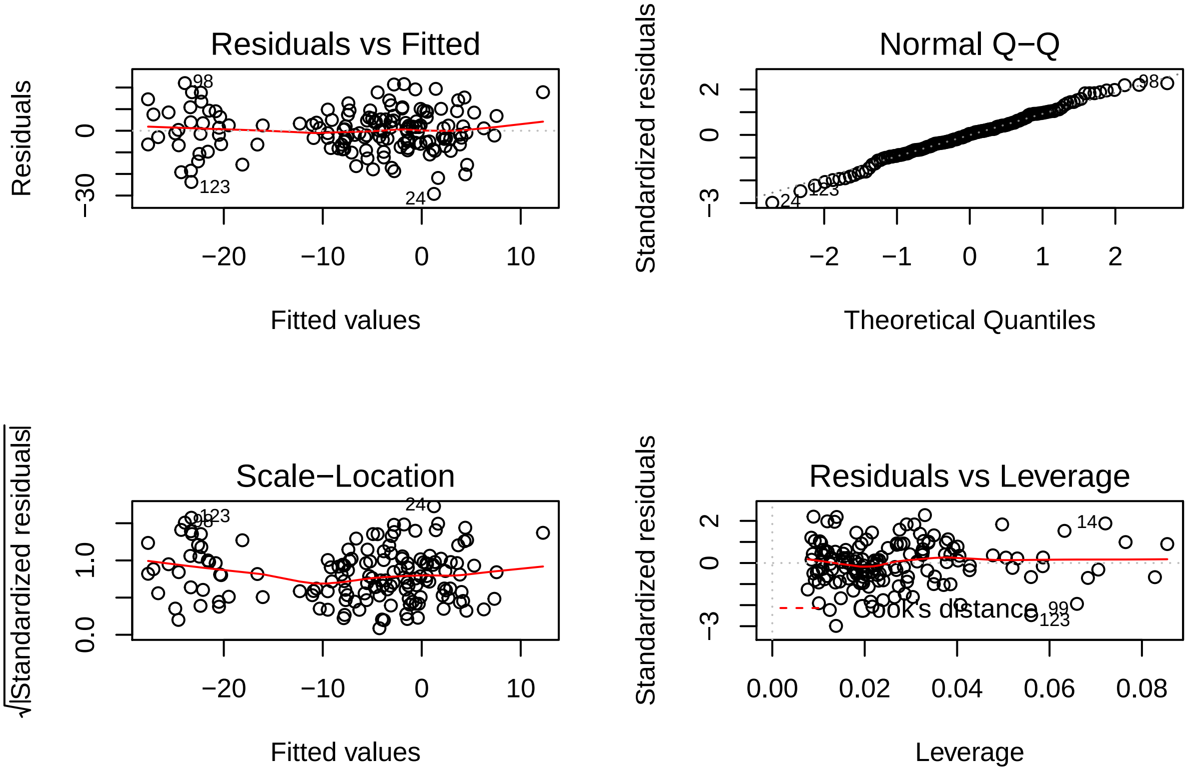
4 Question 4

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
```

4.1 Answer 4 is E

Let's draw the regression diagnostics plots.



- There's no substantial problem with Non-Normality of errors, certainly, from the Normal Q-Q plot on the top right.
- None of the points are highly leveraged or influential, from the plot on the bottom right.
- There's no clear indication of non-linearity in the residuals vs. fitted plot. We clearly have two groups of fitted values (those below about -20 and those between -10 and +10), but no real curve there.
- The scale-location plot doesn't really suggest much of a problem with the assumption of

constant variance, either. Nor is there any meaningful fan shape in the plot of residuals vs. fitted values.

- What about collinearity? Let's look at the variance inflation factors.

```
      x1       x2       x5 
1.187698 1.005022 1.187434 
```

So, my conclusion is that there are no serious violations of regression assumptions, nor is there substantial collinearity in this model.

4.2 Grading 4: no partial credit / 3 available points

Points Awarded: 60.5% of those available.

The most common incorrect response was d. Heteroscedasticity of errors. If you look at that scale-location plot and declare heteroscedasticity to be a substantial problem, then I think you need to recalibrate a bit. Sorry.

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.
```

5.1 Answer 5 is C

```
# A tibble: 1 x 1
  RMSE_model
       <dbl>
1       9.97
```

- The answer is 9.97, which is between 9 and 11.
- Note that this is a little larger than the original estimated residual standard error, which, as you can see below, was 9.86.

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.481         0.471  9.86      45.2 1.04e-20     4  -554. 1118. 1133.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

5.2 Grading 5: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

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?

6.1 Answer 6 is 97 rows.

```
[1] 903   6
```

We see that we have 903 rows with complete data. Since we started with 1000 rows, we have 97 with missing data.

6.2 Grading 6: no partial credit / 3 available points

Points Awarded: 86.8% of those available.

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.

7.1 Answer 7

You'll need to have:

- done multiple imputation, and
- included alive, treated, age, female and comor in the imputation model, and
- fit the outcome model using `fit.mult.impute`

Here's what I used.

```
set.seed(43237)

d <- datadist(quiz2C)
options(datadist = "d")

imp_fit7 <- aregImpute(~ alive + treated + age + female + comor,
                       nk = c(0,3), tlinear = TRUE, data = quiz2C,
                       B = 10, n.impute = 20, pr = FALSE)

m7_imp <- fit.mult.impute(alive ~ treated + age + female + comor,
                          fitter = lrm, xtrans = imp_fit7,
                          data = quiz2C, x = T, y = T)
```

The result of applying this is:

```
set.seed(43237)

d <- datadist(quiz2C)
options(datadist = "d")
```

```
imp_fit7 <- aregImpute(~ alive + treated + age + female + comor,
                        nk = c(0,3), tlinear = TRUE, data = quiz2C,
                        B = 10, n.impute = 20, pr = FALSE)

m7_imp <- fit.mult.impute(alive ~ treated + age + female + comor,
                           fitter = lrm, xtrans = imp_fit7,
                           data = quiz2C, x = T, y = T)
```

```
Variance Inflation Factors Due to Imputation:

Intercept   treated       age    female     comor
     1.11      1.11      1.22      1.05      2.83

Rate of Missing Information:

Intercept   treated       age    female     comor
     0.10      0.10      0.18      0.04      0.65

d.f. for t-distribution for Tests of Single Coefficients:

Intercept   treated       age    female     comor
  2102.80   1914.56    587.89   9632.75     45.52

The following fit components were averaged over the 20 model fits:

  stats linear.predictors
```

```
m7_imp
```

```
Logistic Regression Model

 fit.mult.impute(formula = alive ~ treated + age + female + comor,
     fitter = lrm, xtrans = imp_fit7, data = quiz2C, x = T, y = T)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           1000    LR chi2     347.39    R2       0.429    C       0.856
  0             738    d.f.             4    g        2.073    Dxy     0.712
  1             262    Pr(> chi2) <0.0001    gr       7.980    gamma   0.712
 max |deriv| 8e-06                           gp       0.276    tau-a   0.275
                                             Brier    0.122

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept  6.1439 0.6471   9.49 <0.0001
```

```
treated    1.3984 0.1970   7.10 <0.0001
age       -0.1863 0.0154 -12.09 <0.0001
female    -0.0457 0.1801  -0.25 0.7995
comor      0.3690 0.1052   3.51 0.0005
```

7.2 Grading 7 (out of 6 points)

Points Awarded	0	1	2	3	4	5	**6**
Students	0	0	3	3	4	10	**18**

Points Awarded: 82.9% of those available.

- There was no reason to filter to complete cases on `alive` before doing this imputation.
- You needed to use `lrm` as the fitter in `fit.mult.impute` rather than `glm`.
- You needed to use the `set.seed` seed that I specified, which was 43237.
- You needed to include precisely the variables I did in each model.
- You needed to use the data frame called `quiz2C` and not something else, like `quiz2c`.
- You needed to fit both `aregImpute` and `fit.mult.impute`.

8 Question 8 (6 points)

Using your model 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 a complete English sentence or two.

8.1 Answer 8 is 4.05, with 95% CI (2.75, 5.96) for the odds ratio.

To receive full credit, you'd have to describe this as an odds ratio appropriately, mentioning the key predictor (and the direction of the effect) and the outcome, and correctly specify the variables that are adjusted for in the model.

We can read the odds ratio estimate comparing `treated` to `control`, while adjusting for the effects of `age`, `female` and `comor` using the summary of the imputation model displayed below.

```r
summary(m7_imp)
```

```
             Effects              Response : alive
```

```
Factor       Low High Diff. Effect    S.E.    Lower 0.95 Upper 0.95
treated       0   1    1     1.398400 0.19704  1.012200   1.78460
 Odds Ratio   0   1    1     4.048800      NA  2.751700   5.95730
age          43  57   14    -2.607800 0.21566 -3.030500  -2.18510
 Odds Ratio  43  57   14     0.073697      NA  0.048292   0.11247
female        0   1    1    -0.045743 0.18008 -0.398690   0.30721
 Odds Ratio   0   1    1     0.955290      NA  0.671200   1.35960
comor         1   3    2     0.737980 0.21035  0.325700   1.15030
 Odds Ratio   1   3    2     2.091700      NA  1.385000   3.15900
```

8.2 Grading 8 (out of 6 points)

Points Awarded	0	1	2	3	4	5	**6**
Students	0	0	8	2	2	11	**15**

Points Awarded: 76.8% of those available.

