

# 432 Quiz 2, Answer Sketch and Grading Results

*Thomas E. Love*

*Due 2018-05-01 at Noon. Version: 2018-05-03*

## Contents

0.1	Setup . . . . .	3
<b>1</b>	<b>Question 1</b>	<b>3</b>
1.1	Display for Question 1 . . . . .	3
<b>2</b>	<b>Question 2</b>	<b>5</b>
2.1	Display for Question 2 . . . . .	5
<b>3</b>	<b>Question 3</b>	<b>6</b>
<b>4</b>	<b>Question 4</b>	<b>6</b>
<b>5</b>	<b>Question 5</b>	<b>7</b>
5.1	Display for Question 5 . . . . .	7
<b>6</b>	<b>Question 6</b>	<b>8</b>
6.1	Display 1 for Question 6 . . . . .	8
6.2	Display 2 for Question 6 . . . . .	9
<b>7</b>	<b>Question 7</b>	<b>10</b>
7.1	Display 1 for Question 7 . . . . .	10
7.2	Display 2 for Question 7 . . . . .	10
<b>8</b>	<b>Question 8</b>	<b>11</b>
8.1	Display for Question 8 . . . . .	11
<b>9</b>	<b>Question 9</b>	<b>13</b>
9.1	Display for Question 9 . . . . .	13
<b>10</b>	<b>Question 10</b>	<b>14</b>
10.1	Display for Question 10 . . . . .	14
<b>11</b>	<b>Question 11</b>	<b>14</b>
<b>12</b>	<b>Question 12</b>	<b>15</b>
12.1	Display for Question 12 . . . . .	15
<b>13</b>	<b>Question 13</b>	<b>15</b>
13.1	Display for Question 13 . . . . .	15
<b>14</b>	<b>Question 14</b>	<b>16</b>
<b>15</b>	<b>Question 15</b>	<b>17</b>
<b>16</b>	<b>Question 16</b>	<b>17</b>
16.1	Display for Question 16 . . . . .	17

<b>17 Question 17</b>	<b>18</b>
17.1 Display for Question 17 . . . . .	18
<b>18 Question 18</b>	<b>20</b>
<b>19 Question 19</b>	<b>21</b>
19.1 Display 1 for Question 19 . . . . .	21
19.2 Display 2 for Question 19 . . . . .	22
<b>20 Question 20</b>	<b>22</b>
20.1 Display 1 for Question 20 . . . . .	22
20.2 Plot A for Question 20 . . . . .	23
20.3 Plot B for Question 20 . . . . .	24
20.4 Plot C for Question 20 . . . . .	25
20.5 Plot D for Question 20 . . . . .	26
<b>21 Question 21</b>	<b>27</b>
21.1 Display for Question 21 . . . . .	27
<b>22 Question 22</b>	<b>28</b>
22.1 Display for Question 22 . . . . .	28
<b>23 Question 23</b>	<b>28</b>
<b>24 Question 24</b>	<b>30</b>
24.1 Display for Question 24 . . . . .	30
<b>25 Question 25</b>	<b>31</b>
<b>26 Question 26</b>	<b>31</b>
26.1 Display for Question 26 . . . . .	31
<b>27 Question 27</b>	<b>31</b>
<b>28 Question 28</b>	<b>32</b>
<b>29 Question 29</b>	<b>32</b>
<b>30 Question 30</b>	<b>32</b>
<b>31 Question 31</b>	<b>32</b>
<b>32 Question 32</b>	<b>33</b>
<b>33 Question 33</b>	<b>33</b>
<b>34 Question 34</b>	<b>33</b>
<b>35 OPTIONAL BONUS Question 35</b>	<b>34</b>
<b>36 OPTIONAL BONUS Question 36</b>	<b>34</b>
<b>37 Answers</b>	<b>35</b>
37.1 Answer 1 is d . . . . .	35
37.2 Answer 2 is d . . . . .	35
37.3 Answer 3 is c . . . . .	35
37.4 Answer 4 is c . . . . .	36

37.5 Answer 5 is d . . . . .	36
37.6 Answer 6 is a . . . . .	36
37.7 Answer 7 is technically b but I also accepted f . . . . .	37
37.8 Answer 8 is e . . . . .	37
37.9 Answer 9 is b . . . . .	38
37.10 Answer 10 is b . . . . .	38
37.11 Answer 11 is e . . . . .	38
37.12 Answer 12 is c . . . . .	40
37.13 Answer 13 is c . . . . .	40
37.14 Answer 14 is d . . . . .	40
37.15 Answer 15 is a . . . . .	41
37.16 Answer 16 is a . . . . .	41
37.17 Answer 17 is e . . . . .	41
37.18 Answer 18 is b . . . . .	41
37.19 Answer 19 is c . . . . .	42
37.20 Answer 20 is b . . . . .	42
37.21 Answer 21 is h . . . . .	42
37.22 Answer 22 is e . . . . .	44
37.23 Answer 23 is d . . . . .	45
37.24 Answer 24 is b . . . . .	45
37.25 Answer 25 is d . . . . .	45
37.26 Answer 26 is a . . . . .	46
37.27 Answer 27 is d . . . . .	46
37.28 Answer 28 is d . . . . .	46
37.29 Answer 29 is c . . . . .	46
37.30 Answer 30 is d . . . . .	47
37.31 Answer 31 is c . . . . .	47
37.32 Answer 32 is c . . . . .	47
37.33 Answer 33 is b . . . . .	47
37.34 Answer 34 is 100 rows. . . . .	48
37.35 Answer 35 is a long presentation of R code . . . . .	48
37.36 Answer 36 is 4.8, with 95% CI (3.0, 7.6) for the odds ratio. . . . .	50

## 38 Grades on Quiz 2 50

### 0.1 Setup

## 1 Question 1

### 1.1 Display for Question 1

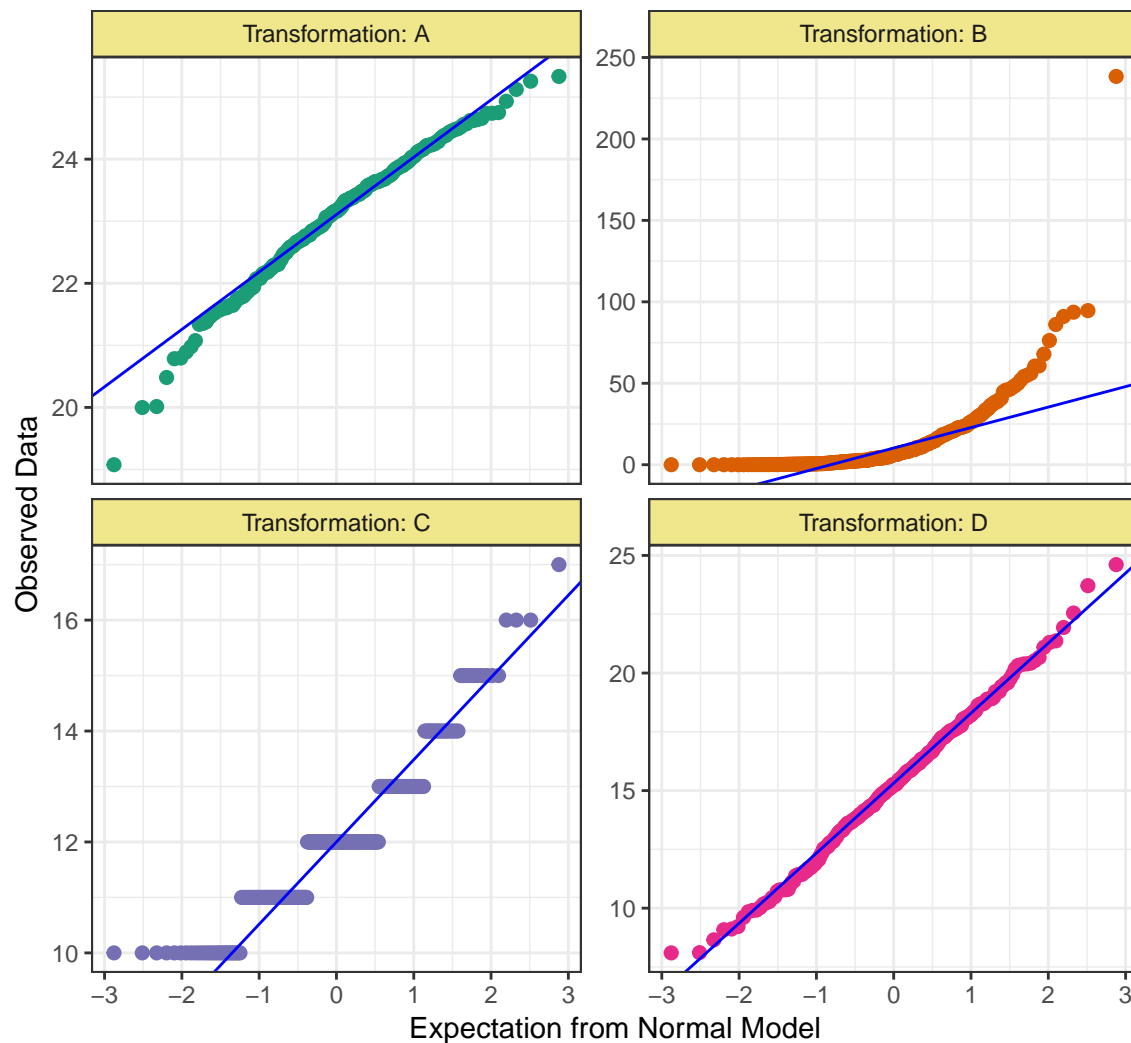
```
ints1 <- data01 %>% group_by(Transformation) %>%
  summarize(q25 = quantile(value, 0.25),
            q75 = quantile(value, 0.75),
            norm25 = qnorm(0.25),
            norm75 = qnorm(0.75),
            slope = (q25 - q75) / (norm25 - norm75),
            int = q25 - slope * norm25) %>%
  select(Transformation, slope, int)

ggplot(data01, aes(sample = value, col = Transformation)) +
  geom_qq(size = 2) +
```

```
geom_abline(data = intsl, aes(intercept = int, slope = slope), col = "blue") +
facet_wrap(~ Transformation, scales = "free_y", labeller = "label_both") +
theme_bw() +
guides(col = FALSE) +
theme(strip.background =element_rect(fill="khaki")) +
scale_color_brewer(palette = "Dark2") +
labs(y = "Observed Data", x = "Expectation from Normal Model",
     title = "Question 1: Normal Q-Q plots",
     subtitle = "Comparing Four Transformations (A, B, C, and D)")
```

## Question 1: Normal Q-Q plots

Comparing Four Transformations (A, B, C, and D)



```
ggsave("displays/display01.png", height = 6, width = 6)
```

