432 Quiz 1 with Answer Sketch and Rubric

Thomas E. Love

Due 2020-03-02 at 2 PM. Version: 2020-03-03 11:59:05

Instructions

Please select or type in your best response (or responses, as indicated) for each question.

The deadline for completing the Quiz is 2 PM on Monday 2020-03-02, and this is a firm deadline, without the grace period we allow for in Homework assignments.

Each question's point value is specified. Total available points on the Quiz = 75.

There are 14 questions in all. The questions are not arranged in any particular order. Your score is based on your correct responses, so there's no chance a blank response will be correct, and a guess might be, so you should definitely answer all of the questions.

If you wish to work on some of the quiz and then return later, you can do this by [1] completing the final question (the affirmation) which asks you to type in your full name, and then [2] submitting the quiz. You will then receive a link at your CWRU email which will allow you to return to the quiz as often as you like without losing your progress.

There are two data files (`lind.Rds` and `riff.csv`) provided as part of the Quiz 1 materials on our Shared Google Drive. You will need these files to complete the Quiz.

A PDF version of Quiz 1 is also available to you our Shared Google Drive, but all of your answers must be returned using the Answer Sheet Google Form, which is located at http://bit.ly/432-2020-quiz1-answer-form.

You are welcome to consult the materials provided on the course website, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants and you must do so by sending an email to 431-help at case dot edu. That way, we can track what's been asked. We will not guarantee that questions received after 9 AM on Monday 2020-03-02 will be answered in a timely fashion, but otherwise, we will try to keep up.

Thank you, and good luck.

Setup

Here are the packages you can assume have been loaded into R to help us do this work. We do not guarantee that all of these packages are actually necessary to complete the Quiz successfully.

```
library(here)
library(knitr)
library(magrittr)
library(janitor)
library(patchwork)
library(tableone)
library(rms)
library(leaps)
library(caret)
library(simputation)
library(naniar)
library(broom)
library(tidyverse)

theme_set(theme_bw())
```

Question 1 (6 points)

The table below describes the result of using 10-fold cross-validation to compare seven candidate linear regression models (labeled model1, model2, model3, model4, model5, model6, and model7) for a data set predicting a quantitative outcome. The table summarizes cross-validation R-square (labeled `Rsquared`), the root mean squared prediction error (labeled `RMSE`), and the mean absolute prediction error (labeled `MAE`).

```
q01_display <- bind_rows(
    res1$results,
    res2$results,
    res3$results,
    res4$results,
    res5$results,
    res6$results,
    res7$results) %>%
    mutate(model = c("model1", "model2", "model3", "model4",
                     "model5", "model6", "model7")) %>%
    select(model, Rsquared, RMSE, MAE)

q01_display %>% kable(digits = 4)
```

model	Rsquared	RMSE	MAE
model1	0.5938	5.5902	4.4872
model2	0.5943	5.5508	4.4634
model3	0.5954	5.5676	4.4529
model4	0.5923	5.5534	4.4494
model5	0.5538	5.8391	4.7508
model6	0.5929	5.5711	4.4611
model7	0.5487	5.8320	4.7363

According to the table, which model shows the strongest results in terms of:

Rows:

a. cross-validated R-square
b. root mean squared prediction error
c. mean absolute prediction error

Columns:

1. model1
2. model2
3. model3
4. model4
5. model5
6. model6
7. model7

Questions 2-7 use the `lind` data

The `lind` data provided to you in the **lind.Rds** file describe 970 of the subjects in an observational study of adults receiving an initial Percutaneous Coronary Intervention (PCI) at Ohio Heart Health, Christ Hospital, Cincinnati in 1997 and followed for at least 6 months by the staff of the Lindner Center. The patients thought to be more severely diseased were assigned to treatment with abciximab (an expensive, high-molecular-weight IIb/IIIa cascade blocker); while the rest of the patients received usual-care-alone with their initial PCI. The data elements we're using in Quiz 1 are:

Variable	Description
`ptid`	subject ID (assigned by Dr. Love for this Quiz)
`cardbill`	Cardiac related costs incurred within 6 months of patient's initial PCI; numeric value in 1998 dollars
`abcix`	Treatment indicator: 0 means usual PCI care alone; 1 means usual PCI care augmented by treatment with abciximab.
`stent`	Coronary stent deployment, with 1 meaning YES and 0 meaning NO.
`acutemi`	Acute myocardial infarction in the previous 7 days, with 1 meaning YES and 0 meaning NO.
`ejecfrac`	Left ventricular ejection fraction; numeric value from 0 percent to 90 percent.
`ves1proc`	Number of vessels involved in the patient's initial PCI procedure; integer from 0 to 5.
`diabetic`	Diabetes mellitus diagnosis, with 1 meaning YES and 0 meaning NO.

```
lind <- readRDS(here("data/lind.Rds"))

summary(lind)
```

```
     ptid              cardbill          abcix            stent       
 Length:970         Min.   :  2216   Min.   :0.0000   Min.   :0.0000  
 Class :character   1st Qu.: 10172   1st Qu.:0.0000   1st Qu.:0.0000  
 Mode  :character   Median : 12395   Median :1.0000   Median :1.0000  
                    Mean   : 15496   Mean   :0.7082   Mean   :0.6691  
                    3rd Qu.: 16597   3rd Qu.:1.0000   3rd Qu.:1.0000  
                    Max.   :178534   Max.   :1.0000   Max.   :1.0000  
    acutemi          ejecfrac        ves1proc        diabetic      
 Min.   :0.0000   Min.   : 0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:45.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :0.0000   Median :55.00   Median :1.000   Median :0.0000  
 Mean   :0.1412   Mean   :51.18   Mean   :1.385   Mean   :0.2186  
 3rd Qu.:0.0000   3rd Qu.:56.00   3rd Qu.:2.000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :90.00   Max.   :5.000   Max.   :1.0000  
```

Question 2 (10 points)

Provide R code to produce the result shown below, which I obtained using the `lind` data. You can assume that all commands shown in this document prior to these words have already been run, including loading all of the packages listed above and loading the `lind` data set as indicated above.

Be sure that your table shows exactly the same result as this one. Note that this will require you to create a new variable called `treatment` that contains the information in `abcix` in a revised form.

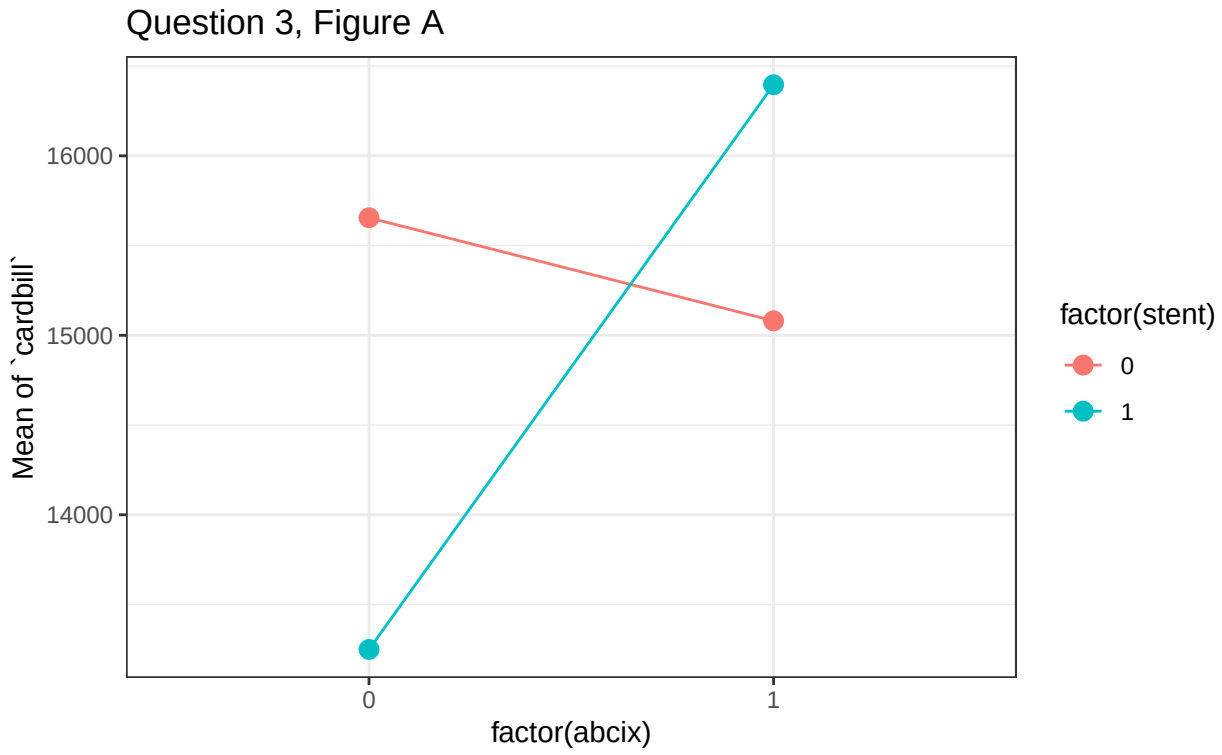
```
                     Stratified by treatment
                           abcix                          usual_care                    p   test
n                            687                            283
stent = 1 (%)                484 (70.5)                     165 (58.3)               <0.001
acutemi = 1 (%)              121 (17.6)                      16 ( 5.7)               <0.001 exact
diabetic = 1 (%)             139 (20.2)                      73 (25.8)                0.060 exact
ejecfrac (mean (SD))       50.46 (10.38)                  52.93 (9.62)                0.001
ves1proc (mean (SD))        1.46 (0.70)                    1.20 (0.47)               <0.001
cardbill (median [IQR]) 12901.00 [10882.50, 17067.50] 10169.00 [8282.50, 15684.00] <0.001 nonnorm
```