- 4.05 higher odds isn't the same thing as 4.05 times the odds, or 4.05 times higher odds, either of which might be OK.
- If you had an answer for the odds ratio that rounded to 4 or 4.1, that was OK.
- Getting the wrong direction for the result was a big problem.
- Trying to force a statistical significance interpretation was not the right path to follow here.
- Some people wrote "4.05" in one place and "4.11" in another, which is a problem.
- If you have a grammatical or syntax error, you lost a point.
- You should have specified the variables that the model adjusts/controls for. If you just wrote "holding everything else constant" without specifying what "everything else" was, you lost a point.
- Numerous people got seriously wrong answers for the odds ratio. My guess there was that you had either built the wrong model in the first place, or misinterpreted which estimate you should be writing about. In either case, this led to losing about 4 points.
- Explicitly reporting the log of the odds ratio (rather than the odds ratio) is a poor choice, but I only removed 1-2 points if you did it completely correctly.
- If you used an example with two subjects of the same sex, but didn't specify that this would work if you picked either sex, I let that go without deleting points here. If you specified (incorrectly) that they both had to be female, then you lost a point.
- Not just in this question, but everywhere, it's a **confidence** interval, and not a *confidential* interval.

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.
```

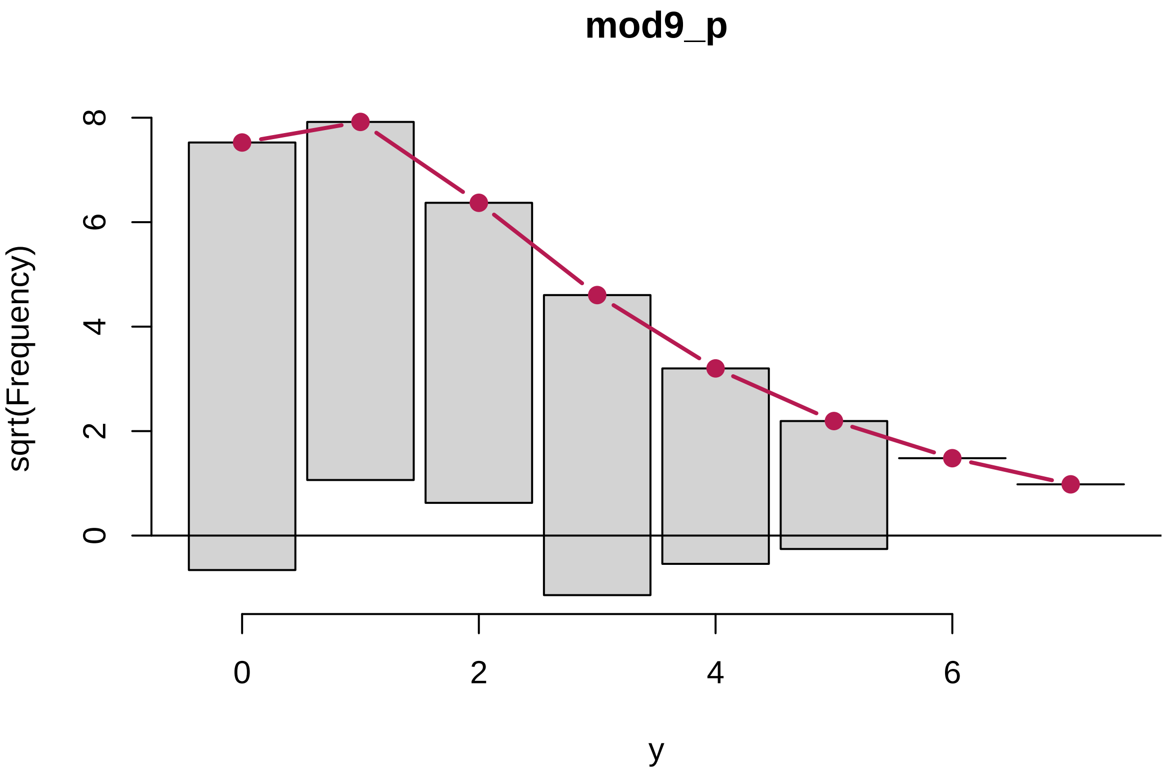
9.1 Answer 9 is H

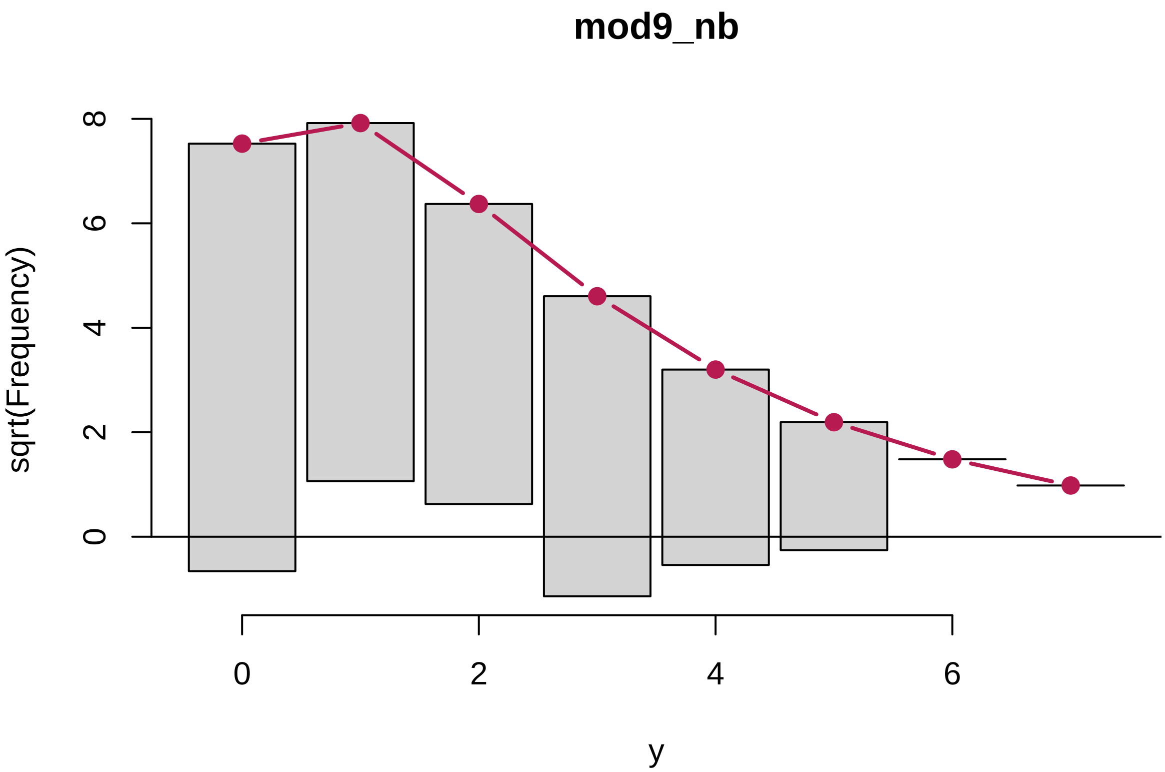
- None of the statements are true. The Poisson and Negative Binomial regression models are nearly identical, and show no significant difference (the difference in the log likelihood functions is essentially zero) in the likelihood ratio test, and there are no meaningful differences between the rootograms.

```
'log Lik.' -296.0254 (df=4)
```

```
'log Lik.' -296.0262 (df=5)
```

```
'log Lik.' 1 (df=5)
```


mod9_p
sqrt(Frequency)
0
2
4
6
8
0
2
4
6
y


mod9_nb
sqrt(Frequency)
0
2
4
6
8
0
2
4
6
y

9.2 Grading 9: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

10 Question 10

Display for Question 10

The three new subjects are Abigail, Brad and Chen

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
```

10.1 Answer 10 is F

Let's make predictions from the Poisson model.

```
        1         2         3 
0.9738522 1.4725208 2.7020633 
```

- Chen is highest, then Brad, then Abigail.