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

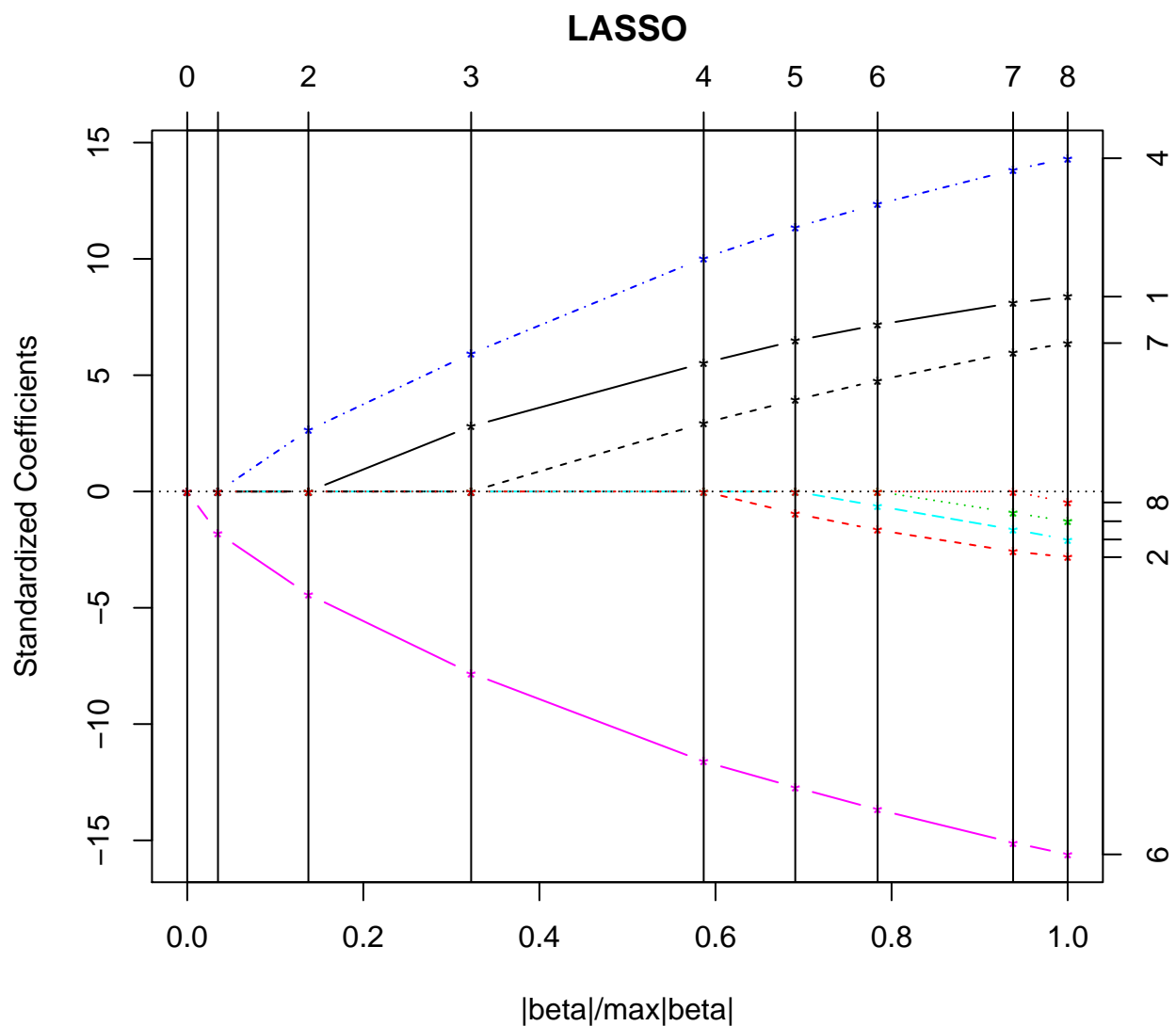
- A
- B

- c. C
- d. D
- e. None of the above.

## 2 Question 2

### 2.1 Display for Question 2

```
preds02 <- with(data02, cbind(x1, x2, x3, x4, x5, x6, x7, x8))
lasso02 <- lars(preds02, data02$y, type = "lasso")
plot(lasso02)
```



```
#set.seed(432002)
#lassocv02 <- cv.lars(preds02, data02$y, K=10)
#frac02 <- lassocv02$index[which.min(lassocv02$cv)]
```

#frac02

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

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

### 3 Question 3

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

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

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

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

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

### 4 Question 4

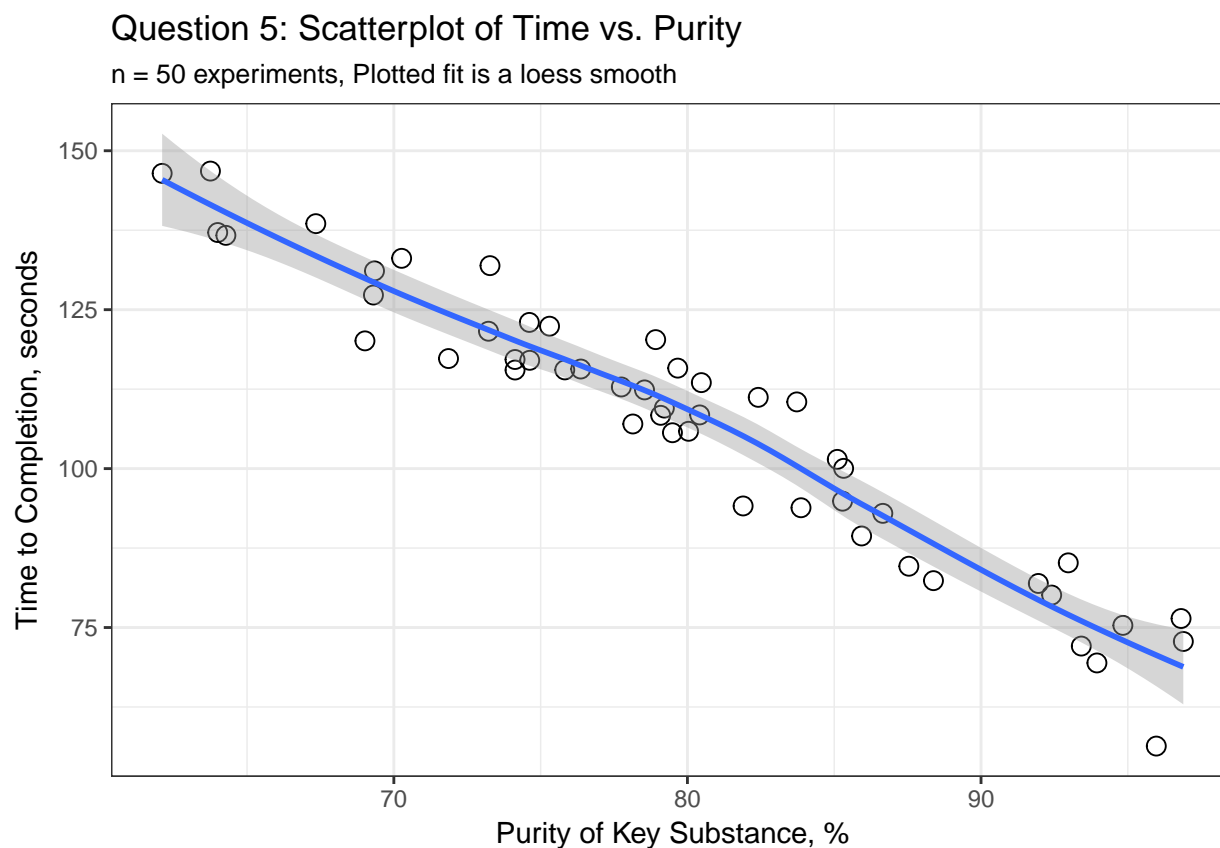
Suppose you are trying to build a regression model to predict a patient’s self-reported overall health (where the available responses are Excellent, Very Good, Good, Fair or Poor) where you want to treat the health assessments as categorical. Which of the following models would be most appropriate?

- a. An ordinary least squares model.
- b. A Cox proportional hazards model.
- c. A proportional odds logistic regression model.
- d. A zero-inflated negative binomial model.
- e. None of these models would be appropriate.

## 5 Question 5

### 5.1 Display for Question 5

```
ggplot(data05, aes(x = purity, y = time)) +  
  geom_point(size = 3, shape = 1) +  
  geom_smooth(method = "loess") +  
  theme_bw() +  
  labs(y = "Time to Completion, seconds", x = "Purity of Key Substance, %",  
        title = "Question 5: Scatterplot of Time vs. Purity",  
        subtitle = "n = 50 experiments, Plotted fit is a loess smooth")
```



```
ggsave("displays/display05.png", height = 6, width = 6)
```

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

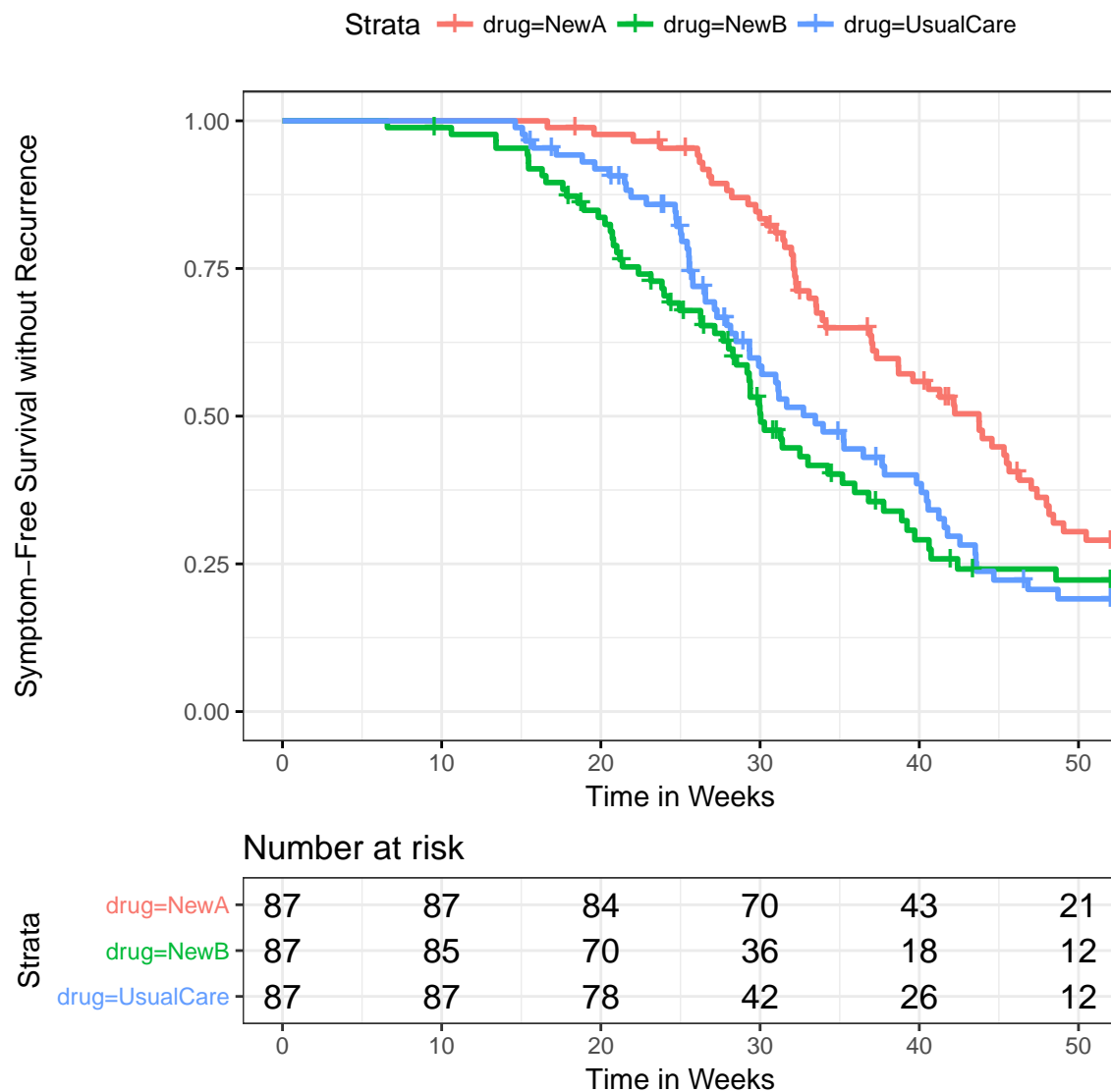
- a. Model 5 will have an R-squared value of about 0.10
- b. Model 5 explains between 25% and 50% of the variation in completion time.
- c. Model 5 is not helpful, since we should be fitting a Cox model instead.
- d. Model 5 explains more than 50% of the variation in completion time.
- e. Model 5 fits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