Question 3 (4 points)

Consider the Figure A I built using this code. . .

```
summ_q3 <- lind %>% group_by(abcix, stent) %>%
    summarize(n = n(), mean = mean(cardbill))

ggplot(summ_q3, aes(x = factor(abcix), y = mean,
                    col = factor(stent))) +
    geom_point(size = 3) +
    geom_line(aes(group = factor(stent))) +
    labs(y = "Mean of `cardbill`",
         title = "Question 3, Figure A")
```



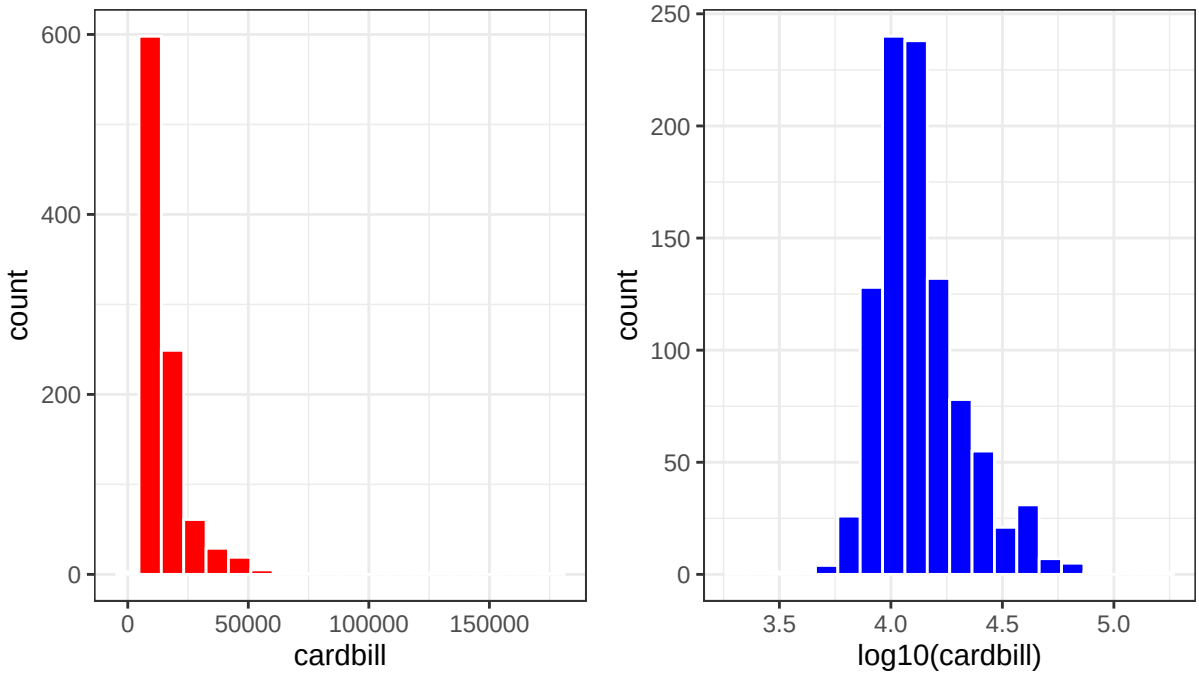
Now, it turns out that the `cardbill` data are highly right skewed. See Question 3, Figure B on the next page, which was developed using the code below.

```
p1 <- ggplot(lind, aes(x = cardbill)) +
    geom_histogram(bins = 20, fill = "red", col = "white")

p2 <- ggplot(lind, aes(x = log10(cardbill))) +
    geom_histogram(bins = 20, fill = "blue", col = "white")

p1 + p2 +
    plot_annotation(title = "Question 3, Figure B")
```

Question 3, Figure B



Develop an appropriate interaction plot describing the effect of `abcix` and `stent` on the base-10 logarithm of `cardbill`. Please note the use of the base-10 logarithm here, rather than our more typical use of the natural logarithm (base e). So if we needed to back out of that transformation, to move from `log10(cardbill)` back to `cardbill`, we'd have to use `10^(log10(cardbill))`, rather than the `exp` function.

For Question 3, we want to have a standard for how we'll define an interaction that is large enough for us to call it "substantial". So, if the difference in the mean of the base-10 logarithm of cardiac related costs for those with a stent as compared to those without a stent is as least twice as large in one `abcix` category than it is in the other `abcix` category, we'll call that a substantial interaction.

Which of the following conclusions is most appropriate, based on the plot you developed, and the definition of "substantial" interaction specified above?

a. There is no substantial interaction between treatment received (abcix vs. usual care) and stent placement (yes or no) on the mean of log(cardiac related costs). Subjects with stents have higher means, regardless of whether they received abcix.
b. There is no substantial interaction between treatment received (abcix vs. usual care) and stent placement (yes or no) on the mean of log(cardiac related costs). Subjects with stents have lower means, regardless of whether they received abcix.
c. There is a substantial interaction, and among subjects receiving abcix, those with stents exhibit higher means, but among those subjects not receiving abcix, those with stents exhibit lower means.
d. There is a substantial interaction, and those with stents exhibit higher means, regardless of their abcix status.
e. There is a substantial interaction, and among subjects receiving abcix, those with stents exhibit lower means, but among those subjects not receiving abcix, those with stents exhibit higher means.
f. There is a substantial interaction, and those with stents exhibit lower means, regardless of their abcix status.
g. None of these conclusions are appropriate.

Question 4 (4 points)

Fit an appropriate ANOVA model (taking into account what you learned in Question 3 regarding whether the interaction term is substantial) to describe the impact of `abcix` and `stent` on the base-10 logarithm of `cardbill`. Specify the η^2 value for the combined impact of **all** predictors in your model.

Note: The answer you type into the Google Form Answer Sheet should just be a number, expressed as a percentage, and rounded to a single decimal place. Do not type the symbol %.

Question 5 (6 points)

Consider model `mod5` as tabulated below.

```
mod5 <- lind %$% lm(log10(cardbill) ~ acutemi + ejecfrac)

tidy(mod5, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error,
           conf.low, conf.high, p.value) %>%
    kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	4.264	0.034	4.208	4.319	0.000
acutemi	-0.042	0.019	-0.072	-0.011	0.024
ejecfrac	-0.002	0.001	-0.003	-0.001	0.000

Please provide an appropriate interpretation for the meaning of the coefficient estimate for `acutemi` provided above. Use complete English sentences. Be sure to interpret both the point estimate and confidence interval provided, and explain what they mean in context.

Question 6 (5 points)

Consider the model `mod6` summarized below.