10.2 Grading 10: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

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.
```

11.1 Answer 11 is B

Let's get the predictions from the proportional odds logistic regression model:

```
[1] 0 1 3
Levels: 0 1 2 3 4 5
```

- For Abigail, the polr model predicts 0, and the Poisson predicts 0.97, which rounds to 1.
- For Brad, the polr model predicts 1, and the Poisson predicts 1.47, which rounds to 1.
- For Chen, the polr model predicts 3, and the Poisson predicts 2.70, which rounds to 3.
- So two of the predictions are the same, and one (for Abigail) is different.
- The Poisson model would predict 1.84, which rounds to 2 for Amy, and the polr also predicts 2.
- The Poisson model would predict 0.32, which rounds to 0 for Bart, and the polr also predicts 0.
- The Poisson model predicts 3.18 (which rounds to 3) for Chris, but the polr predicts 4.
- So, only one subject, specifically Chris, gets a new predicted count.

11.2 Grading 11: no partial credit / 3 available points

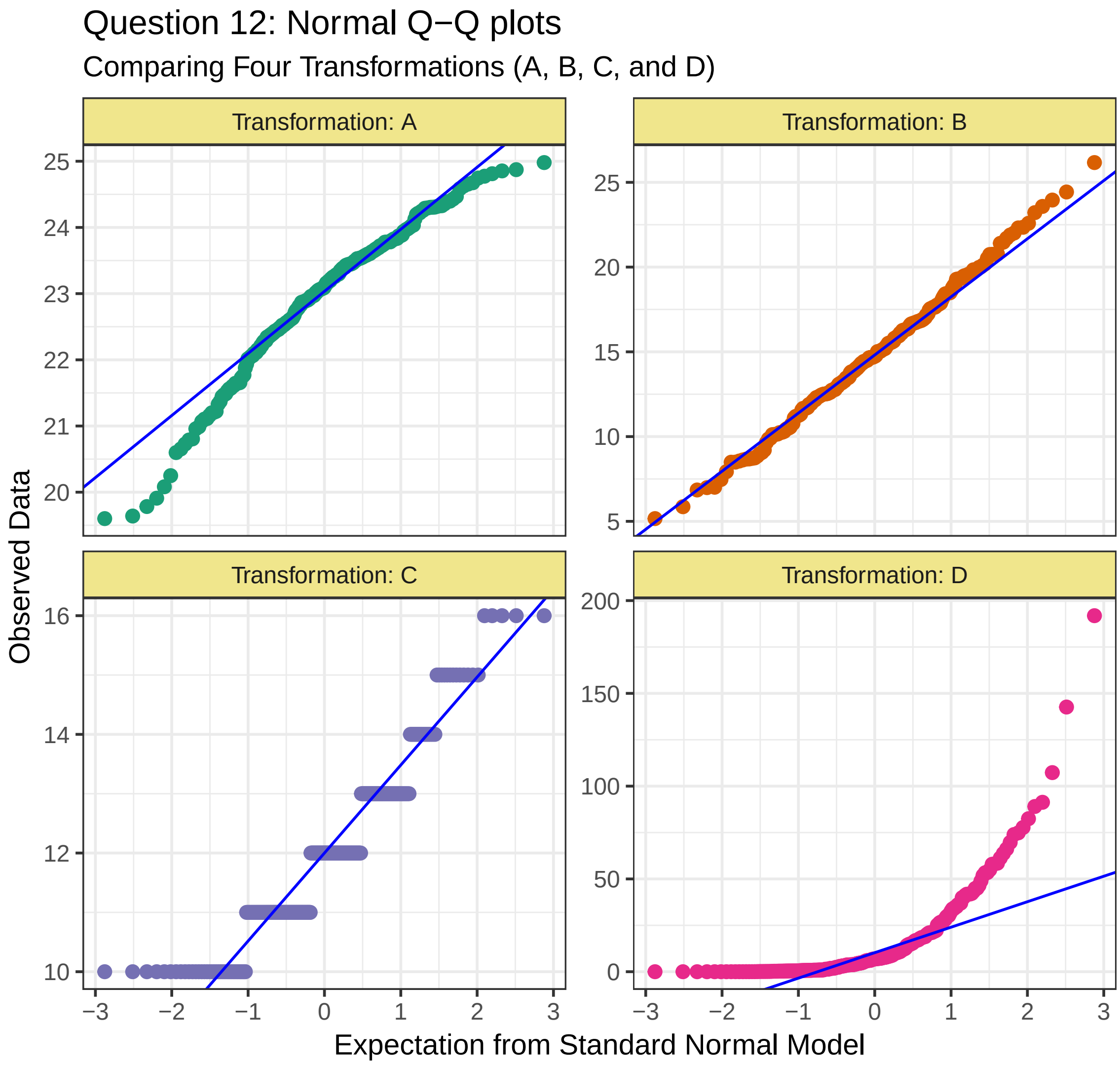
Points Awarded: 65.8% of those available.

- Incorrect responses were more or less evenly dispersed among all other responses.

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.
```

12.1 Answer 12 is B

- The plot of Transformation B describes an approximately Normal distribution, following the 45 degree line in the normal Q-Q plot. The others do not.

12.2 Grading 12: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

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.
```

13.1 Answer 13 is B

- A binary outcome (rehospitalized or not rehospitalized) is what we have, so a plain old binary logistic regression is the best choice of model from these options.

13.2 Grading 13: no partial credit / 3 available points

Points Awarded: 44.7% of those available.

- Everyone who got this wrong selected d. A Cox proportional hazards model, which would imply that we were interested in how many days it took for someone to be rehospitalized, as opposed to what was specfied in the question.
- Sorry. Some questions are deliberately a bit trickier. This was one of those.

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.
```

14.1 Answer 14 is C

- We want to look at the interaction effect for urgency and intervention, so that's either `c` or `d`. In addition to `lrm` giving us the comparison we want (at the 25th and 75th percentiles of `urgency`) the `glm` function doesn't take `x = TRUE, y = TRUE` options, while the `lrm` approach requires them to build the summary estimates of effect size we'd want.

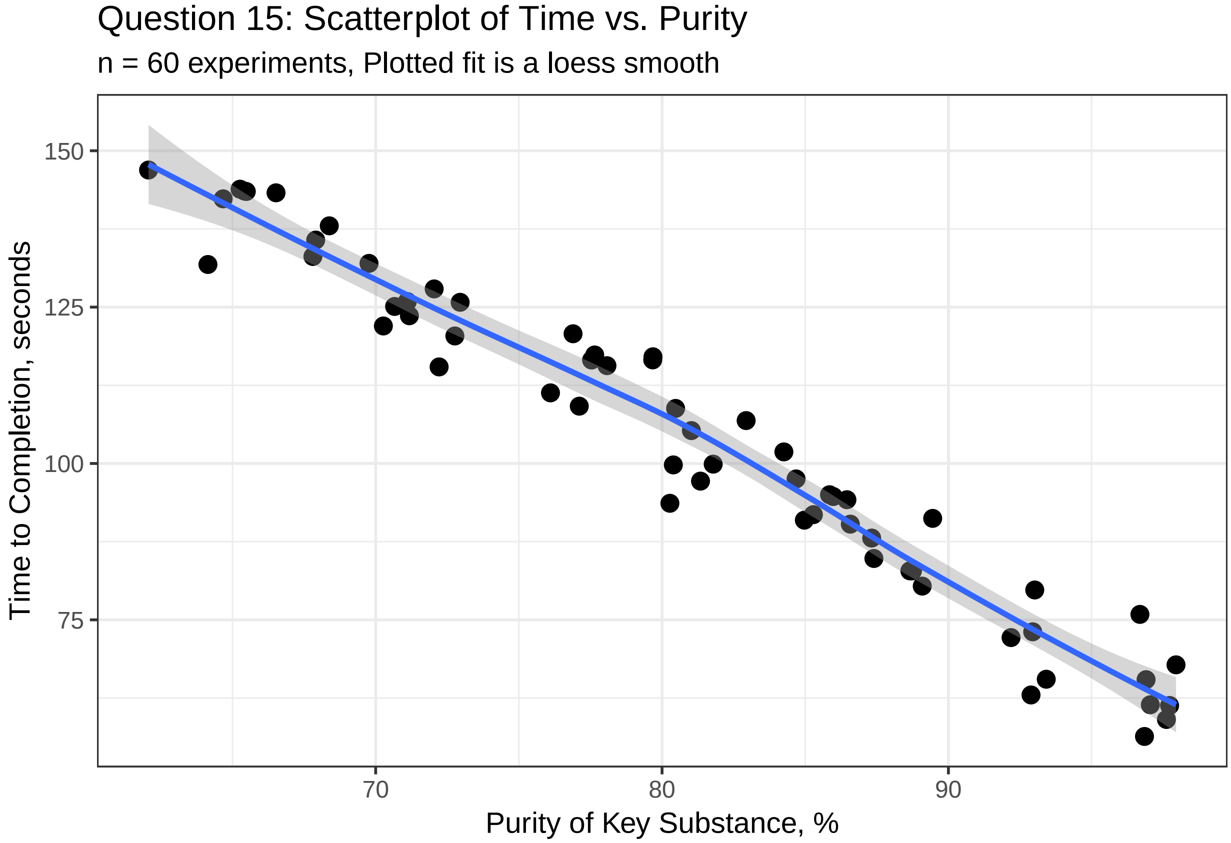
14.2 Grading 14: no partial credit / 3 available points

Points Awarded: 60.5% of those available.

- Most of the incorrect folks chose `a` which doesn't let us look at the interaction term.