restricted cubic spline in purity.

## 6 Question 6

### 6.1 Display 1 for Question 6

```
data06$$S <- Surv(time = data06$time, event = data06$recur)
fit06 <- survfit(data06$$S ~ data06$drug)
ggsurvplot(fit06, data = data06,
  risk.table = TRUE,
  risk.table.height = 0.25,
  risk.table.y.text.col = TRUE,
  xlab = "Time in Weeks",
  ylab = "Symptom-Free Survival without Recurrence",
  ggtheme = theme_bw())
```





## 6.2 Display 2 for Question 6

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

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

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

\* restricted mean with upper limit = 52

```
survdifff(data06$S ~ data06$drug)
```

```
Call:
```

```
survdifff(formula = data06$S ~ data06$drug)
```

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

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

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

- Drug A
- Drug B
- The Usual Care drug
- It is impossible to tell from the output provided.

## 7 Question 7

NOTE: THIS QUESTION CONTAINED AN ERROR IN CHUNK III. SEE ANSWERS FOR DETAILS.

### 7.1 Display 1 for Question 7

```
data07

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

### 7.2 Display 2 for Question 7

Chunk I

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

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

## Chunk II

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

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

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

## Chunk III (NOTE THIS IS FIXED IN THE ANSWERS)

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

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

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

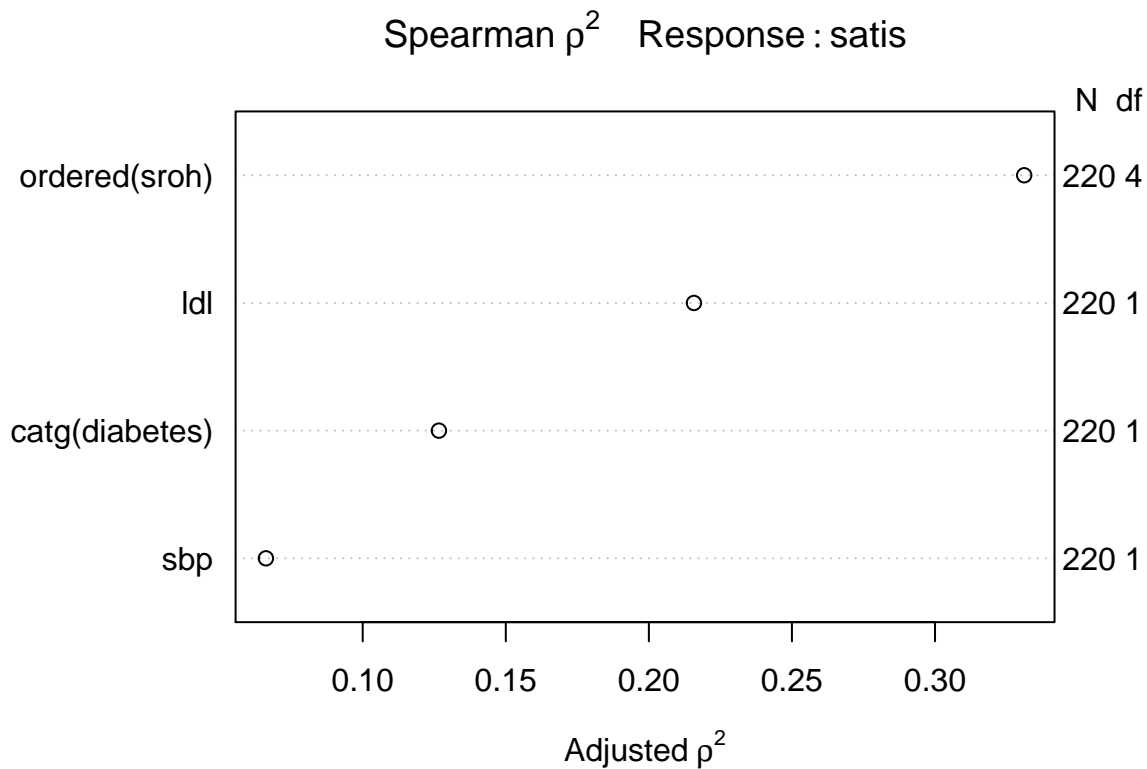
Given the data set `data07` as shown in Display 1 for Question 7, suppose you want to remove all rows containing missing values, then create a training sample containing 80% of the rows without missing data, and a test sample containing the other 20% of the values after missingness is removed. Which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

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

## 8 Question 8

### 8.1 Display for Question 8

```
plot(spearman2(satis ~ sbp + ldl + catg(diabetes) + ordered(sroh), data = data08))
```



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

For each subject, you also have information on

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

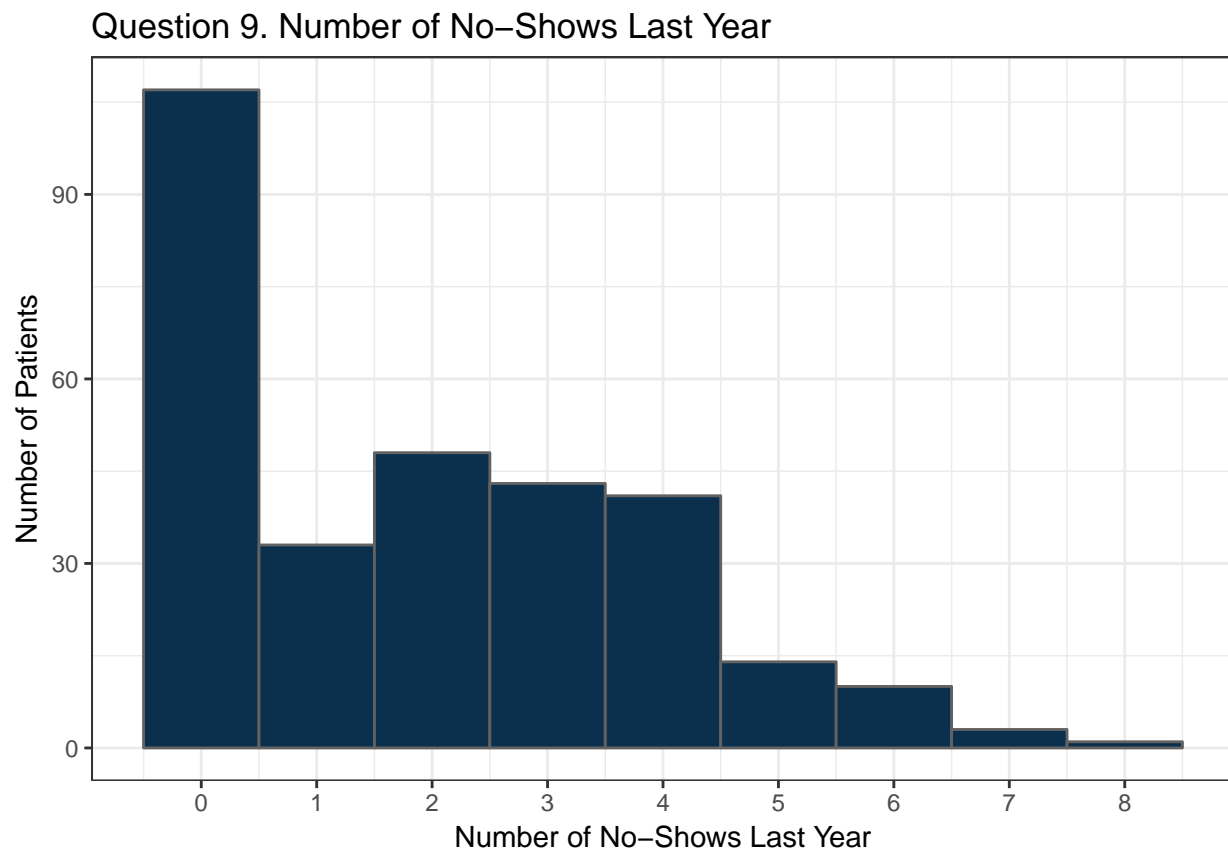
The Display for Question 8 shows a Spearman rho-squared plot for these subjects. Assuming you wish to include all of the main effects for these predictors in your model, and you can afford to add an additional four degrees of freedom to the model, which of the following augmentations to a “main effects”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

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

## 9 Question 9

### 9.1 Display for Question 9

```
ggplot(data09, aes(x = noshow)) +  
  geom_histogram(binwidth = 1, fill = "#0a304e", col = "#626262") +  
  scale_x_continuous(breaks = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)) +  
  theme_bw() +  
  labs(y = "Number of Patients", x = "Number of No-Shows Last Year",  
       title = "Question 9. Number of No-Shows Last Year")
```



```
ggsave("displays/display09.png", height = 5, width = 7)
```

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

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

## 10 Question 10

### 10.1 Display for Question 10