```
mod6 <- lind %$% lm(log10(cardbill) ~
                        ejecfrac + abcix * acutemi)

summary(mod6)
```

```
Call:
lm(formula = log10(cardbill) ~ ejecfrac + abcix * acutemi)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72970 -0.12950 -0.04775  0.08744  1.14604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.1867803  0.0351328 119.170  < 2e-16 ***
ejecfrac      -0.0020274  0.0006241  -3.248   0.0012 **
abcix          0.0838100  0.0144988   5.780 1.01e-08 ***
acutemi       -0.1195489  0.0501452  -2.384   0.0173 *
abcix:acutemi  0.0707196  0.0538871   1.312   0.1897
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1948 on 965 degrees of freedom
Multiple R-squared:  0.05854,   Adjusted R-squared:  0.05464
F-statistic:    15 on 4 and 965 DF,  p-value: 6.686e-12
```

Suppose we have two patients, named Abby and Bob.

- Abby has an ejection fraction of 50, was treated with `abcix` in addition to usual PCI care, and had an acute MI within the previous 7 days.
- Bob has an ejection fraction of 60, and was also treated with `abcix` in addition to usual PCI care, but had not had an acute MI within the previous 7 days.

Question 6 has two parts. (6a is worth 2 points, and 6b is worth 3 points.)

6a. Please specify which patient (Abby or Bob) has the higher predicted cardiac related costs (according to model 6) incurred within 6 months of their initial PCI.

6b. Tell me the difference between the two model 6 predictions, in 1998 dollars, rounded to the nearest 100 dollars. (For example, if you found that the difference was $4,230, you would report this as $4,200 after rounding to the nearest 100 dollars.) Do not type in the dollar sign in the Google Form. Just the number, please.

Note: I'll remind you that the logarithm used here is a base-10 logarithm, and not a natural (base e) logarithm.

NOTE: This was corrected on 2020-02-27. All of question 6 should refer to model 6.

Question 7 (5 points)

I have used the `lind.Rds` file provided to predict `stent` using `ves1proc` and `abcix` in two models, one called `mod7a` and one called `mod7b`. as shown below.

```
mod7a <- lind %$% lm(stent ~ abcix + ves1proc)

tidy(mod7a, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error,
           conf.low, conf.high, p.value) %>%
    kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	0.630	0.039	0.565	0.695	0.000
abcix	0.132	0.034	0.076	0.187	0.000
ves1proc	-0.039	0.023	-0.078	-0.001	0.091

```
mod7b <- lind %$% glm(stent ~ abcix + ves1proc,
                      family = binomial())

tidy(mod7b, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error,
           conf.low, conf.high, p.value) %>%
    kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	1.735	0.175	1.301	2.315	0.002
abcix	1.791	0.150	1.399	2.292	0.000
ves1proc	0.836	0.105	0.704	0.995	0.088

For each statement below, identify whether it is true about `mod7a`, `mod7b`, both models, or neither model.

Columns:

1. mod7a
2. mod7b
3. both
4. neither

Rows:

a. The model predicts the probability that a subject will receive a stent.
b. If subjects A and B have the same `abcix` status, but A has one more `ves1proc` than B, A is predicted to have a larger probability of receiving a stent.
c. This is a linear probability model.
d. If subjects A and B have the same `abcix` status, but A has one more `ves1proc` than B, A is predicted to have a smaller probability of receiving a stent.
e. This is a logistic regression model.

Question 8 (4 points)

Consider model `mod8`, as specified below.

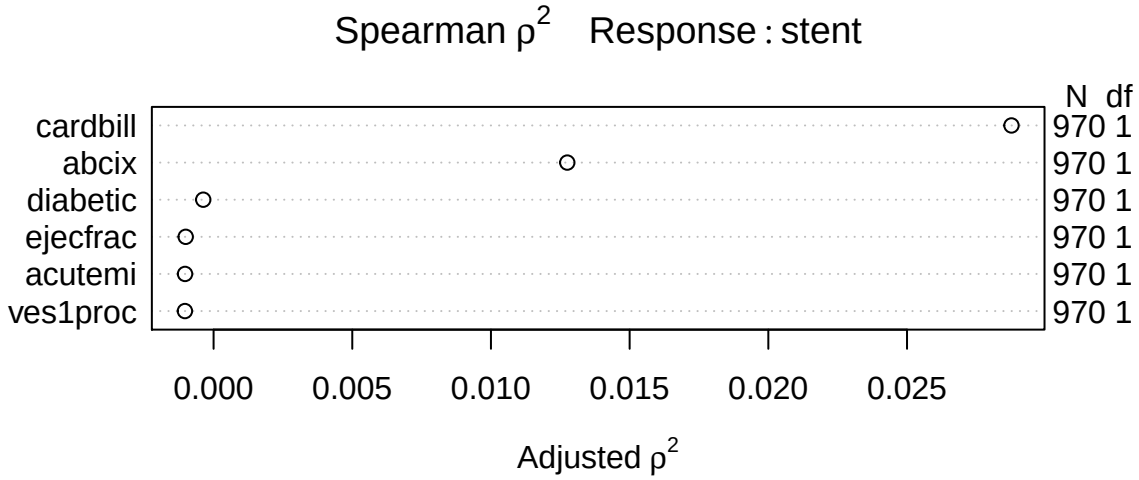
```
mod8 <- lind %$% lrm(stent ~ cardbill + abcix + acutemi +
                     ejecfrac + ves1proc + diabetic)
anova(mod8)
```

```
                Wald Statistics          Response: stent

 Factor     Chi-Square d.f. P
 cardbill     0.07       1    0.7969
 abcix       14.80       1    0.0001
 acutemi      0.24       1    0.6267
 ejecfrac     0.03       1    0.8703
 ves1proc     2.89       1    0.0889
 diabetic     0.24       1    0.6244
 TOTAL       16.58       6    0.0110
```

As you can see, Model `mod8` uses exactly 6 degrees of freedom beyond the intercept term. Now, consider the Spearman ρ^2 plot below.

```
plot(spearman2(stent ~ cardbill + abcix + acutemi + ejecfrac + ves1proc + diabetic, data = lind))
```



Suppose we plan to add some non-linear terms to `mod8` so as to now use exactly 10 degrees of freedom beyond the intercept. According to the Spearman ρ^2 plot, which of the following strategies is most appropriate?

a. Add a polynomial term of degree 4 for cardbill, as well as the interaction of abcix and ejecfrac.
b. Add a restricted cubic spline term for the cardbill predictor, using five knots.
c. Add restricted cubic spline terms with 4 knots each for ves1proc and ejecfrac.
d. Add a cubic polynomial for abcix, and a cubic polynomial for cardbill.
e. Add an interaction term for abcix and cardbill, plus a restricted cubic spline in cardbill with 5 knots.
f. Add interaction terms for abcix and diabetic, for diabetic and cardbill, and for abcix and cardbill.

Question 9 (6 points)

I created the output for this question using a data set which I am not providing to you, called `q9`. These data contain 450 observations on 7 potential predictors, labeled a, b, c, d, e, f and g, of a quantitative outcome (called `outcome`) that has been rounded to 1 decimal place.

- Variable `a` takes on integer values between 1 and 10
- Variable `b` can take any integer value between 0 and 1000
- Variables `c`, `d` and `e` are quantities (rounded to 2 decimal places, each)
- Variables `f` and `g` are binary categorical variables

Here's a quick look at the data.

```
summary(q9)
```

```
       a                b               c                d         
 Min.   : 1.000   Min.   :  4.0   Min.   : 6.470   Min.   : 86.18  
 1st Qu.: 3.000   1st Qu.:496.2   1st Qu.: 9.685   1st Qu.: 98.20  
 Median : 5.000   Median :595.5   Median :10.525   Median :102.06  
 Mean   : 5.398   Mean   :586.8   Mean   :10.443   Mean   :102.13  
 3rd Qu.: 8.000   3rd Qu.:690.8   3rd Qu.:11.217   3rd Qu.:105.90  
 Max.   :10.000   Max.   :927.0   Max.   :13.530   Max.   :117.67  
       e                f                g             outcome      
 Min.   : 35.53   Min.   :0.0000   Min.   :0.0000   Min.   :39.20  
 1st Qu.: 99.22   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:54.52  
 Median :109.98   Median :0.0000   Median :1.0000   Median :60.60  
 Mean   :108.73   Mean   :0.4733   Mean   :0.5156   Mean   :60.91  
 3rd Qu.:119.91   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:67.47  
 Max.   :155.95   Max.   :1.0000   Max.   :1.0000   Max.   :87.40  
```

Consider the output provided below.