15 Question 15

Display for Question 15



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
```

15.1 Answer 15 is C

- The plot shows a highly linear association, with a negative Pearson correlation that is very strong. In fact, the Pearson correlation here is $r = -0.977$, so a simple linear model will account for considerably more than half of the variation in `time`. The actual R^2 for such a model is 0.95.

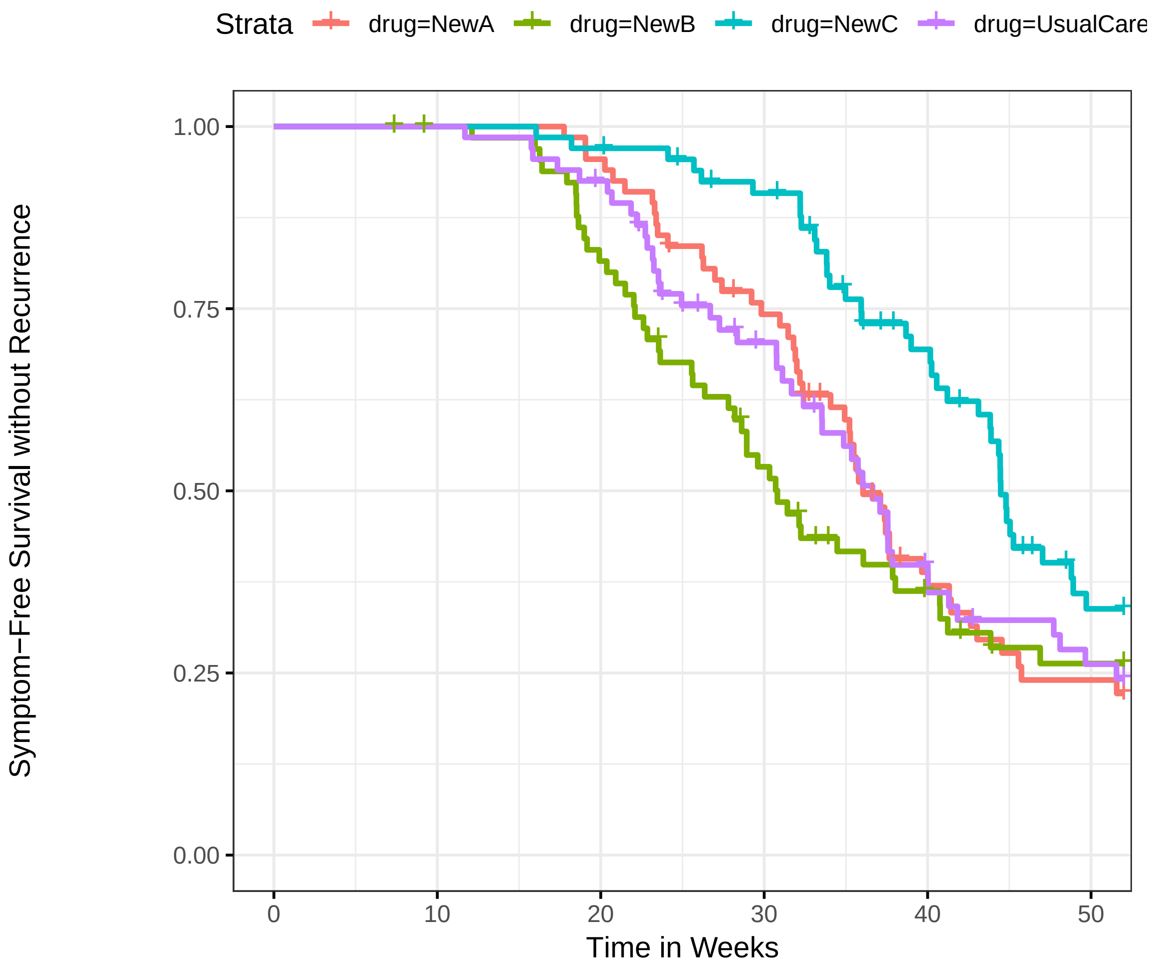
15.2 Grading 15: no partial credit / 3 available points

Points Awarded: 78.9% of those available.

- Incorrect responses were mostly either `a` or `b`. Neither of those options is correct. There is no survival object here, and the model's existing association shows no sign of a substantial curve that might be exploited by a spline.

16 Question 16

Display 1 for Question 16



Number at risk

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

Time in Weeks

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.
```

16.1 Answer 16 is C

- The blue line (for New Drug `C`) in the survival plot displays the best results, in terms of the longest survival times before recurrence of symptoms. We can also see that the comparison across the four drugs shows statistically significant differences ($p = 0.01$) between the drug groups, and that new drug C has a substantially higher median and restricted mean time to recurrence than do the other drugs.

16.2 Grading 16: no partial credit / 3 available points

Points Awarded: 100% of those available. Everyone got this right. Congratulations!

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.
```

17.1 Answer 17 is D

- We're looking for the index-corrected values from the `validate` output. The index-corrected Somers' d is 0.5782, so the index-corrected C statistic is $0.5 + d/2 = 0.79$, and the index-corrected Nagelkerke R-square is 0.3255, so the correct answer is `D`.

17.2 Grading 17: no partial credit / 3 available points

Points Awarded: 94.7% of those available.

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.
```

18.1 Answer 18 is C

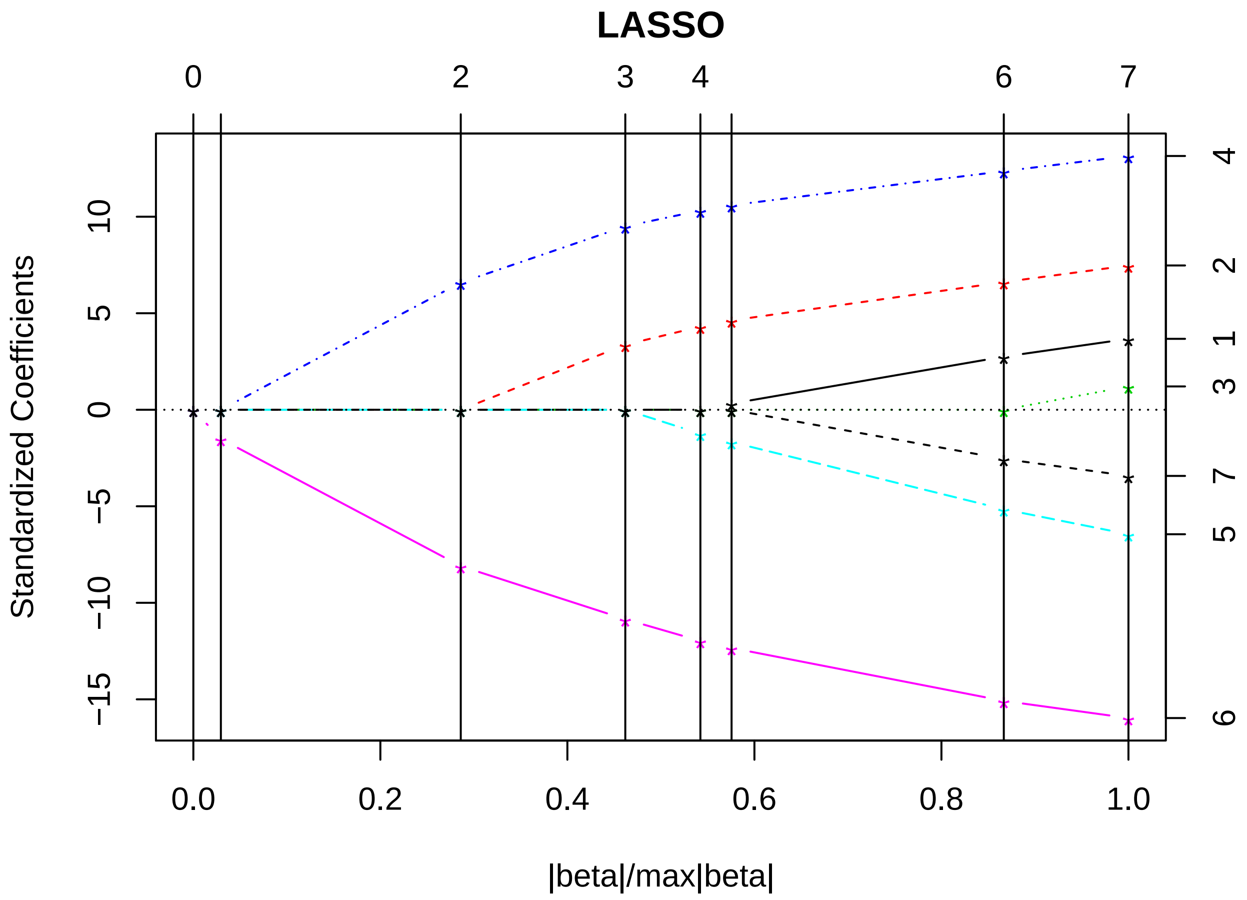
- A proportional odds logistic regression model is used to describe ordered multi-categorical outcomes. The others are not.

18.2 Grading 18: no partial credit / 3 available points

Points Awarded: 92.1% of those available.

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.

19.1 Answer 19 is C

- At the fraction 0.38 on the x-axis ($|\beta|/max(|\beta|)$ scale), three of the seven predictors are included in the model by the lasso, specifically predictors 2, 4, and 6.

19.2 Grading 19: no partial credit / 3 available points

Points Awarded: 60.5% of those available.

- Everyone who got this wrong selected `b`, which is close but not correct. Sorry.