```
data10

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

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

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

## 11 Question 11

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

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

## 12 Question 12

### 12.1 Display for Question 12

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

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

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

## 13 Question 13

### 13.1 Display for Question 13

```
q13_modelA <- glm(out ~ x1,
  family = poisson(),
  data = data13)

q13_modelB <- zeroinfl(out ~ x2 + x3,
  data = data13)

q13_modelC <- zeroinfl(out ~ err1,
  family = poisson(),
  data = data13)

q13_modelD <- MASS::glm.nb(out ~ x1 + x2,
  data = data13)

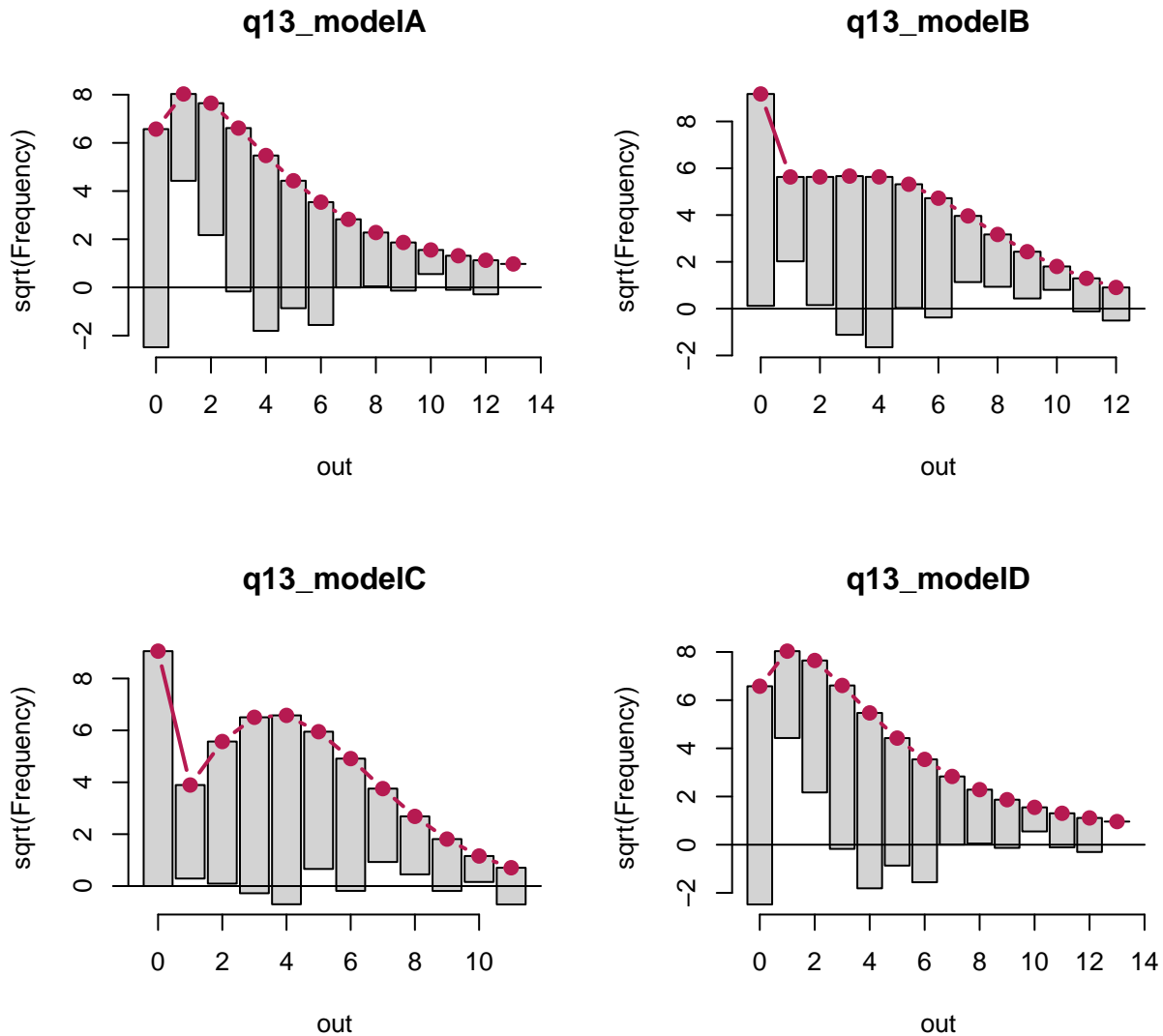
par(mfrow=c(2,2))

rootogram(q13_modelA)

rootogram(q13_modelB)

rootogram(q13_modelC)

rootogram(q13_modelD)
```



```
par(mfrow=c(1,1))
```

The Display for Question 13 shows four rootgrams, 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
- B
- C
- D
- It is impossible to tell from the information provided.

## 14 Question 14

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



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

## 15 Question 15

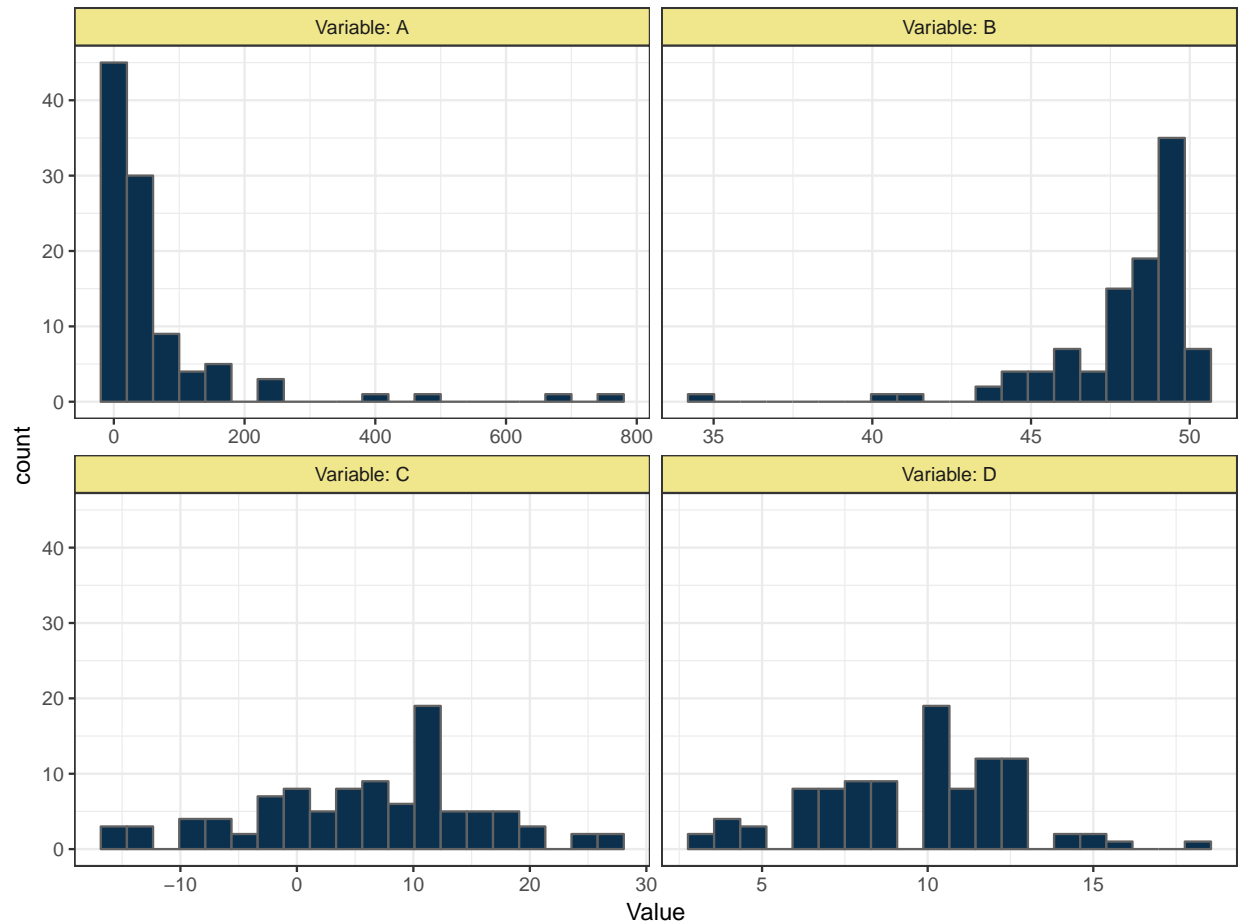
Suppose you want to build a plot to describe the relationship between a child's score on a measure of depression, and the child's favorite color. The depression scores emerge from a questionnaire, whose final score is standardized to have a mean of 50 and standard deviation of 10 across a prior large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

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

## 16 Question 16

### 16.1 Display for Question 16

```
ggplot(data16, aes(x = Value)) +  
  geom_histogram(bins = 20, fill = "#0a304e", col = "#626262") +  
  facet_wrap(~ Variable, scales = "free_x", labeller = "label_both") +  
  theme_bw() +  
  theme(strip.background =element_rect(fill="khaki"))
```



```
ggsave("displays/display16.png", height = 6, width = 8)
```

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

- A
- B
- C
- D
- It is impossible to tell from the information provided.

## 17 Question 17

### 17.1 Display for Question 17

```
d <- datadist(data17)
options(datadist = "d")

m17 <- lrm(y ~ x1 + rcs(x2,3) + x3, data = data17, x = TRUE, y = TRUE)

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

index.orig	training	test	optimism	index.corrected	n
------------	----------	------	----------	-----------------	---

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

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

- Area under the ROC curve will be about 0.28, Nagelkerke R-square about 0.08
- Area under the ROC curve will be about 0.31, Nagelkerke R-square about 0.11
- Area under the ROC curve will be about 0.36, Nagelkerke R-square about 0.14
- Area under the ROC curve will be about 0.39, Nagelkerke R-square about 0.17
- Area under the ROC curve will be about 0.64, Nagelkerke R-square about 0.08
- Area under the ROC curve will be about 0.66, Nagelkerke R-square about 0.11
- Area under the ROC curve will be about 0.68, Nagelkerke R-square about 0.14
- Area under the ROC curve will be about 0.69, Nagelkerke R-square about 0.17

## 18 Question 18



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

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

```

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

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

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

```

Figure 1:

## 19 Question 19

### 19.1 Display 1 for Question 19

```

summary(data19)
skim(data19)

```

## 19.2 Display 2 for Question 19

### Chunk I for Question 19

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

### Chunk II

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

### Chunk III

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

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

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

## 20 Question 20

### 20.1 Display 1 for Question 20