```
rs9 <- regsubsets(outcome ~ a + b + c + d + e + f + g,
                  data = q9, nvmax = 7, nbest = 1)

rs9_winners <-
    tbl_df(summary(rs9)$which) %>%
    mutate(preds = 1:(rs9$nvmax - 1),
           r2 = summary(rs9)$rsq,
           adjr2 = summary(rs9)$adjr2,
           cp = summary(rs9)$cp,
           bic = summary(rs9)$bic) %>%
    select(preds, r2, adjr2, cp, bic, a, b, c, d, e, f, g, "int" = "(Intercept)")
```

```
rs9_winners %>%
    kable(digits = c(0, 5, 5, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0))
```

preds	r2	adjr2	cp	bic	a	b	c	d	e	f	g	int
1	0.53917	0.53814	54.27	-336.4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
2	0.58445	0.58259	7.11	-376.8	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
3	0.58841	0.58564	4.82	-375.0	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE
4	0.59078	0.58710	4.24	-371.5	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
5	0.59191	0.58732	5.01	-366.7	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
6	0.59281	0.58729	6.04	-361.5	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
7	0.59285	0.58640	8.00	-355.5	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Question 9 (continued)

Which predictors are included in the models recommended by a best subsets approach applied to the data?

Rows:

1. by adjusted R^2
2. by Mallows' Cp
3. by BIC

Columns:

a. a
b. b
c. c
d. d
e. e
f. f
g. g

Questions 10-12 use the `riff` data set

286 male patients were examined. Each exhibited one of several reasons to suspect problems with their prostate glands. These data are available in the `riff.csv` data set. For each patient, the following data are provided:

- `ptnum` = patient identification code
- `age` = age (in years)
- `dre` = digital rectal examination result (0 = negative, 1 = positive)
- `tru` = transurethral ultrasound result (0 = negative, 1 = positive)
- `psa` = prostate-specific antigen level (in ng/ml)
- `vol` = volume of prostate (in ml)
- `psad` = prostate-specific antigen density level (this is just psa / vol)
- `biopsy` = biopsy result (0 = negative, 1 = positive)

```
summary(riff)
```

```
     ptnum             age             dre              tru        
 Min.   :  1.00   Min.   :47.00   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 72.25   1st Qu.:61.00   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :143.50   Median :66.00   Median :1.0000   Median :0.0000  
 Mean   :143.50   Mean   :66.72   Mean   :0.6084   Mean   :0.4755  
 3rd Qu.:214.75   3rd Qu.:72.75   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :286.00   Max.   :91.00   Max.   :1.0000   Max.   :1.0000  
      psa               vol              psad            biopsy      
 Min.   :  0.300   Min.   :  3.30   Min.   :0.0100   Min.   :0.0000  
 1st Qu.:  3.125   1st Qu.: 24.59   1st Qu.:0.0800   1st Qu.:0.0000  
 Median :  5.850   Median : 32.80   Median :0.1600   Median :0.0000  
 Mean   :  8.928   Mean   : 36.67   Mean   :0.2729   Mean   :0.3147  
 3rd Qu.:  8.000   3rd Qu.: 43.80   3rd Qu.:0.2800   3rd Qu.:1.0000  
 Max.   :221.000   Max.   :114.03   Max.   :4.5500   Max.   :1.0000  
```

The outcome which we are interested in predicting is the `biopsy` result, which we will assume indicates the "truth" in this case as to whether the patient actually has prostate cancer.

Question 10 (5 points)

To begin, use the `riff` data to build a regression model to predict whether the patient actually has prostate cancer on the basis of their PSA level, prostate volume, transurethral ultrasound result, digital rectal examination result and age.

Which predictors show a statistically detectable effect (at the 5% level) on the model, using Wald tests? (Note that more than one response may be selected.)

a. the subject's age
b. the subject's prostate-specific antigen level
c. the subject's prostate volume
d. the result of the subject's transurethral ultrasound
e. the result of the subject's digital rectal exam
f. None of the above

Question 11 (6 points)

Suppose you decide to use a cutpoint of a fitted probability of **0.3 or higher** for `biopsy` as your prediction rule to predict that the patient actually should be further screened for prostate cancer. Create a confusion matrix for the model you developed in Question 10. Use that matrix to specify:

a. the sensitivity
b. the specificity
c. the positive predictive value

under the prediction rule we've specified above. Specify your responses as **proportions** rounded to two decimal places.

Question 12 (4 points)

Below, you'll see the results of running a validation for the five-predictor model fit back in Question 10.

```
set.seed(2019); validate(mod10)
```

```
          index.orig training    test optimism index.corrected  n
Dxy           0.4714   0.5091  0.4639   0.0452          0.4262 40
R2            0.2399   0.2653  0.2246   0.0407          0.1992 40
Intercept     0.0000   0.0000 -0.1226   0.1226         -0.1226 40
Slope         1.0000   1.0000  0.8268   0.1732          0.8268 40
Emax          0.0000   0.0000  0.0644   0.0644          0.0644 40
D             0.1838   0.2058  0.1708   0.0350          0.1489 40
U            -0.0070  -0.0070  0.0067  -0.0137          0.0067 40
Q             0.1908   0.2128  0.1641   0.0487          0.1422 40
B             0.1772   0.1698  0.1809  -0.0111          0.1883 40
g             1.3651   1.5372  1.2782   0.2589          1.1062 40
gp            0.2015   0.2129  0.1926   0.0203          0.1813 40
```

Which of the following responses specifies an appropriate estimate of the area under the ROC curve that we would expect to see in new data, based on these cross-validations, and after rounding to three decimal places.

a. 0.971
b. 0.736
c. 0.732
d. 0.713
e. 0.509
f. 0.426
g. 0.199
h. None of these.

Question 13 (6 points)

Consider the following two-predictor model for the same `riff` data. In a sentence, state **and interpret** the odds ratio for `psa` in this model. Be sure to carefully describe the comparison made by the odds ratio.

```
model_13 <- riff %$%
    glm(biopsy ~ dre + psa, family = binomial())

summary(model_13)
```

```
Call:
glm(formula = biopsy ~ dre + psa, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0681  -0.8019  -0.7140   1.1847   1.8275

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.84442    0.28427  -6.488 8.68e-11 ***
dre          0.58289    0.28512   2.044 0.040921 *
psa          0.08909    0.02305   3.865 0.000111 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 356.24  on 285  degrees of freedom
Residual deviance: 319.11  on 283  degrees of freedom
AIC: 325.11

Number of Fisher Scoring iterations: 6
```

Question 14 (4 points)

Find a 90% confidence interval for the odds ratio associated with `dre` in the model fit in Question 13. Specify the lower and upper bounds for the confidence interval after rounding to two decimal places.

Quiz 1 Results

In all, students earned 76% of the available points.

- The class median was 58/75.
- The 75th percentile was 65/75.

If I had to put letters on the Quiz right now,

- the cutoff for an A grade is around 64 (approximately 85% of the 75 points available)
- the cutoff for a B grade is around 52.5 (approximately 70%)

In all, about three-quarters of the class fell in the A or B range, and several people were within a couple of points of either an A or a B.

If you didn't do well, remember that this Quiz is 75 of the 250 quiz points available this term, and that the projects and homework are also a major part of the course grade. I would focus my energy, were I you, on (a) figuring out what went wrong on the Quiz, and then (b) putting it out of my mind and doing the best job possible on Project 1.

Answer Sketch for Quiz 1

I added the `praise` package to help me with the Answer Sketch.

```
library(praise)
```

Answer 1 is a = 3, b = 2 and c = 4

- We want to maximize adjusted R^2 and model 3 does this.
- We want to minimize RMSE and model 2 does this.
- We also want to minimize MAE and model 4 does this.

Rubric

2 points for each correct response, so the maximum score is 6.

Results

Everyone got parts b and c right. At least 45/49 got a right, too.

```
praise("${EXCLAMATION}! This is ${adjective}!")
```

```
[1] "HUH! This is praiseworthy!"
```

Answer 2 is a few lines of code which produce the right Table 1

Here is what I used.