Setup for Question 20

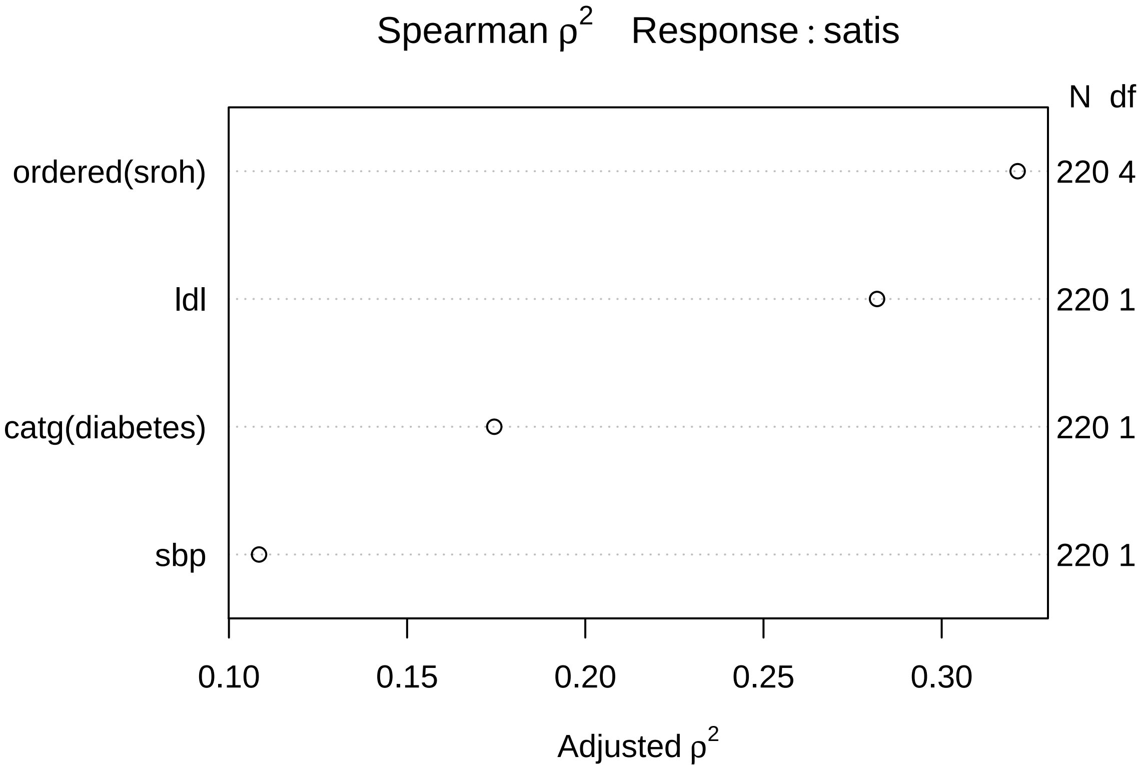
Suppose you plan to fit a model to predict the level of a patient's satisfaction (`satis`, measured on a 0-100 scale, where `satis` = 100 indicates that a patient is extremely satisfied) with their health care, using a sample of 220 subjects. For each subject, you have information on:

- 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.
```

20.1 Answer 20 is B

- We need an ols model, rather than lrm because the outcome is quantitative.
- The Spearman plot indicates that non-linear terms built using (first) sroh (and second) ldl will have the largest impact on the model if they turn out to be useful, so that's where we should start. Since sroh is a 5-category variable, its interaction with ldl will use all four allowed additional degrees of freedom, so we stop there.
- Fitting an ols model including a spline in ldl is the other possibility that might seem attractive, but sroh is well to the right of ldl in the Spearman plot, so this isn't actually the best choice.

20.2 Grading 20: no partial credit / 3 available points

Points Awarded: 73.7% of those available.

- Several folks selected choice d, as I'd anticipated in writing up the sketch above.

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`
```

21.1 Answer 21 is E

- fct_collapse can be used to put D and E together into an "other" category and fct_relevel can be used to resort the resulting levels of that collapsed factor by the costs. All of the other options can only do part of the job.

21.2 Grading 21: no partial credit / 3 available points

Points Awarded: 86.8% of those available.

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)))
```

22.1 Answer 22 is E

- Chunks I and III each accomplish the desired partitioning.

- Chunk II has multiple problems, including not selecting unique observations to go in the two parts of the partition, and using replace = TRUE, rather than FALSE in the test sample, so that individual observations may be repeated in the test sample.

22.2 Grading 22: partial credit awarded / 3 available points

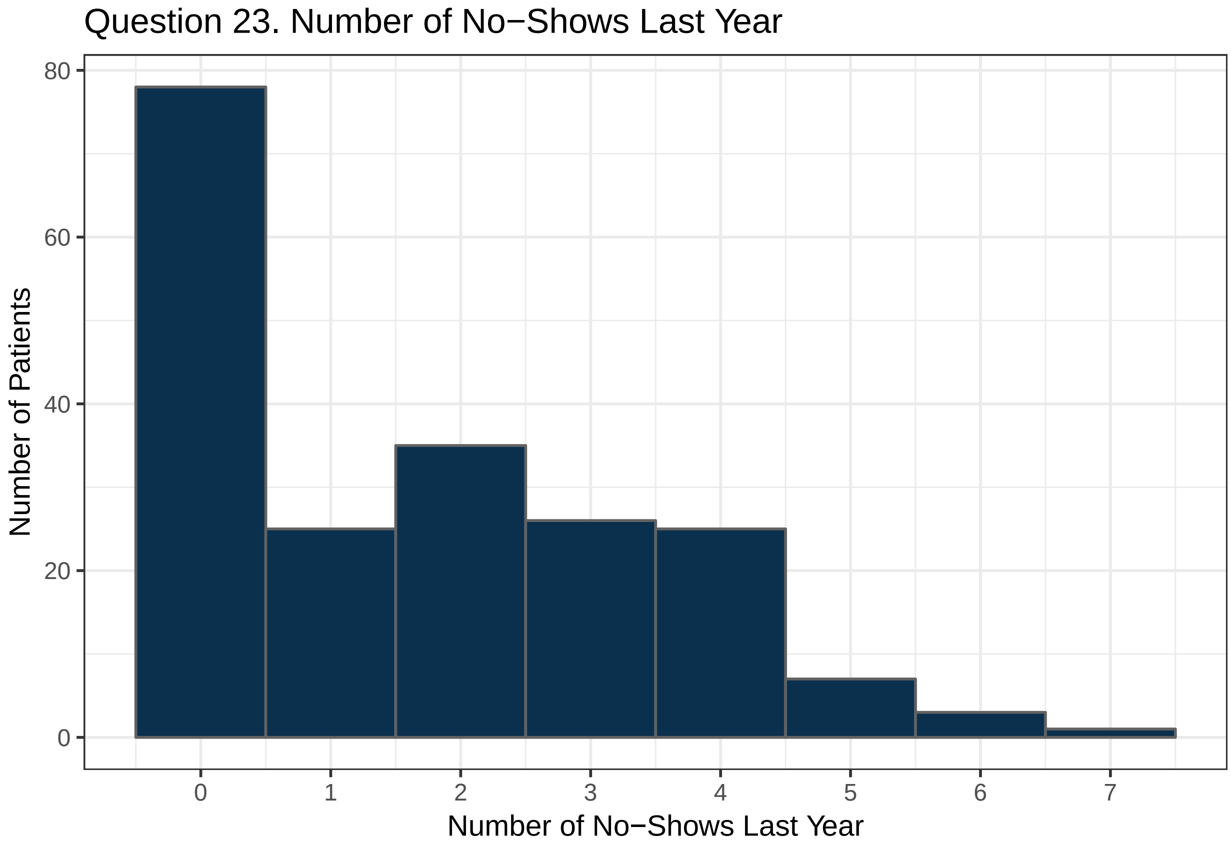
Points Awarded: 86.0% of those available.

Response	a	d	e	g	h
Points	2	1	**3**	2	1
Students	9	2	**25**	1	1

- I gave 1 point for each correct (I, II, III) decision made.

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.
```