```
d <- datadist(data20)  
options(datadist = "d")  
m20 <- lrm(outcome ~ catg(x1) + x3 + rcs(x2, 3) +  
          x1 %ia% x2,  
          data = data20, x = TRUE, y = TRUE)  
  
m20
```

Logistic Regression Model

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

Model Likelihood  
Ratio Test

Discrimination  
Indexes

Rank Discrim.  
Indexes

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

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

## 20.2 Plot A for Question 20

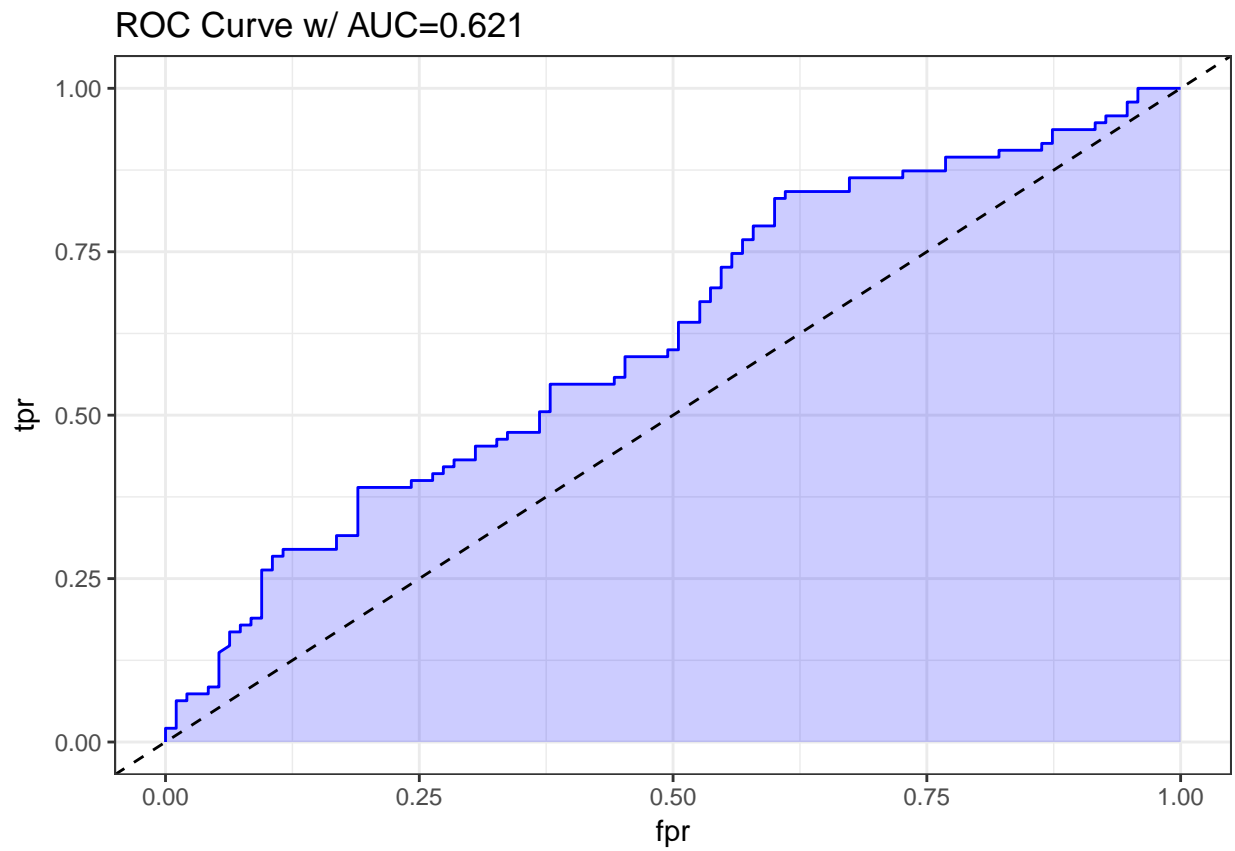
```
m20d <- glm(outcome ~ x3,
             data = data20, family = "binomial")

library(ROCR)

prob <- predict(m20d, data20, type = "response")
pred <- prediction(prob, data20$outcome)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
auc <- performance(pred, measure = "auc")

auc <- round(auc@y.values[[1]],3)
roc.data <- data.frame(fpr=unlist(perf@x.values),
                      tpr=unlist(perf@y.values),
                      model="GLM")

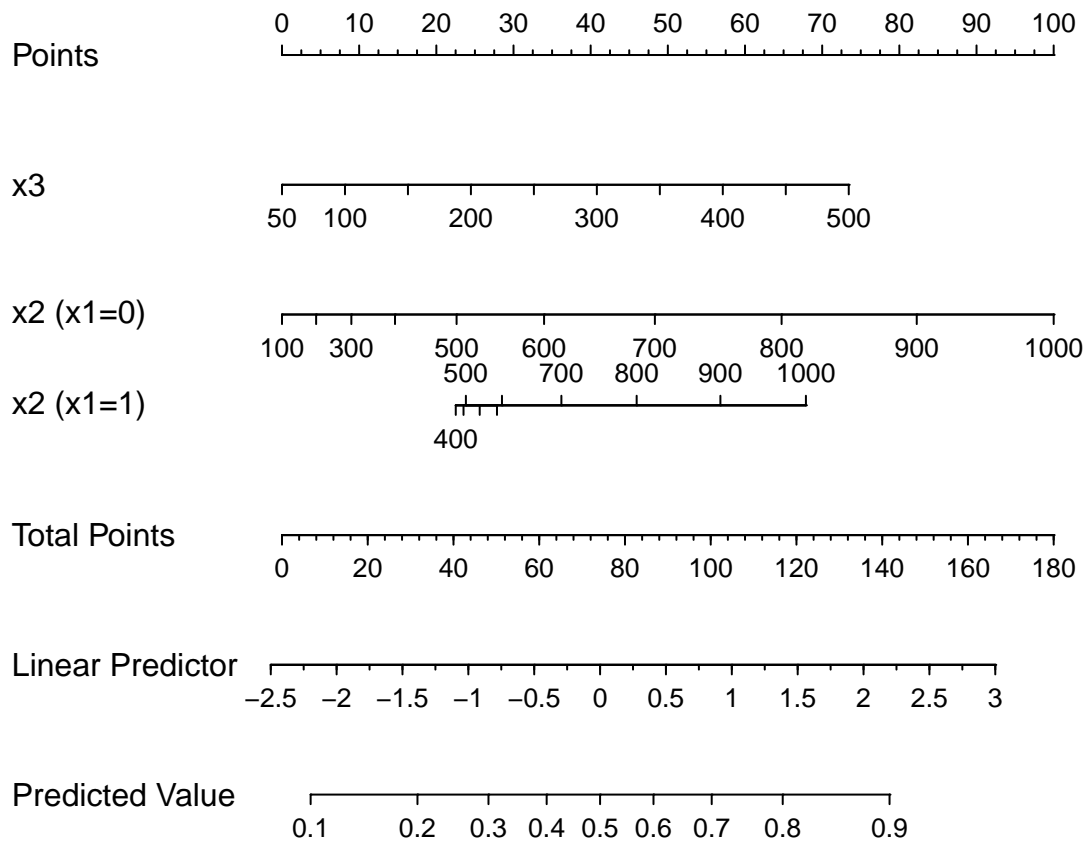
ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
  geom_ribbon(alpha=0.2, fill = "blue") +
  geom_line(aes(y=tpr), col = "blue") +
  geom_abline(intercept = 0, slope = 1, lty = "dashed") +
  labs(title = paste0("ROC Curve w/ AUC=", auc)) +
  theme_bw()
```



### 20.3 Plot B for Question 20

```
plot(nomogram(m20, fun = plogis))
```

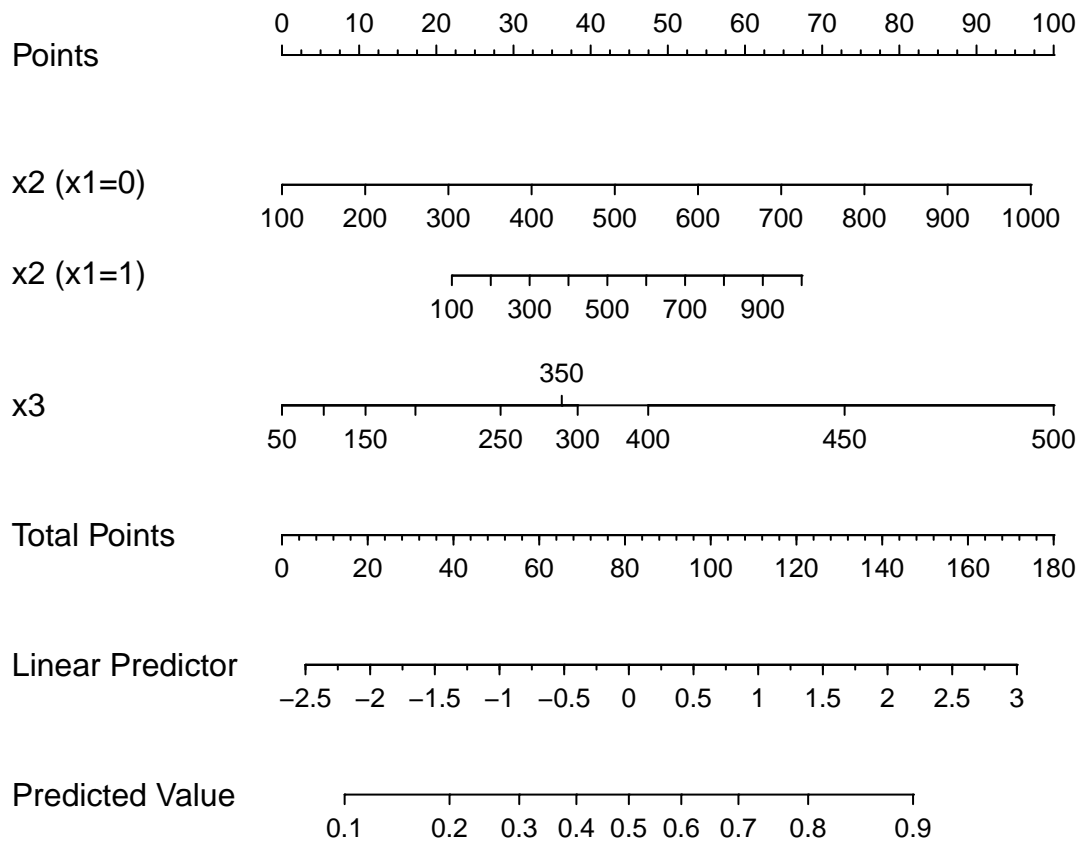




## 20.4 Plot C for Question 20

```
m20b <- lrm(outcome ~ x1 * x2 + rcs(x3,5),
             data = data20, x = TRUE, y = TRUE)