```
lind <- lind %>%
    mutate(treatment = fct_recode(factor(abcix),
                                  "abcix" = "1", "usual_care" = "0")) %>%
    mutate(treatment = fct_relevel(treatment, "abcix"))

temp_vars <- c("stent", "acutemi", "diabetic", "ejecfrac", "ves1proc", "cardbill")
temp_fac <- c("stent", "acutemi", "diabetic")
temp_str <- c("treatment")
tab1 <- CreateTableOne(data = lind, vars = temp_vars,
                       factorVars = temp_fac,
```

```
                       strata = temp_str)
print(tab1,
      nonnormal = c("cardbill"),
      exact = c("acutemi", "diabetic"))
```

```
                         Stratified by treatment
                          abcix
  n                            687
  stent = 1 (%)                484 (70.5)
  acutemi = 1 (%)              121 (17.6)
  diabetic = 1 (%)             139 (20.2)
  ejecfrac (mean (SD))       50.46 (10.38)
  ves1proc (mean (SD))        1.46 (0.70)
  cardbill (median [IQR]) 12901.00 [10882.50, 17067.50]
                         Stratified by treatment
                          usual_care                  p      test
  n                            283
  stent = 1 (%)                165 (58.3)             <0.001
  acutemi = 1 (%)               16 ( 5.7)             <0.001 exact
  diabetic = 1 (%)              73 (25.8)              0.060 exact
  ejecfrac (mean (SD))       52.93 (9.62)              0.001
  ves1proc (mean (SD))        1.20 (0.47)             <0.001
  cardbill (median [IQR]) 10169.00 [8282.50, 15684.00] <0.001 nonnorm
```

```
rm(temp_vars, temp_fac, temp_str)
```

Rubric

If your code produced my result, you should receive full credit, so long as you don't include anything which breaks the code.

- Note that you didn't need to remove the temporary variables you created.
- A correct table has ...
  - 5 columns (variable names, abcix, usual_care, p and test)
  - 6 rows (n, stent = 1 (%), acutemi = 1 (%), diabetic = 1 (%), ejecfrac (mean (SD)), ves1proc (mean(SD)), and cardbill (median [IQR]))
  - two exact tests (`acutemi` and `diabetic`)
  - 1 non-normal comparison (`cardbill`)
- As I saw it, you needed to provide code which would do seven things, without creating any extraneous output.
  - creating the `treatment` variable
  - sorting the `treatment` factor so that `abcix` appears first, not second
  - ensuring that `stent`, `acutemi` and `diabetic` are each treated as factors
  - ensuring that `acutemi` and `diabetic` are tested using exact tests, and that `stent` isn't
  - ensuring that `cardbill` is tested using the non-normal approach, but that `ejecfrac` and `ves1proc` are not
  - successfully print the table

Essentially, you got 10 points if you did all of these things correctly and lost two points for each of the items here that your code fails to complete, down to a minimum of 0 points. Also ...

- You lost at least 2 points if your code created extraneous output. This definitely includes warning messages.
- You lost at least 2 points if you wrote code that was incorrect, but didn't affect the final result.
- You lost at least 2 points if your code created more than one new tibble/data frame in the process of doing your work. (You didn't actually need to create any new tibbles, but creating one was OK. More

than one is wasteful, and unhelpful to those trying to follow your work.)

Results

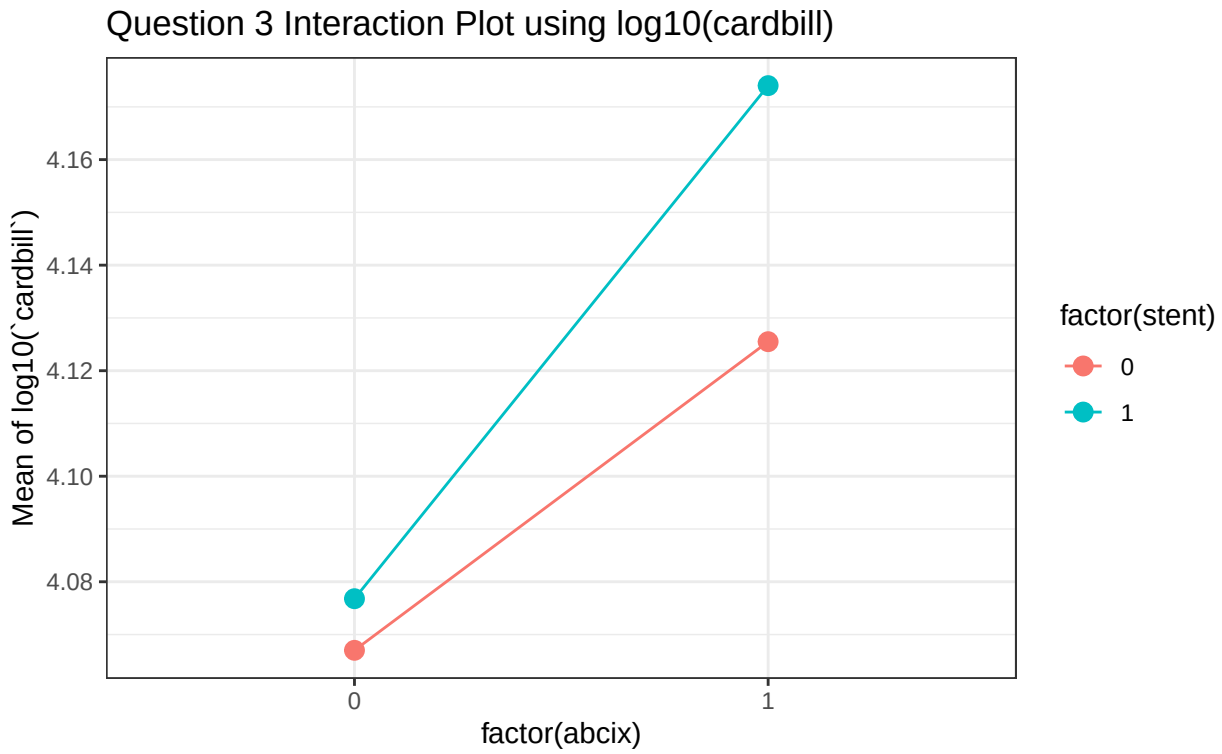
- 69% of students matched the table and did other things well enough to get the full 10 points.
- 88% of the possible points were awarded, including partial credit.

I have a R Markdown file with each student's code, and some comments. You'll find it posted to our Quiz web page. It's the `quiz1_checkquestion2code.Rmd` file. That way, you can see what I thought of your code.

Answer 3 is d.

```
summ_q3log <- lind %>% group_by(abcix, stent) %>%
    summarize(n = n(), mean = mean(log10(cardbill)))

ggplot(summ_q3log, aes(x = factor(abcix), y = mean,
                       col = factor(stent))) +
    geom_point(size = 3) +
    geom_line(aes(group = factor(stent))) +
    labs(y = "Mean of log10(`cardbill`)",
         title = "Question 3 Interaction Plot using log10(cardbill)")
```



We can look directly at the table of means, to verify that the observed difference meets our definition of "substantial":

```
summ_q3log %>%
  kable(digits = 3)
```

abcix	stent	n	mean
0	0	118	4.067

abcix	stent	n	mean
0	1	165	4.077
1	0	203	4.126
1	1	484	4.174

- The difference in mean (log costs) is 4.077 - 4.067 = 0.010 in the `abcix` = 0 group, and is
- The difference in mean (log costs) is 4.174 - 4.126 = 0.048 in the `abcix` = 1 group, and
- 0.048 is more than four times as large as 0.010 so the difference in means is substantial by our definition.

Rubric

This is a multiple-choice item. You got 4 points for a correct response, 2 points for choosing `a` (incorrectly declaring the interaction not to be substantial) and 0 otherwise.

Results

- 63% of students matched the best response.
- 77% of the possible points were awarded, including partial credit.

The most common incorrect response was `a`, naturally.

Answer 4 is 4.8, or 4.8%.

If you correctly identified the interaction term as substantial, then here's the result.

```
mod4 <- lind %$% lm(log10(cardbill) ~ abcix * stent)

anova(mod4)
```

```
Analysis of Variance Table

Response: log10(cardbill)
             Df Sum Sq Mean Sq F value    Pr(>F)
abcix         1  1.516 1.51622 39.5357 4.872e-10 ***
stent         1  0.274 0.27360  7.1340  0.007691 **
abcix:stent   1  0.070 0.06973  1.8183  0.177829
Residuals   966 37.047 0.03835
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The η^2 value is (1.516 + 0.274 + 0.070) / (1.516 + 0.274 + 0.070 + 37.047), or 0.0478063, which, expressed as a percentage would be 4.8%.

Rubric

4 points for a correct response, assuming your answer to Question 3 said that the interaction was substantial.

- 3 points for 0.048 (not expressing the response as a percentage.)
- you also got full credit (4 points) in Question 4 if you were incorrect in Question 3, and concluded that the interaction was not substantial, but then followed up on that choice appropriately and got η^2 = 4.6% here. (Actually, since some people had suggested in Question 3 that they couldn't decide, I just gave full credit for 4.6 to everyone in Question 4.)
- In fact, I also gave 3/4 points for people with 4.5 and 4.7.

Specifically, if you thought the interaction was not substantial, then you should have fit the model below...

```
mod4_noint <- lind %$% lm(log10(cardbill) ~ abcix + stent)

anova(mod4_noint)
```

```
Analysis of Variance Table

Response: log10(cardbill)
           Df Sum Sq Mean Sq F value   Pr(>F)
abcix       1  1.516 1.51622  39.502 4.95e-10 ***
stent       1  0.274 0.27360   7.128 0.007716 **
Residuals 967 37.116 0.03838
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and that no-interaction model leads to an η^2 value of (1.516 + 0.274) / (1.516 + 0.274 + 37.116), or 0.0460083, which, expressed as a percentage would be 4.6%.

Results

- 57% of students matched the best response.
- 63% of the possible points were awarded, including partial credit.

Answer 5 is some reasonable description of the effect size.

Consider model `mod5` as tabulated below.