23.1 Answer 23 is D

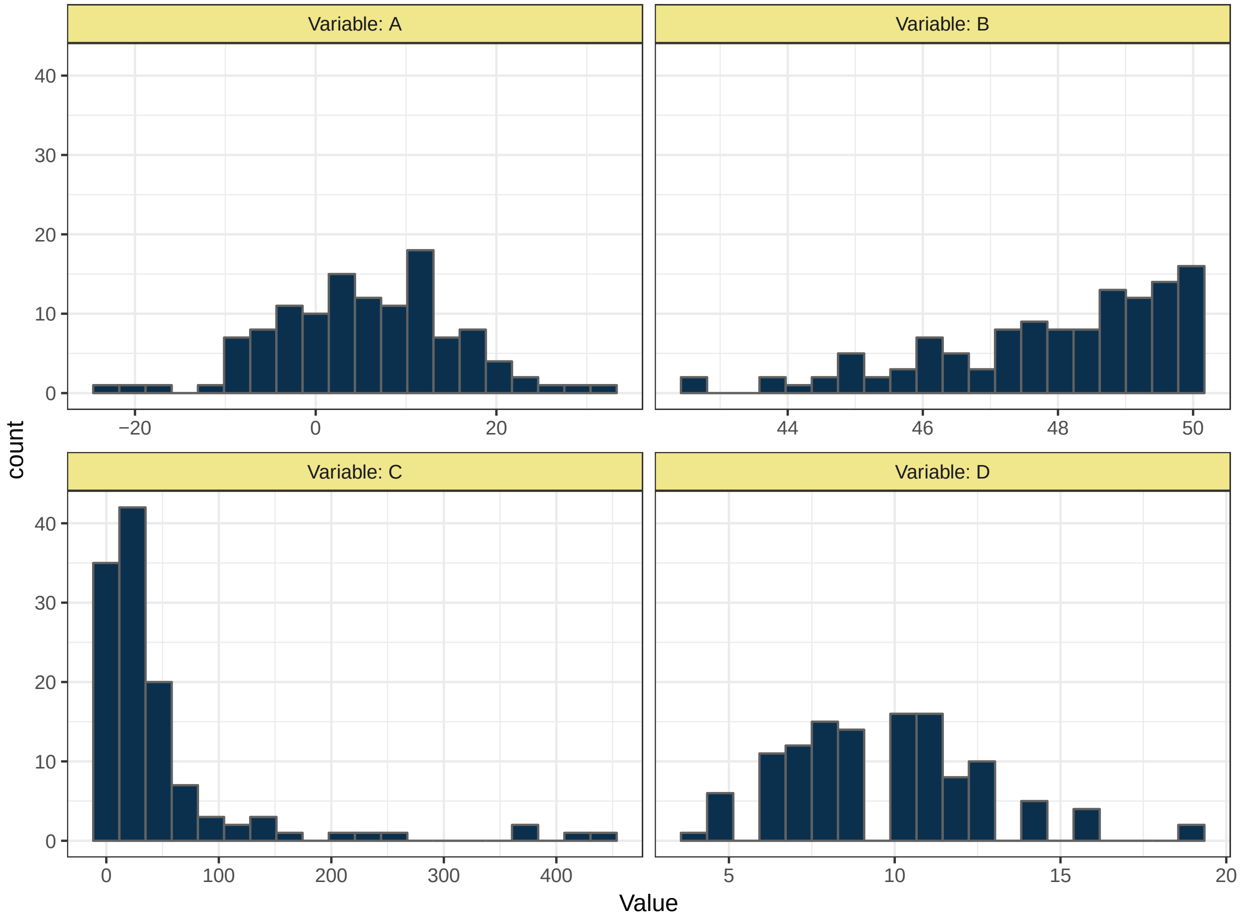
- From the Display, this is a count outcome, with (it appears) some extra zeros. That looks like a zero-inflated Poisson regression would be the best choice.

23.2 Grading 23: no partial credit / 3 available points

Points Awarded: 89.5% of those available.

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.
```

24.1 Answer 24 is C

- The logarithm is an excellent transformation to deal with right skew, in positive values. The histogram of variable C fits those specifications well, but those of the other variables do not.

24.2 Grading 24: no partial credit / 3 available points

Points Awarded: 78.9% of those available.

- Most incorrect responses were either A or D. Neither plot A nor D shows substantial skew.

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
```

25.1 Answer 25 is D

From the scatterplot matrix (Item 3 in the output), we have the following correlations with mass: forearm (0.90), bicep (0.73), chest (0.78), neck (0.80), waist (0.86), height (0.48), so the weakest of these is height.

25.2 Grading 25: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

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.
```

26.2 Answer 26 is A

From the VIF for Model R, the full model (Item 6 in the output) we see at least one VIF exceeding 5, so Model R does show a substantial collinearity problem. From the VIF for the four predictor model (Model S) with forearm, chest, waist and height shown in Item 10 of the output, we see all VIFs below 5. Since Model T is a proper subset of Model S, if Model S has no collinearity problem, Model T cannot, either.

26.3 Grading 26: no partial credit / 3 available points

Points Awarded: 92.1% of those available.

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
```

27.1 Answer 27 is D

The relevant output is in Item 12. The bias-corrected AIC is clearly minimized with the model using five fitted coefficients, including the intercept term.

27.2 Grading 27: no partial credit / 3 available points

Points Awarded: 94.7% of those available.

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
```

28.1 Answer 28 is D

The correct model is the one with 5 coefficients, including the intercept, from Item 12. From item 11, we can see that this model is the one with (the intercept), forearm, chest, waist and height.

28.2 Grading 28: no partial credit / 3 available points

Points Awarded: 92.1% of those available.

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.
```

29.1 Answer 29 is C

See Item 4 of the output. from the t tests, as last predictor in, `forearm` (p = 0.0031), `waist` (p = 0.0006) and `height` (p = 0.0213) are the predictors whose confidence intervals will not meet this standard. The other three (`biceps`, `chest`, and `neck`) will include 0 in their confidence intervals.

29.2 Grading 29: no partial credit / 3 available points

Points Awarded: 89.5% of those available.

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
```

30.1 Answer 30 is D

The relevant material is contained in Item 7 of the output. The model selected by the stepwise backwards elimination procedure includes `forearm`, `chest`, `waist` and `height`.

30.2 Grading 30: no partial credit / 3 available points

Points Awarded: 100% of those available. Congratulations, all!

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.
```

31.1 Answer 31 is C

From Item 12 - the Cp plot suggests that a model with four coefficients (specifically, those in c, plus the intercept) is the best choice.

31.2 Grading 31: no partial credit / 3 available points

Points Awarded: 81.6% of those available.

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
```

32.1 Answer 32 is C

- From the anova output in Item 5, `bicep` doesn't add value when `forearm` is already in the model.
- From Item 9, we see that `chest` adds significant value (p = 0.028) when `forearm` is already in the model.
- From Item 14, we see that `waist` does add significant value (p = 0.0002) when `forearm` is in the model.
- So two of these three variables add significant value after `forearm` is included.

32.2 Grading 32: no partial credit / 3 available points

Points Awarded: 63.2% of those available.

- The most common incorrect response was `b. 1`.

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.
```

33.1 Answer 33 is B

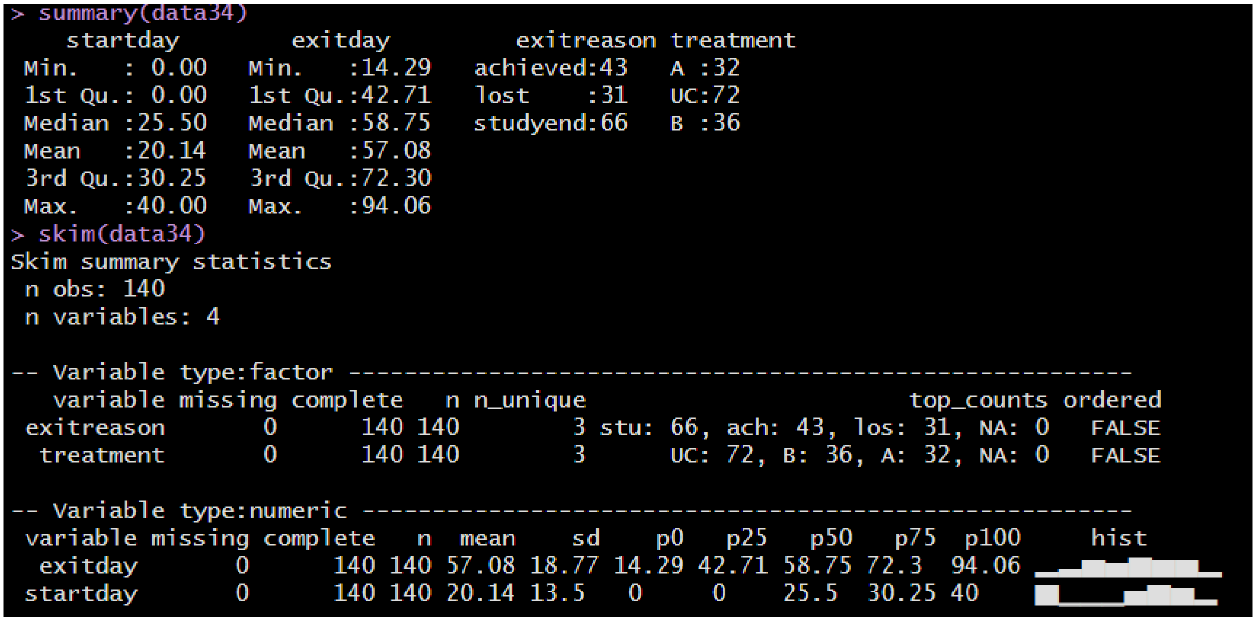
The model suggested by the adjusted R^2 plot in Item 12 includes four predictors: forearm, chest, height and waist. Relevant output is found in Item 20 (summary of effects for model m4). So the estimated effect is an increase of 6.51 kilograms.

33.2 Grading 33: no partial credit / 3 available points

Points Awarded: 97.4% of those available.

34 Question 34

Display 1 for Question 34



```
> summary(data34)
    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(data34)
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)
```

34.2 Answer 34 is C

- Chunk I creates the wrong survival object, flipping the designation of censored and observed times.
- Chunk II doesn't create a survival object, so that won't work.
- Chunk III gets everything right.