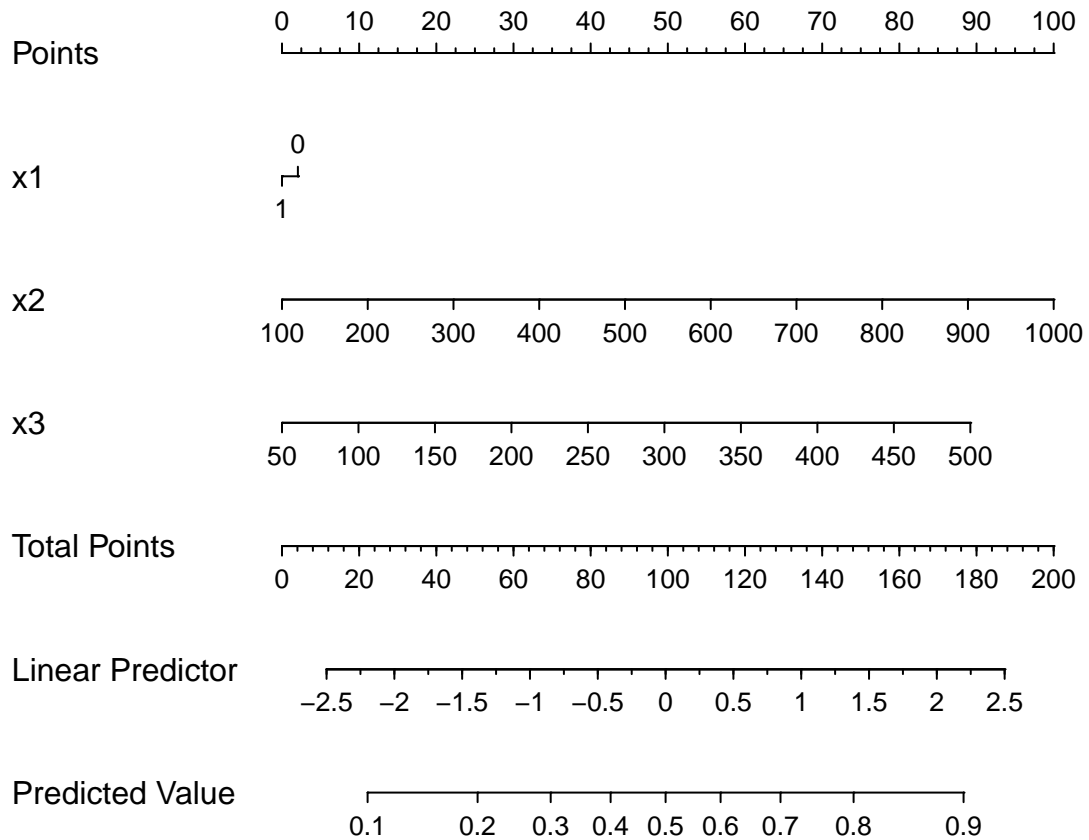
plot(nomogram(m20b, fun = plogis))
```



## 20.5 Plot D for Question 20

```
m20c <- lrm(outcome ~ x1 + x2 + x3,
             data = data20, x = TRUE, y = TRUE)

plot(nomogram(m20c, fun = plogis))
```



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

- Plot A
- Plot B
- Plot C
- Plot D
- It is impossible to tell from the information provided.

## 21 Question 21

### 21.1 Display for Question 21

- Statement I. A main effects model fit with Poisson regression provides a statistically significantly worse fit (at the 95% confidence level) than a model fit with Negative Binomial regression.

- Statement II. The rootogram for the Poisson model indicates a substantially better fit than the rootogram for the Negative Binomial model.
- Statement III. The rootogram for the Poisson model indicates a substantially worse fit than the rootogram for the Negative Binomial model.

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

- I only.
- II only.
- III only.
- I and II
- I and III
- II and III
- All three statements.
- None of these three statements.

## 22 Question 22

### 22.1 Display for Question 22

The three new subjects are Amy, Bart and Chris.

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

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

- Amy has the highest predicted `y`, then Bart then Chris
- Amy is highest, then Chris then Bart
- Bart is highest, then Amy then Chris
- Bart is highest, then Chris then Amy
- Chris is highest, then Amy then Bart
- Chris is highest, then Bart then Amy

## 23 Question 23

Now, instead of treating `y` in `data21` as a count variable, treat it as an ordinal category, and fit a new model that is appropriate for such an outcome using again the main effects of `x1`, `x2` and `x3` as predictors. Use that model to predict the actual category that our three new subjects (Amy, Bart and Chris) will fall into, and

compare that to the results you found in Question 22. How many of the three new subjects get a different predicted count with this ordinal categorical regression model, than they do when you round the predicted count made with the Poisson model to an integer?

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

## 24 Question 24

### 24.1 Display for Question 24

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

Morgan-Tatar search since family is non-gaussian.

```
res24_BIC$BestModels
```

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

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

Morgan-Tatar search since family is non-gaussian.

```
res24_CV
```

```
CVd(d = 82, REP = 1000)
```

```
BICq equivalent for q in (0.133590203067706, 0.468490321674287)
```

```
Best Model:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.62670316	1.26666101	-3.652677	0.0002595209
cov4	0.03065343	0.01131937	2.708051	0.0067679590

An “all subsets” approach as implemented in the `bestglm` package was used to fit a logistic regression model to describe the relationship between a binary outcome,  $y$ , and six predictors labeled  $x_1$  through  $x_6$ , using two different approaches to variable selection, as shown in the Display for Question 24. The first approach shown in the Display was an exhaustive search for the best possible BIC result. The second approach shown in the Display involved cross-validation. Which of the following statements is true?

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

### Setup for Questions 25-33

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

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

## 25 Question 25

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

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

## 26 Question 26

### 26.1 Display for Question 26

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

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

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

## 27 Question 27

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

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

## 28 Question 28

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

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

## 29 Question 29

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

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

## 30 Question 30

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

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

## 31 Question 31

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

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



## 32 Question 32

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

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

## 33 Question 33

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

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

### Setup for Questions 34-36

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

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

## 34 Question 34

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

### Questions 35 and 36 are BONUS questions

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

#### Setup for Questions 34-36

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

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

### 35 OPTIONAL BONUS Question 35

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

### 36 OPTIONAL BONUS Question 36

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

## 37 Answers

I'm going to try using the `praise` package in R to provide comments on some of these responses. Let's see how that goes...

```
library(praise)
```

### 37.1 Answer 1 is d

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

#### 37.1.1 Responses for Question 1

All 41 students got this right. YAY! That is just shining!

### 37.2 Answer 2 is d

- At the fraction 0.424 on the x-axis ( $|\beta|/\max(|\beta|)$  scale), four of the eight predictors are included in the model by the lasso, specifically predictors 1, 4, 6 and 7.

#### 37.2.1 Responses for Question 2

Selection	Count	%
d	23	56
c	16	39
Others	2	5

I assume the people who chose 3 misread the plot, perhaps forgetting that the plot works back from the right rather than forward from the left. Perhaps they didn't catch on that the marking of 3 on the top of the plot would identify the *maximum* fraction where 3 predictors would be used as being below 0.4, so that the 4-predictor solution would be needed with a fraction of 0.424.

### 37.3 Answer 3 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.

#### 37.3.1 Responses for Question 3

Selection	Count	%
c	26	63
a	6	15
d	6	15

Selection	Count	%
All Others	3	7

I was disappointed that people chose **a**, as that choice should have been ruled out quickly for not including an interaction term. **d**, as indicated above, is not a proper command in R - the `glm` model isn't built like that.

### 37.4 Answer 4 is c

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

#### 37.4.1 Responses for Question 4

All 41 students got this question right. You are first-class!

### 37.5 Answer 5 is d

- The plot shows a highly linear association, with a negative Pearson correlation that is very strong. In fact, the Pearson correlation here is  $r = -0.96$ , 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 0.93.

#### 37.5.1 Responses for Question 5

Selection	Count	%
<b>d</b>	<b>30</b>	<b>73</b>
<b>b</b>	6	15
All Others	5	12

Choosing **b** suggests a need to recalibrate your understanding of what a correlation means in terms of a scatterplot. The points follow nearly a straight line.

### 37.6 Answer 6 is a

- The red line (for drug A) 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 three groups shows highly statistically significant differences between the drug groups, and that drug has a substantially higher median and restricted mean time to recurrence.

#### 37.6.1 Responses for Question 6

All 41 students got this question right. AYE! You have done this really!

## 37.7 Answer 7 is technically b but I also accepted f

- A student pointed out my error. I'd intended Chunks II and III to each work, but I made a silly mistake in the Chunk III code. So I wound up giving full credit to the 34 students who had either II alone or II and III.
- The Chunk III code should have in the slice commands used row numbers, which can be accomplished by adding `which` to the code, as follows:

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

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

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

- Chunk I 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.

### 37.7.1 Responses for Question 7

Selection	Count	%
<b>b or f</b>	<b>34</b>	<b>83</b>
d	5	12
All Others	2	5

## 37.8 Answer 8 is e

- The Spearman plot indicates that non-linear terms built using (first) `sroh` (and second) `ld1` 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 `ld1` will use all four allowed additional degrees of freedom, so we stop there.

### 37.8.1 Responses for Question 8

Selection	Count	%
<b>e</b>	<b>31</b>	<b>76</b>
a	7	17
All Others	3	7

- Fitting an `ols` model including a spline in `ld1` is the other possibility that seemed popular. But `sroh` is well to the right of `ld1` in the Spearman plot, so I don't agree.

### 37.9 Answer 9 is b

- 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.

#### 37.9.1 Responses for Question 9

At least 90% of 41 students got this question right. HURRAH!

### 37.10 Answer 10 is b

- The stepwise model suggests x1, x2, x4 and x5.

```
step(lm(y ~ x1 + x2 + x3 + x4 + x5, data = data10))
```

Start: AIC=534.48

y ~ x1 + x2 + x3 + x4 + x5

	Df	Sum of Sq	RSS	AIC
- x3	1	20.80	9355.2	532.74
<none>			9334.4	534.48
- x1	1	162.97	9497.4	534.55
- x4	1	559.04	9893.5	539.46
- x2	1	570.76	9905.2	539.60
- x5	1	738.74	10073.2	541.62

Step: AIC=532.74

y ~ x1 + x2 + x4 + x5

	Df	Sum of Sq	RSS	AIC
<none>			9355.2	532.74
- x2	1	554.0	9909.2	537.65
- x4	1	573.3	9928.5	537.88
- x1	1	978.5	10333.7	542.68
- x5	1	11221.3	20576.6	625.33

Call:

```
lm(formula = y ~ x1 + x2 + x4 + x5, data = data10)
```

Coefficients:

(Intercept)	x1	x2	x4	x5
-8.6534	-0.1519	-0.7196	0.7014	0.3131

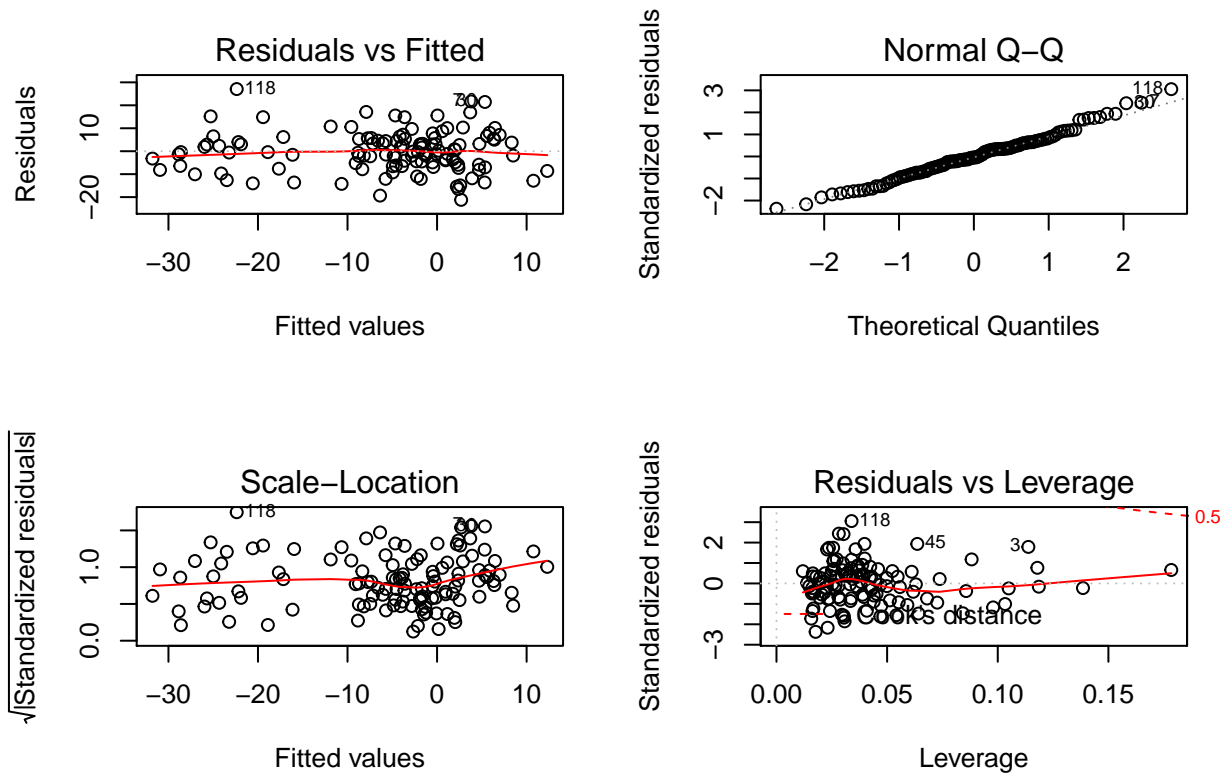
#### 37.10.1 Responses for Question 10

All 41 students got this question right. You are breathtaking!

### 37.11 Answer 11 is e

- There are no serious violations of regression assumptions, nor is there substantial collinearity in this model.

```
par(mfrow=c(2,2))
plot(lm(y ~ x1 + x2 + x4 + x5, data = data10))
```



```
par(mfrow=c(1,1))
vif(lm(y ~ x1 + x2 + x4 + x5, data = data10))
```

```
      x1      x2      x4      x5
1.215090 1.004946 1.016625 1.198333
```

### 37.11.1 Responses for Question 11

Selection	Count	%
e	13	32
d	23	56
All Others	5	12

I attribute this mainly to the fact that people are often reluctant to select “none of the above.” If you look at that scale-location plot and declare heteroscedasticity to be a substantial problem, then you need to recalibrate yourself a bit. I guess people are over-reacting to the little uptick in the smooth red curve between fitted values of 0 and 10 but the points show very little sign of that being an issue.

### 37.12 Answer 12 is c

- The answer is 9.3, which is between 9 and 10, and which is just a bit larger than the original estimated residual standard error of 9.02.

```
set.seed(432)
q12_models <- data10 %>%
  modelr::crossv_kfold(k = 10) %>%
  mutate(model = map(train, ~ lm(y ~ x1 + x2 + x4 + x5, data = .)))

q12_preds <- q12_models %>%
  unnest(map2(model, test, ~ augment(.x, newdata = .y)))

q12_preds %>%
  summarize(RMSE_model = sqrt(mean((y - .fitted) ^ 2)))

# A tibble: 1 x 1
  RMSE_model
    <dbl>
1      9.31
```

#### 37.12.1 Responses for Question 12

At least 90% of 41 students got this question right. HA!

### 37.13 Answer 13 is c

- c 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.

#### 37.13.1 Responses for Question 13

All 41 students got this question right. You are wonderful!

### 37.14 Answer 14 is d

- 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.

#### 37.14.1 Responses for Question 14

Selection	Count	%
d	29	71
b	10	24
All Others	2	5

I think that the 10 people who suggested this should be handled with a Cox model were assuming we were looking at *time* to rehospitalization, rather than just a binary measure of whether someone was re-hospitalized.



### 37.15 Answer 15 is a

We have a quantitative outcome (y) and a categorical predictor (x). Of the listed set of `geoms`, only a violin plot would be appropriate in this setting for comparing the four colors in terms of the distribution of depression in the kids.

#### 37.15.1 Responses for Question 15

All 41 students got this question right. You are lovely!

### 37.16 Answer 16 is a

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

#### 37.16.1 Responses for Question 16

Selection	Count	%
a	22	54
b	9	22
d	7	17
All Others	3	5

Neither **b**, which is left-skewed, nor **d** which isn't so far from Normal to begin with, look like good candidates for a log-Normal model.

### 37.17 Answer 17 is e

- We're looking for the index-corrected values from the `validate` output. The index-corrected Somers' d is 0.2848, so the index-corrected C statistic is  $0.5 + d/2 = 0.6424$ , and the index-corrected Nagelkerke R-square is 0.0778, so the correct answer is **e**.

#### 37.17.1 Responses for Question 17

Selection	Count	%
e	35	85
All Others	6	15

There was no obvious pattern in the incorrect responses. The most common error was to select **g**, which describes the original (prior to any validation) results, but that was just a few people.

### 37.18 Answer 18 is b

- `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.

### 37.18.1 Responses for Question 18

At least 90% of 41 students got this question right. WOWIE!

### 37.19 Answer 19 is c

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

#### 37.19.1 Responses for Question 19

Selection	Count	%
<b>c</b>	<b>27</b>	<b>66</b>
<b>b</b>	7	17
<b>f</b>	5	12
All Others	2	5

Choices **b** and **f** are each excited about Chunk II, but this doesn't work, because it defines the censored as observed and vice versa.

### 37.20 Answer 20 is b

- This is best done by process of elimination. Our model contains an interaction in  $x_1$  and  $x_2$ , and a non-linear term (spline) in  $x_2$ , with a C statistic of 0.741. Plot A cannot be right, since it shows the wrong ROC value (0.621). Plot C cannot be right because  $x_3$  includes a non-linear term. Plot D cannot be right because it doesn't show any interaction of  $x_1$  and  $x_2$ . So it must be that Plot B, which does show an interaction of  $x_1$  and  $x_2$ , and a linear effect of  $x_3$ , is the correct one. And, it is. I built Plot B from the model shown in Display 1, and the other Plots from other models.

#### 37.20.1 Responses for Question 20

Selection	Count	%
<b>b</b>	<b>33</b>	<b>80</b>
<b>c</b>	8	20

Everyone either picked **b** or **c**. The problem with **c** is that there is clearly a non-linear term in  $x_3$  shown in the nomogram (note, for instance, the distance between 50 and 150 as compared to 150 and 250, or, if you prefer, note the turnaround from 300 to 350 to 400). But the model doesn't include such a term.

### 37.21 Answer 21 is h

- None of the statements are true. The Poisson and Negative Binomial regression models are nearly identical, and show no significant difference in the likelihood ratio test, and no meaningful difference

between their rootograms.

```
mod21_p <- glm(y ~ x1 + x2 + x3, family = poisson(), data = data21)
mod21_nb <- MASS::glm.nb(y ~ x1 + x2 + x3, link = log, data = data21)
```

```
logLik(mod21_p)
```

```
'log Lik.' -285.7795 (df=4)
```

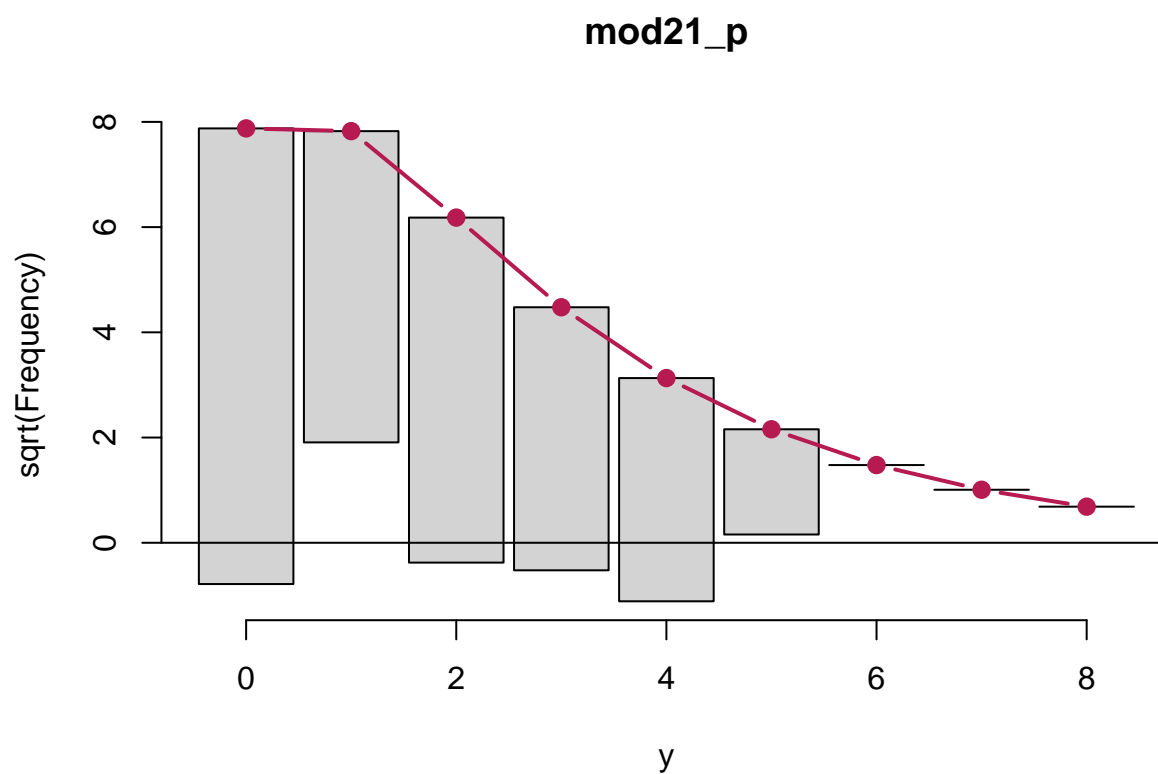
```
logLik(mod21_nb)
```

```
'log Lik.' -285.7802 (df=5)
```