```
mod5 <- lind %$% lm(log10(cardbill) ~ acutemi + ejecfrac)

tidy(mod5, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error,
           conf.low, conf.high, p.value) %>%
    kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	4.264	0.034	4.208	4.319	0.000
acutemi	-0.042	0.019	-0.072	-0.011	0.024
ejecfrac	-0.002	0.001	-0.003	-0.001	0.000

The value -0.042 indicates the estimated effect of having an acute MI in the past 7 days on the base 10 logarithm of cardiac costs, assuming that the ejection fraction is held constant. The 90% confidence interval (-0.072, -0.011) is an indication of the range of effect sizes that are reasonably consistent with the observed data. Since the confidence interval does not include zero, we can conclude that there is a statistically detectable effect (at the 10% significance level) of `acutemi` on `log10(cardbill)` after accounting for `ejecfrac`.

Equivalently, if we have two subjects, Harry and Sally, with the same ejection fraction (`ejecfrac`), but Harry had an acute MI in the past 7 days (so his `acutemi` = 1), but Sally did not (so her `acutemi` = 0), our model `mod5` predicts that the base 10 logarithm of Harry's cardiac costs will be 0.042 lower (with 90% CI 0.011 lower to 0.072 lower) than Sally's. Those effect sizes would have units of `log10(dollars)`, I suppose, but I didn't need you to specify the units here.

A third, trickier, approach was to anti-log the (-0.042) coefficient, with $10^(-0.042) = 0.908$, and recognize that this is describing the **multiplicative** effect on the raw (untransformed) `cardbill` costs.

- Specifically, the value $10^(-0.042) = 0.908$ is the estimated multiplier of `cardbill` costs for someone who had an acute MI in the past 7 days as compared to someone who did not, assuming the ejection

fraction is held constant.
- The person with `acutemi` = 1 will have 0.908 times the costs of the person with `acutemi` = 0 (and the same `ejecfrac`) according to the model.
- Alternatively, we could say that the first subject is modeled to have 90.8% of the costs of the second subject.
- If we anti-log the bounds of the confidence interval, we get $10^(-0.072) = 0.847$ and $10^(-0.011) = 0.974$ so our 90% confidence interval for the multiplier is (0.847, 0.974), or if you prefer, the 90% confidence interval for the multiplier associated with `acutemi` = 1 rather than 0 is (84.7%, 97.4%) of the `cardbill` costs.

Summarizing the third option:

- The model describes changes in the log of cardiac costs using a linear model.
- To understand the changes in terms of costs, you could anti-log the coefficient for `acutemi` but then you are describing the multiplicative effect on the original untransformed cardiac costs.

Many of you started your interpretation by finding the anti-log of -0.042 to be 0.908, and then explained what 0.908 means in terms of that being related to a *linear* effect (a difference) in the original costs somehow, and that was an approach which was going to fail. The most common incorrect response included a statement like this:

> ... if two people had the same left ventricular ejection fraction, then the one person who had a myocardial infarction in the last 7 days would have a card bill increased by 0.908 (as compared to one who has not had a myocardial infarction in the last 7 days).

Another common approach started OK, but well, if they'd just left off the last sentence here...

> ... For patient A and patient B with the same ejecfrac, if patient A has had an acute myocardial infarction in the past 7 days, our model predicts that the base 10 logarithm of their cardiac related bills will be 0.042 (-0.019, 0.072) lower than patient B who has not had an acute myocardial infarction. This means that patient B is predicted to pay $1685.13 (-817.53, 2792.83) more than patient A.

If you still don't see the problem with this approach, consider two subjects with `ejecfrac` = 55 (which was the median value observed in the `lind` data.) One of them (Art) has had an acute MI, and the other (Bill) has not.

- Our predicted value for Art in terms of log(costs) is then 4.264 - 0.042 (1) - 0.002 (55) = 4.112, and so Art's predicted costs are 10^{4.112} = $12,941.96.
- Our predicted value for Bill in terms of log(costs) is then 4.264 - 0.042 (0) - 0.002 (55) = 4.154, and so Art's predicted costs are 10^{4.112} = $14,256.08.

So for two subjects with `ejecfrac` = 55, the difference in log(costs) associated with `acutemi` = 1 instead of 0 is 0.042, just as the model expects, and my interpretation above indicates. But the difference in predicted costs is ($12,941.96 - $14,256.08) = -$1314.12.

Now, what if the two subjects had `ejecfrac` = 50 instead (close the mean value in `lind`). Then:

- Our predicted value for Art in terms of log(costs) is then 4.264 - 0.042 (1) - 0.002 (50) = 4.122, and so Art's predicted costs are 10^{4.122} = $13,243.42.
- Our predicted value for Bill in terms of log(costs) is then 4.264 - 0.042 (0) - 0.002 (50) = 4.164, and so Art's predicted costs are 10^{4.164} = $14,588.14.

So now, for two subjects with `ejecfrac` = 50, the difference in log(costs) associated with `acutemi` = 1 instead of 0 is still 0.042, just as the model expects and my interpretation above confirms, but the difference in predicted costs is now ($13,243.42 - $14,588.14) = -$1344.72. This is because the changes in costs are not linear in the `acutemi` or `ejecfrac` variables. They are in fact, non-linear.

So how can we describe these terms in terms of the predictors? Well, with a little arithmetic, we can see that

- when Art and Bill each have `ejecfrac` = 50, the *ratio* of their predicted costs is 12941.96/14256.08 = 0.90782, and this (of course) rounds to 0.908, which is 10^(-0.042).
- when Art and Bill each have `ejecfrac` = 50, the *ratio* of their predicted costs is 13243.42/14588.14 = 0.90782, and this (of course) also rounds to 0.908, which is 10^(-0.042).

and, in fact, this is the key. The use of the logarithmic transformation sets up a multiplicative effect of the `acutemi` variable on untransformed `cardbill` costs when the `ejecfrac` is held constant.

So, this is what leads me to my third description of the model.

- The model describes changes in the log of cardiac costs using a linear model.
- To understand the changes in costs, you could anti-log the coefficient for `acutemi` but then you are describing the multiplicative effect on the original untransformed cardiac costs.

Rubric

This is mostly about getting the details right. You got the full six points only if you:

- **either** correctly described the result in terms of a difference in `log10(costs)` depending on whether the indicator was yes or no, as long as the ejection fraction is held constant,
- **or** correctly described the result in terms of a multiplier for `costs` depending on whether the indicator was yes or no, as long as the ejection fraction is held constant
- and did all of the following things, too:
  - related your response back to the problem, identifying the outcome and predictors appropriately
  - described the confidence interval appropriately (including back-transforming if you're using the multiplier approach)
  - wrote in grammatically correct English sentences
  - didn't write anything that was incorrect (in particular, not applying incorrect units of measurement to your interpretation.)

You got at most 4/6 points (but probably less) if you failed to do any of these things, and your score was lower still if you failed to accomplish more than one of those things.

If you calculated 0.908 and misinterpreted it as some sort of a difference, the largest score you would have received if everything else was perfect was 3/6 points but most people with that going on also tried to put units to it that didn't make sense, and wrote other things that were incorrect.

There was no need to define what a confidence interval is in creating your response, but several people did try to do that, and if you didn't quite get that right, it also left you open to losing meaningful points. If you skipped the confidence interval entirely, you lost 3 points for that. If you described the confidence interval as meeting a standard for statistical significance, but didn't specify why you knew that (for instance that 0 was not in the interval), then you lost points for that, too.

Lots of people were focused on whether the confidence interval contained 0 or not which is not especially interesting in this case because we're not focused on statistically detectable findings, but it was really a problem if you made the transformation of the coefficient to 0.908, since then, we're interested in whether or not the coefficient is 1 to determine statistical significance, right? So that was another place where people tended to lose some points. Saying that there was a statistically significant `acutemi` effect without making it clear that was after adjustment for `ejecfrac` cost you a point, too. Writing that because of this interval we're 90% certain our model is true was also a major problem, since that's not at all what a confidence interval does. Also, what I presented and asked you to interpret was a 90% confidence interval, and not a 95% CI as some of you wrote.

I really wanted to see you specify that the ejection fractions needed to remain the same, not just that the subjects were "equal in every way" or "have the same factors" or "the same values of all other variables" or anything generic like that, so you lost at least a point for that.