34.3 Grading 34: partial credit awarded / 3 available points

Points Awarded: 76.3% of those available.

Response	a	c	d	e	h
Points	1	3	0	2	2
Students	8	21	1	6	2

- I gave 1 point for each correct (I, II, III) decision made.

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
```

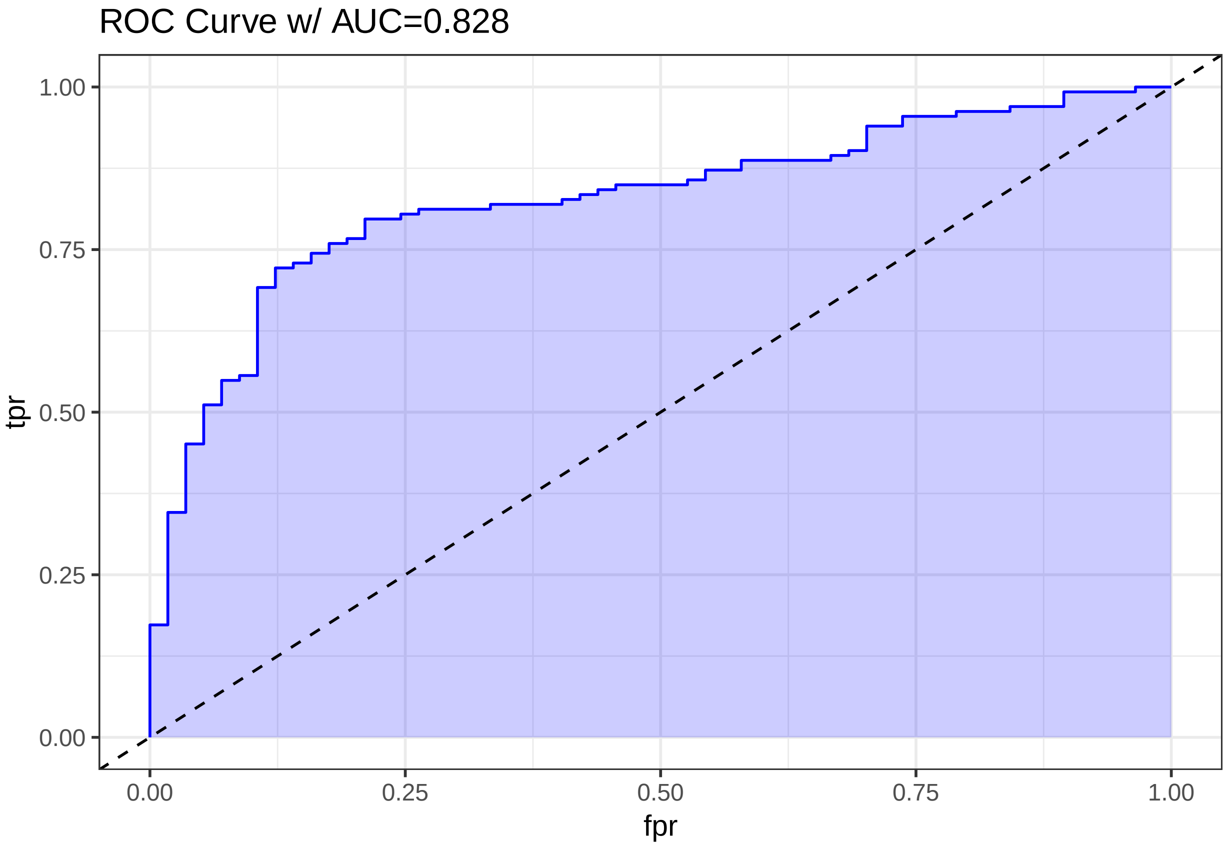
35 Question 35

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?

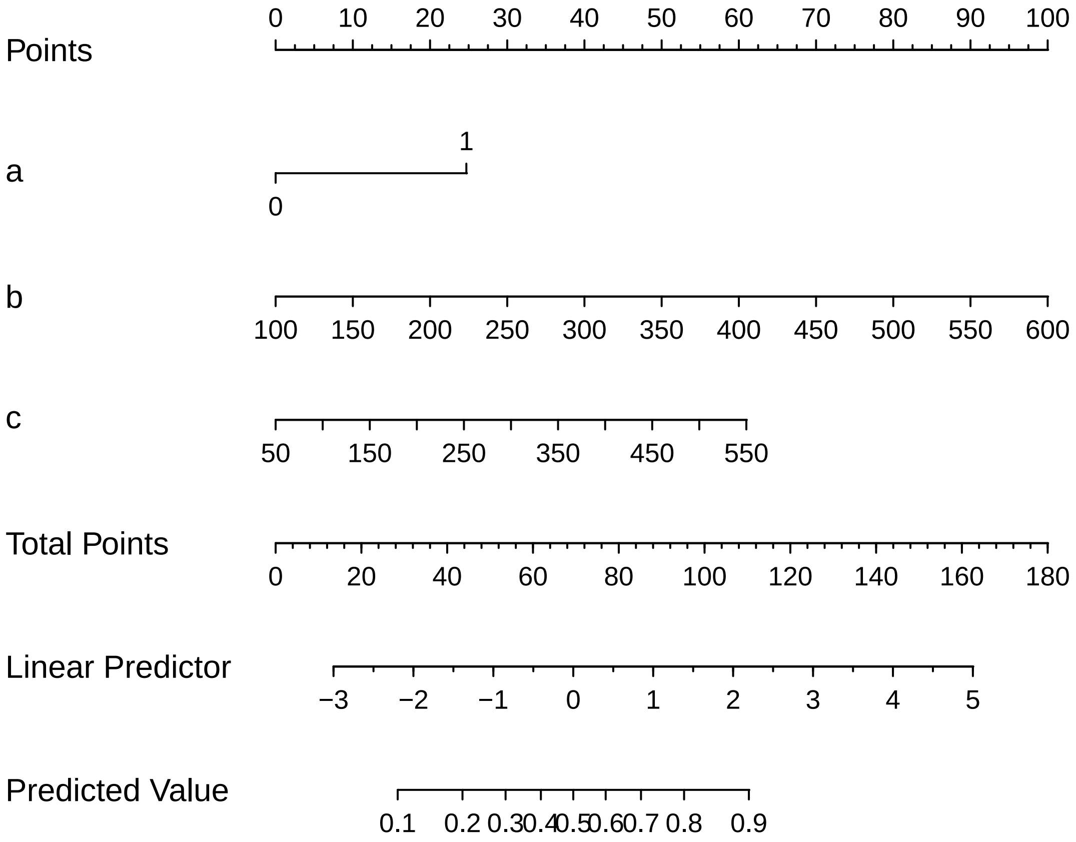
(Hint: the nomograms in Plots B, C, and D all show the estimated probability of the outcome being 1 as the “Predicted Value”.)

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

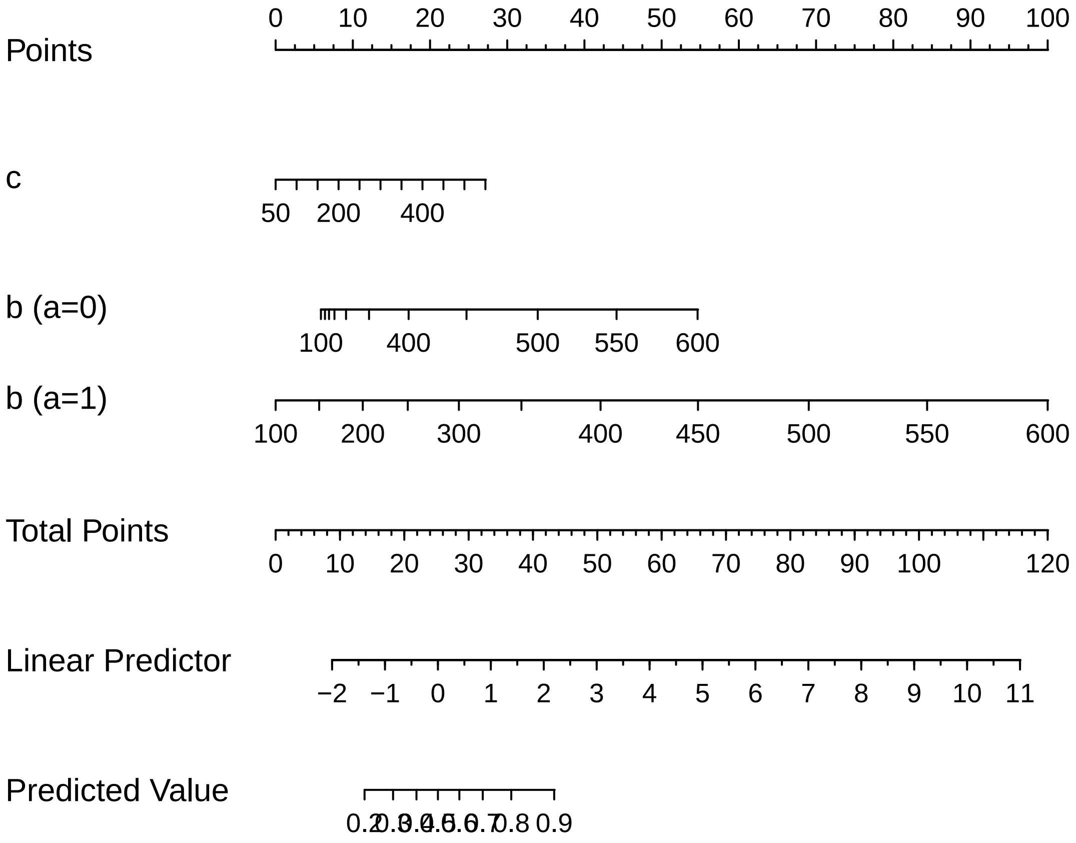
Plot A for Question 35



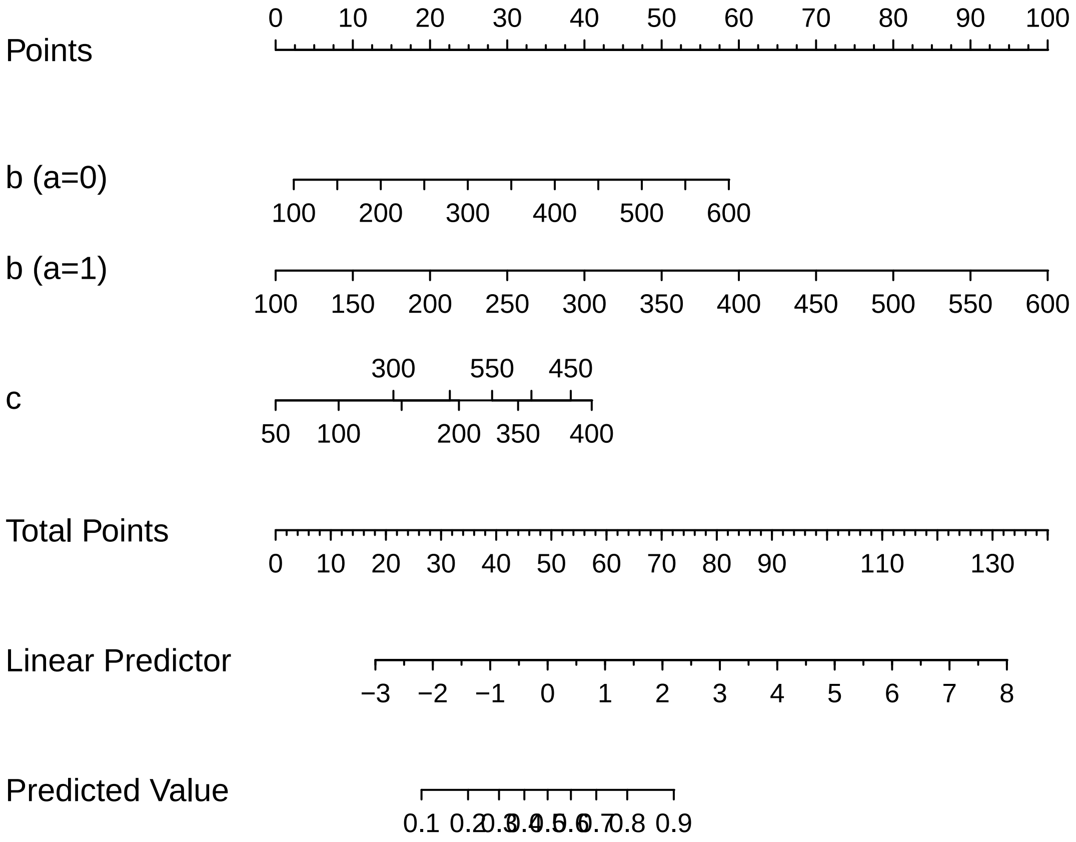
Plot B for Question 35



Plot C for Question 35



Plot D for Question 35



35.1 Answer 35 is C

This can be done by process of elimination.

- Our model contains an interaction in `a` and `b`, and a non-linear term (spline) in `b`, with a C statistic of 0.858.
- Plot A cannot be right, since it shows the wrong ROC value (0.621).
- Plot B cannot be right, because it doesn't show any interaction of `a` and `b`.
- Plot D cannot be right because x3 includes a non-linear term in variable `c`.