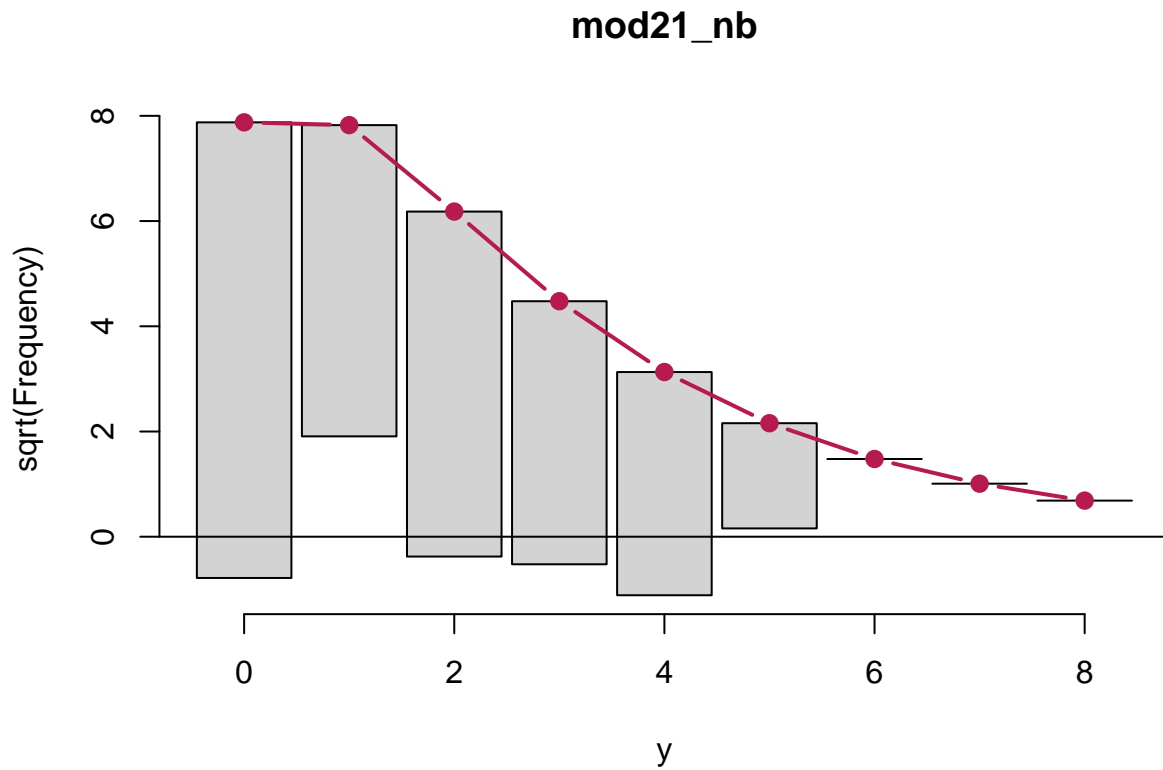
```
pchisq(2 * (logLik(mod21_nb) - logLik(mod21_p)), df = 1, lower.tail = FALSE)
```

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

```
rootogram(mod21_p)
```



```
rootogram(mod21_nb)
```



### 37.21.1 Responses for Question 21

Selection	Count	%
<b>h</b>	<b>35</b>	<b>85</b>
a and b	3 each	7 each

Three people chose Statement I and three people chose Statement II. Neither is true.

### 37.22 Answer 22 is e

- Chris is highest, then Amy, then Bart.

```
nd22 = data_frame(name = c("Amy", "Bart", "Chris"),
                  x1 = c(3, 2, 4),
                  x2 = c(1, 0, 1),
                  x3 = c(4, 0, 6))

predict(mod21_p, newdata = nd22, type = "response")
```

```
      1      2      3
1.8382583 0.3177623 3.1760489
```

### 37.22.1 Responses for Question 22

At least 90% of 41 students got this question right. YOW!

### 37.23 Answer 23 is d

- 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.

```
mod23_polr <- polr(factor(y) ~ x1 + x2 + x3, data = data21, Hess = TRUE)

predict(mod23_polr, nd22)
```

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

### 37.23.1 Responses for Question 23

Selection	Count	%
d	24	59
a	8	20
e	5	12
All Others	4	10

I assume people who got this wrong made a mistake somewhere in their calculations, but without seeing them, it's hard to identify where.

### 37.24 Answer 24 is b

- From the Display, the model fit by cross-validation includes only `cov4`, which is the model with the second lowest BIC across all subsets of predictors, as identified by the output on the BIC in the Display.

### 37.24.1 Responses for Question 24

At least 90% of 41 students got this question right. HURRAY!

### 37.25 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.

### 37.25.1 Responses for Question 25

All 41 students got this question right. You are breathtaking!

### 37.26 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.

#### 37.26.1 Responses for Question 26

Selection	Count	%
<b>a</b>	<b>36</b>	<b>88</b>
All Others	5	12

I didn't observe any clear pattern to the incorrect responses, which included **d**, **e** and **f**.

### 37.27 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.

#### 37.27.1 Responses for Question 27

At least 90% of 41 students got this question right. YAHOO!

### 37.28 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.

#### 37.28.1 Responses for Question 28

At least 90% of 41 students got this question right. HO-HO!

### 37.29 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.

#### 37.29.1 Responses for Question 29

At least 90% of 41 students got this question right. WOW!

### 37.30 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**.

#### 37.30.1 Responses for Question 30

At least 90% of 41 students got this question right. WHEE!

### 37.31 Answer 31 is c

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

#### 37.31.1 Responses for Question 31

Selection	Count	%
<b>c</b>	<b>34</b>	<b>83</b>
All Others	7	17

The other responses were about evenly split between d and e, which might indicate that people misread the  $C_p$  plot, but could be a sign of something else.

### 37.32 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.

#### 37.32.1 Responses for Question 32

Selection	Count	%
<b>c</b>	<b>33</b>	<b>80</b>
<b>b</b>	6	15
All Others	2	5

Perhaps some people only looked at Item 9.

### 37.33 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.

### 37.33.1 Responses for Question 33

Selection	Count	%
b	35	80
All Others	6	15

Most of the incorrect folks chose a.

### 37.34 Answer 34 is 100 rows.

```
data34 %>% filter(complete.cases(.)) %>% dim(.)
```

```
[1] 900 6
```

We see that we have 900 rows with complete data. Since we started with 1000 rows, we have exactly 100 with missing data.

### 37.34.1 Responses for Question 34

At least 90% of 41 students got this question right. YOW!

### 37.35 Answer 35 is a long presentation of R code

Here's what I used.

```
set.seed(43237)

d <- datadist(data34)
options(datadist = "d")

imp_fit35 <- aregImpute(~ alive + treated + age + female + comor,
  nk = c(0,3), tlinear = TRUE, data = data34,
  B = 10, n.impute = 20, pr = FALSE)

m35_imp <- fit.mult.impute(alive ~ treated + age + female + comor,
  fitter = lrm, xtrans = imp_fit35,
  data = data34, x = T, y = T)
```

The result of applying this is:

```
set.seed(43237)

d <- datadist(data34)
options(datadist = "d")

imp_fit35 <- aregImpute(~ alive + treated + age + female + comor,
  nk = c(0,3), tlinear = TRUE, data = data34,
  B = 10, n.impute = 20, pr = FALSE)

m35_imp <- fit.mult.impute(alive ~ treated + age + female + comor,
```



```
fitter = lrm, xtrans = imp_fit35,
data = data34, x = T, y = T)
```

Variance Inflation Factors Due to Imputation:

Intercept	treated	age	female	comor
1.26	1.42	1.97	1.02	4.24

Rate of Missing Information:

Intercept	treated	age	female	comor
0.20	0.29	0.49	0.02	0.76

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

Intercept	treated	age	female	comor
455.51	220.61	77.97	55262.87	32.52

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

```
stats linear.predictors
m35_imp
```

Logistic Regression Model

```
fit.mult.impute(formula = alive ~ treated + age + female + comor,
  fitter = lrm, xtrans = imp_fit35, data = data34, x = T, y = T)
```

		Model Likelihood		Discrimination		Rank Discrim.	
		Ratio Test		Indexes		Indexes	
Obs	1000	LR chi2	459.89	R2	0.527	C	0.888
0	712	d.f.	4	g	2.529	Dxy	0.775
1	288	Pr(> chi2)	<0.0001	gr	12.817	gamma	0.776
max  deriv	1e-08			gp	0.320	tau-a	0.318
				Brier	0.112		

	Coef	S.E.	Wald Z	Pr(> Z )
Intercept	6.8501	0.6950	9.86	<0.0001
treated	1.5673	0.2343	6.69	<0.0001
age	-0.2116	0.0209	-10.13	<0.0001
female	-0.0302	0.1864	-0.16	0.8712
comor	0.5449	0.1254	4.35	<0.0001

### 37.35.1 Grading on Question 35

31/41 students received credit (5 points) on Question 35. Some of those who didn't get credit didn't attempt the question. Those who made an attempt but didn't get credit were those who:

- failed to do multiple imputation, or
- failed to include alive, treated, age, female and comor in the imputation model, or
- failed to fit the outcome model using `fit.mult.impute` (instead perhaps attempting to do so with `glm`)

### 37.36 Answer 36 is 4.8, with 95% CI (3.0, 7.6) for the odds ratio.

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.

```
summary(m35_imp)
```

Effects				Response : alive			
Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
treated	0	1	1	1.567300	0.23428	1.108100	2.026500
Odds Ratio	0	1	1	4.793800	NA	3.028700	7.587500
age	43	58	15	-3.174100	0.31330	-3.788200	-2.560100
Odds Ratio	43	58	15	0.041830	NA	0.022636	0.077299
female	0	1	1	-0.030229	0.18644	-0.395650	0.335190
Odds Ratio	0	1	1	0.970220	NA	0.673240	1.398200
comor	1	4	3	1.634600	0.37615	0.897330	2.371800
Odds Ratio	1	4	3	5.127300	NA	2.453000	10.717000

#### 37.36.1 Grading on Question 36

26/41 students received credit (5 points) on Question 36. Only those who got credit on Question 35 were eligible here, so it's really 26/31. Most of those who didn't get credit here were those who misinterpreted the odds ratio estimate they obtained or who didn't obtain the odds ratio properly, or at all. A common issue was people obtaining an estimate of 4.5 (roughly) rather than 4.8. This happened either because you used R 3.5.0 instead of R 3.4.4, or because you specified the variables as factors in the imputation or outcome model in a different way than I did. I gave full credit to those getting a value near 4.5 or 4.8, so long as they did the rest properly.

## 38 Grades on Quiz 2

While the total available points added up to 112, the highest score on the Quiz was 109. Special congratulations to the three students meeting this mark.

Students with Grades ...	Count
of 100 to 109	13
of 90 to 99	15
of 75 to 89	9
below 75	4

The median grade was 94.