Some folks mixed up the outcome and the predictor in trying to build their explanation of the effect. That's really a problem - and probably led you to no credit on the Question.

Some people mentioned the Gauss-Markov assumptions - not sure what would have led you to do that. This isn't a linear model in costs. It's a linear model in log(costs).

Results

Before the Quiz, I thought Questions 2, 5 and 13 would be the hardest ones to grade, and I thought 5 would be the hardest one for people to do well on. As it turned out, question 5 was indeed tough. I would anticipate that you'll see something like this again later in the term.

- 16% of students met my standard for a thoroughly solid response and received 6 points.
- 55% of the possible points were awarded, including partial credit.

Answer 6: Bob's predicted costs are $900 higher than Abby's.

The prediction for Abby is 4.120391 on the log scale, which works out to be $13,194.44

```
predict(mod6, newdata = tibble(ejecfrac = 50, abcix = 1, acutemi = 1))
```

```
       1
4.120391
```

```
10^predict(mod6, newdata = tibble(ejecfrac = 50, abcix = 1, acutemi = 1))
```

```
       1
13194.44
```

The prediction for Bob is 4.148946 on the log scale, which works out to be $14,091.14

```
predict(mod6, newdata = tibble(ejecfrac = 60, abcix = 1, acutemi = 0))
```

```
       1
4.148946
```

```
10^predict(mod6, newdata = tibble(ejecfrac = 60, abcix = 1, acutemi = 0))
```

```
       1
14091.14
```

The difference is 14091.14 - 13194.44 = 896.70. Rounding to the nearest $100, we have a difference of $900.

Rubric

- 2 points for successfully identifying Bob as having higher predicted costs.
- 3 points for correctly estimating the difference between Bob and Abby appropriately.

If you failed to convert to dollars correctly (perhaps by exponentiating with e instead of 10, or by forgetting to exponentiate entirely) that was a substantial problem and you got very little credit.

Results

- On 6a, 88% of students selected Bob.
- On 6b, 71% of students matched the best response.

There was no obvious pattern to the incorrect responses. No more than two students had the same incorrect response to part 6b.

Answer 7 is a = 3, b = 4, c = 1, d = 3, e = 2

Model 7a is a linear probability model predicting the probability of a stent. Model 7b is a logistic regression model predicting the log odds of a stent (as well as the odds or the probability of a stent.)

- `a` is true for BOTH models.
  - Each model predicts the probability of a stent, although the logistic model requires a little more work to get there.
- `b` is true for NEITHER model
  - In Model 7a, the coefficient for `ves1proc` is negative, so the predicted probability will drop as `ves1proc` increases, holding everything else constant.
  - In Model 7b, the odds ratio for `ves1proc` is less than 1, so the predicted probability will also drop as `ves1proc` increases, holding everything else constant.
- `c` is true for `mod7a` only
  - See note above.
- `d` is true for BOTH models.
  - See explanation under response `b`.
- `e` is true for `mod7b` only
  - See note above.

Rubric

1 point per correct response.

Results

Part	% Correct	Most Common Wrong Answer
7a	63	mod7a only
7b	33	mod7b only
7c	90	–
7d	27	mod7a only
7e	95+	–

Answer 8 is `e`.

Model `d` is unreasonable, but here are the other candidate models, fit using `lrm`.

```
mod8a <- lind %$% lrm(stent ~ poly(cardbill, 4) + abcix * ejecfrac + acutemi + ves1proc + diabetic)
mod8b <- lind %$% lrm(stent ~ rcs(cardbill, 5) + abcix + acutemi + ejecfrac + ves1proc + diabetic)
mod8c <- lind %$% lrm(stent ~ cardbill + abcix + acutemi + rcs(ejecfrac, 4) +
                          rcs(ves1proc, 4) + diabetic)
mod8e <- lind %$% lrm(stent ~ rcs(cardbill, 5) + abcix + abcix %ia% cardbill + acutemi + ejecfrac +
                          ves1proc + diabetic)
mod8f <- lind %$% lrm(stent ~ abcix * diabetic + diabetic * cardbill +
                          abcix * cardbill + acutemi + ejecfrac + ves1proc + diabetic)
```

We clearly want to emphasize `cardbill` and then `abcix` according to the plot. We also have to use exactly four additional degrees of freedom.

- The problem with option `a` is that it spends df on `ejecfrac` which is far down the priority list.
- The problem with option `b` is that it only spends 9 df, rather than 10.
- The problem with option `c` is that it adds non-linearity to low priority predictors, and that it tries to fit a four-knot spline to a variable (`ves1proc`) with only six different values.
- The problem with option `d` is that it tries to fit a cubic polynomial to a binary predictor, which makes no sense.
- The problem with option `f` is that it also only spends 9 df, rather than 10.

Only option `e` takes the advice of the Spearman plot and also adds exactly four degrees of freedom for non-linearity, as demonstrated below.

```
anova(mod8e)
```

```
                Wald Statistics          Response: stent

 Factor                                                Chi-Square d.f. P
 cardbill  (Factor+Higher Order Factors)               64.98       5   <.0001
  All Interactions                                     11.65       1   0.0006
  Nonlinear                                            61.57       3   <.0001
 abcix  (Factor+Higher Order Factors)                  11.68       2   0.0029
  All Interactions                                     11.65       1   0.0006
 abcix * cardbill  (Factor+Higher Order Factors)       11.65       1   0.0006
 acutemi                                               0.03       1   0.8692
 ejecfrac                                              0.56       1   0.4545
 ves1proc                                              6.18       1   0.0129
 diabetic                                              0.17       1   0.6815
 TOTAL NONLINEAR + INTERACTION                        62.07       4   <.0001
 TOTAL                                                78.87      10   <.0001
```

Rubric

Multiple choice question, so 4 points for the correct response. In a moment of weakness, I decided to give 2 points for the responses (`b` and `f`) that were only incorrect because they spent 9 df instead of 10.

Results

- 43% of students matched the best response.
- 62% of the possible points were awarded, including partial credit.

Answer 9 is `abcfg` for adjusted R-sq, `acfg` for Cp and `ag` for BIC

Summary Measure	Recommended Model	# of predictors
adjusted R^2	`a`, `b`, `c`, `f` and `g`	5
C_p	`a`, `c`, `f` and `g`	4
BIC	`a` and `g`	2

Just verifying that I haven't made a mistake here...

```
tibble(AdjR2 = which.max(summary(rs9)$adjr2),
       Cp = which.min(summary(rs9)$cp),
       BIC = which.min(summary(rs9)$bic),)
```

```
# A tibble: 1 x 3
  AdjR2    Cp   BIC
  <int> <int> <int>
1     5     4     2
```

```
coef(rs9, id = 5)
```

```
 (Intercept)            a            b            c            f            g
45.980590887  0.639958059 -0.001960561  0.582267036  0.834668950 11.932634130
```

```
coef(rs9, id = 4)
```

```
(Intercept)           a           c           f           g
 45.0389112   0.6446395   0.5587588   0.8421471  11.9479764
```

```
coef(rs9, id = 2)
```

```
(Intercept)           a           g
 50.9682694   0.6474889  12.5083112
```

Rubric

- Each of the summary statistics is graded separately. If you specified the correct model, that's worth 2 points.
- So, since there are three summaries, the maximum score is 6.

Results

Part	% correct
9a (Adj. R-sq)	92
9b (Cp)	95+
9c (BIC)	90

```
praise("${Exclamation}! You did this ${adverb_manner}!")
```

```
[1] "Gee! You did this kindly!"
```

Answer 10 is b, c and d

```
dd <- datadist(riff)
options(datadist = "dd")

mod10 <- lrm(biopsy ~ age + dre + tru + vol + psa,
               data = riff, x = TRUE, y = TRUE)

mod10
```

```
Logistic Regression Model

 lrm(formula = biopsy ~ age + dre + tru + vol + psa, data = riff,
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           286    LR chi2      53.58    R2       0.240    C       0.736
  0            196    d.f.             5    g        1.365    Dxy     0.471
  1             90    Pr(> chi2) <0.0001    gr       3.916    gamma   0.472
 max |deriv| 5e-09                          gp       0.202    tau-a   0.204
                                            Brier    0.177

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -0.8152 1.1464 -0.71  0.4770
 age       -0.0046 0.0175 -0.26  0.7943
 dre        0.2862 0.3039  0.94  0.3462
 tru        0.7196 0.2867  2.51  0.0121
 vol       -0.0280 0.0097 -2.87  0.0041
 psa        0.1011 0.0259  3.90  <0.0001
```

Rubric

The observed responses (and the points I awarded for them) were:

Response	Points
b, c and d	5
b and c only	3
b and d only	3
b, c and e	3
All other responses	0

Results

- 84% of students matched the correct response.
- 91% of the possible points were awarded, including partial credit.
- The most common incorrect response was `b` and `d` only.