So it must be that Plot C, which does show an interaction of `a` and `b`, and a linear effect of `c`, is the correct one. And, it is. I built Plot C from the model shown in Display 1, and the other Plots from other models.

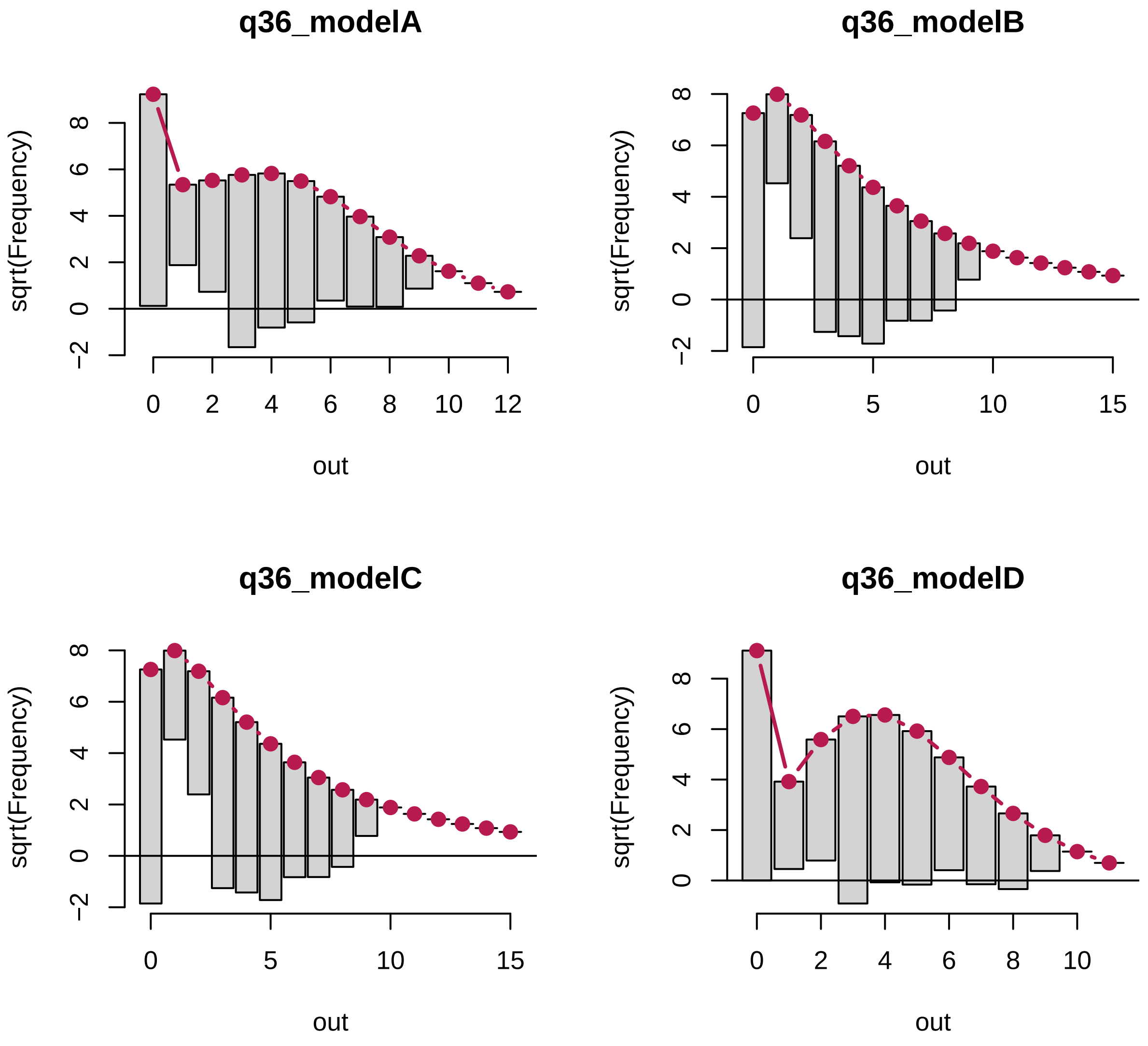
35.2 Grading 35: no partial credit / 3 available points

Points Awarded: 86.8% of those available.

- Most of the incorrect responses were either `d` or `e`.

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.
```

36.2 Answer 36 is D

- Model D clearly shows the best fit to the data, of the four models provided. The modeled counts are very close to the actual counts, across the range.

36.3 Grading 36: no partial credit / 3 available points

Points Awarded: 100% of those available. Again, everyone got this right. Great!

37 Overall Grades

Sum up your score on each of the 36 items on the quiz. The maximum possible score was 120 points. Congratulations to the two students who scored 114/120, which was the maximum observed score. The median score was 101.5, and the mean was 99.5.

Score Range	Rough Letter Grade
108-114	A
102-106	A- or B+
92-101	B
84-90	B-
below 84	Needs Improvement