Answer 11: Sensitivity is 0.66, Specificity is 0.71 and PPV = 0.51

First, we want to obtain the confusion matrix using a cutoff of 0.3. To do so, I'll refit the model using `glm` first, then use `augment` from `broom` and `confusionMatrix` from `caret` to get what I need.

```
mod10_glm <- riff %$%
    glm(biopsy ~ age + dre + tru + vol + psa,
        family = binomial())

m10_aug <- augment(mod10_glm, type.predict = "response")

m10_aug %$% confusionMatrix(
    data = factor(.fitted >= 0.3),
    reference = factor(biopsy == 1),
    positive = "TRUE")
```

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE   140   31
     TRUE     56   59

               Accuracy : 0.6958
                 95% CI : (0.6389, 0.7486)
    No Information Rate : 0.6853
    P-Value [Acc > NIR] : 0.37785

                  Kappa : 0.344

 Mcnemar's Test P-Value : 0.01008

            Sensitivity : 0.6556
            Specificity : 0.7143
         Pos Pred Value : 0.5130
         Neg Pred Value : 0.8187
             Prevalence : 0.3147
```

```
      Detection Rate : 0.2063
Detection Prevalence : 0.4021
   Balanced Accuracy : 0.6849

    'Positive' Class : TRUE
```

Rubric

Generally, you had three statistics to specify, each worth 2 points, for a maximum total of 6 points. But there were a couple of specific errors you could have made that we anticipated and assigned partial credit, as follows...

- If you failed to round, or failed to specify as a proportion, but were otherwise correct, you should have lost 0.5 point for each such problem.
- If you rounded poorly, you lost a point.
- If you used 0.5 instead of 0.3 to define the prediction rule, your answers would have been sensitivity = 0.27, specificity = 0.93 and PPV = 0.65, so if you gave those answers you would have received a total of 4 out of 6 points, assuming you presented them as proportions.
- If you'd used FALSE as the positive group instead of true, you would have gotten sensitivity = 0.71, specificity = 0.66 and PPV = 0.82, for which you received a total of 4 points.
- If you forgot to use `type.predict = "response"` in your setup, and used 0.3, your answers would have been sensitivity = 0.22, specificity = 0.97 and PPV = 0.80, and you would have received 1 point on each part, for a total of 3 out of 6 points.
- If you forgot to use `type.predict = "response"` in your setup, and used 0.5, your answers would have been sensitivity = 0.20, specificity = 0.99 and PPV = 0.90, and you would have received 2 out of 6 points.
- If you forgot to specify that the `glm` should use the `binomial` family, but did everything else correctly, you would have gotten something like sensitivity = 0.68, specificity = 0.64 and PPV = 0.47. As of 1 PM on 2020-03-03, I amended the grades to give students in that situation 2 out of 6 points.

Results

- 53% of students matched the correct response.
- 65% of the possible points were awarded, including partial credit.

A somewhat common incorrect response, for which I expect there's a good explanation, but I don't yet know what it is, was...

- sensitivity = 0.63, specificity = 0.79 and PPV = 0.58

Someone will tell me, I'm sure, how they got those responses, by sending a note to `431-help` before class on Thursday 2020-03-05. Perhaps, after I see how you got these responses, I might consider some partial credit. But for now, no points.

Answer 12 is d.

What we need to do here is first identify the correct value of Somers' d, accounting for the validation. This is the value `d = 0.4262` that we find in the `index.corrected` column.

Then, to convert this to a statement about the area under the ROC curve (designated by C), we use the formula C = 0.5 + d/2, so that the correct response is 0.5 + (0.4262/2) = 0.7131, or 0.713 after rounding to three decimal places. That's option `d`.

Rubric

Multiple choice question. 4 points for the correct response, otherwise 0.

Results

- 69% of students matched the correct response.
- The most common incorrect response was `b`, which is what you got if you used the `index.orig` instead of the `index.corrected` column in the `validate` output, and then applied the formula to convert from Somers' d to the C statistic.

Answer 13 is a few sentences describing the 1.093 odds ratio estimate for `psa`.

Consider the following two-predictor model for the same `riff` data. In a sentence, state **and interpret** the odds ratio for `psa` in this model. Be sure to carefully describe the comparison made by the odds ratio. Of course, we need to exponentiate the coefficients, here with the `tidy` function:

```
model_13 <- riff %$%
    glm(biopsy ~ dre + psa, family = binomial())

tidy(model_13, exponentiate = TRUE) %>%
  select(term, estimate) %>%
  kable(digits = 3)
```

term	estimate
(Intercept)	0.158
dre	1.791
psa	1.093

The odds ratio for `psa` is 1.093, which means that if we have two subjects (A and B) with the same digital rectal exam result, but subject A has a `psa` level that is 1 ng/ml higher than subject B, the odds of a positive biopsy are estimated by this model to be 1.093 times as high for subject A as for subject B. If you prefer, you could say that subject A's odds are 9.3% higher than subject B's.

Rubric

6 points for an appropriate interpretation, as described above. You needed to:

- write in complete and grammatically correct English sentences,
- describe the results in terms of the actual variables under study,
- specify the units of measurement
- correctly identify the size of the effect
- use appropriate language to describe the observed association (this is about a model estimate)

If you did all of those things you got 6 points. If you didn't, you got at most 4/6.

- If you included the CI and it was correct, I ignored it. If you included the CI and it was wrong or your interpretation of it was wrong, you lost points.
- If you didn't specify what the other predictor was besides `psa` and thus provide context, you lost 2 points.
- If you didn't mention that `dre` needed to be held constant in order to interpret the `psa` odds ratio, you lost 3 points. **This was by far the most common problem.**
- If you said the `dre` results needed to be similar, instead of that they needed to be identical, you lost a point for that.
- If you said that the `dre` results had to each be 0, instead of just that they needed to be the same, you lost a point.
- If you wrote that the odds increased by 1.093 instead of "increaed by 9.3% or"would be multiplied by 1.093" or "increased by a factor of 1.093" or "would be 1.093 times as large" or the like, you lost 2 points.

- If you misread the names of the variables, and talked about a variable not included in this model, you lost at least 2 points for that.
- If you thought that the initial output was the odds ratio rather than the log odds, you lost at least 3 points. If you described the log odds and identified it as such, rather than the odds ratio, you lost 3 points.
- If you interpreted an odds ratio as a relative risk, that lost you 3 points.
- Some people didn't mention that this was the conclusion of the model, and draw broader causal conclusions. Don't do that.
- You really should have said a 1 ng/ml increase in `psa` and not just a one unit increase, but I sometimes let that slide, depending on the rest of your response.
- If you used Harry and Sally as the names here, I gave you the benefit of the doubt and assumed that Sally was short for Salvatore.

Results

- 33% of students met my standard for a thoroughly solid response and received 6 points.
- 69% of the possible points were awarded, including partial credit.

Answer 14 is (1.13, 2.89).

```
model_13 <- riff %$%
    glm(biopsy ~ dre + psa, family = binomial())

tidy(model_13, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.9) %>%
    select(term, estimate, conf.low, conf.high)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# A tibble: 3 x 4
  term        estimate conf.low conf.high
  <chr>          <dbl>    <dbl>     <dbl>
1 (Intercept)    0.158   0.0972     0.248
2 dre            1.79    1.13       2.89
3 psa            1.09    1.06       1.14
```

Rubric

- 4 points available: 2 points for the correct lower bound and 2 points for the correct upper bound.
- If you fit the 95% confidence interval that would have been (1.03, 3.17), for which you'd receive 2 points out of 4.
- If you mis-rounded, and got, for instance, 1.12 for the lower bound instead of 1.13 you lost a point.
- If you didn't exponentiate, you would have a point estimate for `dre` of 0.58 a 90% CI of (0.12, 1.06). That was worth 1 point out of 4.
- If you gave the confidence interval for `psa` instead of `dre`, that would have been (1.06, 1.14) and you would have recieved 1 point out of 4 for that. If you forgot to exponentiate, no points.
- You weren't required to interpret this result.
  - If you tried to interpret the result and did so correctly, fine.
  - If you tried to interpret the result and did so incorrectly, you lost at least 2 points.
- You also weren't required to show the point estimate. If you did so correctly, though, OK.

Results

- 71% of students matched the correct response.
- 81% of the possible points were awarded, including partial credit.