432 Class 12

https://thomaselove.github.io/432-2023/

2023-02-23

Today's Agenda

- Using `caret` to help with k-fold cross validation
- Building a Table One
- Setting Up Quiz One

Today's R Setup

```
knitr::opts_chunk$set(comment = NA)

library(janitor)
library(broom)
library(knitr)
library(caret)
library(tableone)
library(tidyverse)

theme_set(theme_bw())
```

Section 1

K-Fold Cross-Validation

The maleptsd data from last time

The maleptsd file on our web site contains information on PTSD (post traumatic stress disorder) symptoms following childbirth for 64 fathers[1]. There are ten predictors and the response is a measure of PTSD symptoms. The raw, untransformed values (ptsd_raw) are right skewed and contain zeros, so we will work with a transformation, specifically, ptsd = log(ptsd_raw + 1) as our outcome, which also contains a lot of zeros.

```
maleptsd <- read_csv("c11/data/maleptsd.csv", show_col_types =
  clean_names() |>
  mutate(ptsd = log(ptsd_raw + 1))
```

[1]Source: Ayers et al. 2007 *J Reproductive and Infant Psychology*. The data are described in more detail in Wright DB and London K (2009) *Modern Regression Techniques Using R* Sage Publications.

Remember the problem

Only 64 observations, 10 predictors. We came up with a lasso model which used 5 of the predictors, specifically `over3`, `bond`, `neg`, `sup` and `aff`.

```
m1 <- lm(ptsd ~ over3 + bond + neg + sup + aff,
         data = maleptsd)

glance(m1) |> select(r2 = r.squared, adjr2 = adj.r.squared,
                     AIC, BIC) |> kable(digits = c(4,4,2,2))
```

r2	adjr2	AIC	BIC
0.2604	0.1966	205.99	221.11

Set up five-fold cross-validation

We'll use the `trainControl()` function from the **caret** package.

```
set.seed(43212345)
ctrl <- trainControl(method = "cv", number = 5)
```

Next, we train our model on those five folds:

```
ptsd_mod <- train(ptsd ~ over3 + bond + neg + sup + aff,
                  data = maleptsd, method = "lm",
                  trControl = ctrl)
```

Results on next slide.

ptsd_mod results

```
> ptsd_mod
Linear Regression

64 samples
 5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 52, 51, 52, 50, 51
Resampling results:

  RMSE      Rsquared   MAE
  1.162811  0.2275055  0.9904286

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Compare this to the nominal R^2 we saw earlier of 0.2604.

A New Model with Two Predictors

Perhaps we can justify a two-predictor model.

```
m2 <- lm(ptsd ~ aff + neg,
         data = maleptsd)

glance(m2) |> select(r2 = r.squared, adjr2 = adj.r.squared,
                     AIC, BIC) |> kable(digits = c(4,4,2,2))
```

r2	adjr2	AIC	BIC
0.2025	0.1764	204.82	213.45

Train our new model on the same 5 folds

Next, we train our new model on our five folds:

```
ptsd_mod2 <- train(ptsd ~ neg + aff,
                   data = maleptsd, method = "lm",
                   trControl = ctrl)
```

Results on next slide...

ptsd_mod2 cross-validation results

```
> ptsd_mod2
Linear Regression

64 samples
 2 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 50, 50, 52, 52, 52
Resampling results:

  RMSE      Rsquared   MAE
  1.174825  0.1594224  0.9971287

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Compare this to the nominal R^2 we saw earlier of 0.2025

Model Summaries within each of the 5 folds

```
ptsd_mod2$resample
```

```
      RMSE   Rsquared       MAE Resample
1 1.1024703 0.10193142 0.8674134    Fold1
2 1.2309474 0.12985532 1.0764762    Fold2
3 0.8744284 0.52695861 0.7348838    Fold3
4 1.4177608 0.00765152 1.2705278    Fold4
5 1.2485165 0.03071503 1.0363423    Fold5
```

Final Model from cross-validation

```
ptsd_mod2$finalModel
```

```
Call:
lm(formula = .outcome ~ ., data = dat)

Coefficients:
(Intercept)          neg          aff
   -0.23311      0.04191      0.10648
```

```
glance(ptsd_mod2$finalModel) |>
  select(r2 = r.squared, adjr2 = adj.r.squared,
         AIC, BIC) |> kable(digits = c(4,4,2,2))
```

r2	adjr2	AIC	BIC
0.2025	0.1764	204.82	213.45

Tidied Coefficients from C-V model 2

```
tidy(ptsd_mod2$finalModel, 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)	-0.233	0.527	-1.113	0.646	0.660
neg	0.042	0.013	0.021	0.063	0.001
aff	0.106	0.047	0.027	0.185	0.028

Learning More about V-fold Cross-Validation

There's another example in section 16.5 of our Course Notes.

- There's more on the caret package at https://topepo.github.io/caret/ although that's older now, and the tidymodels approach will allow us to do a lot of the same things later this term.

Can you do something similar to this with a `glm()` fit in logistic regression?

- Yes, definitely.

Section 2

Building a Table One

An Original Clinical Investigation

Original Investigation | Cardiology

January 18, 2019

Incidence, Risk Factors, and Outcomes Associated With In-Hospital Acute Myocardial Infarction

Steven M. Bradley, MD, MPH[1,2]; Joleen A. Borgerding, MS[3]; G. Blake Wood, MS[3]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2019;2(1):e187348. doi:10.1001/jamanetworkopen.2018.7348

Key Points

Question What are the incidence, risk factors, and outcomes associated with in-hospital acute myocardial infarction (AMI)?

Findings This cohort study of 1.3 million patients hospitalized in US Veterans Health Administration facilities found an incidence of in-hospital AMI of 4.27 per 1000 admissions, and risk factors associated with in-hospital AMI included history of coronary artery disease, elevated heart rate, low hemoglobin level, and elevated white blood cell count. Compared with a matched control group, mortality was significantly higher for in-hospital AMI.

Meaning In-hospital AMI is common and is associated with prior cardiovascular disease, physiological disturbances, and poor survival.

Link to Source

Part of Bradley et al.'s Table 1

Table 1. Patient Characteristics on Admission and In-Hospital Variables Prior to Event for Matched In-Hospital Acute Myocardial Infarction Cases and Controls

Characteristic	No. (%)			*P* Value
	Total (N = 1374)	Cases (n = 687)	Controls (n = 687)	
Age, mean (SD), y	73.3 (10.2)	73.3 (10.1)	73.4 (10.3)	.80
Male	1343 (97.7)	677 (98.5)	666 (96.9)	.05
White race/ethnicity	1073 (78.1)	546 (79.5)	527 (76.7)	.22
Married	666 (48.5)	356 (51.8)	310 (45.1)	.01
Location				
Intensive care unit	251 (18.3)	186 (27.1)	65 (9.5)	<.001
Medical bed	1026 (74.7)	446 (64.9)	580 (84.4)	
Other	97 (7.1)	55 (8.0)	42 (6.1)	

Table Creation Instructions, JAMA: linked here

Creating a Table

Use the table editor of the word processing software to build a table. Do not embed tables as images in the manuscript file or upload tables in image formats. Regardless of which program is used, each piece of data needs to be contained in its own cell in the table. Tables should be single-spaced.

Avoid creating tables using spaces or tabs. For accepted manuscripts, tables created with spaces, tabs, and/or hard returns must be retyped during the editing process, creating delays and opportunities for error. Do not try to align cells with hard returns or extra spaces. Similarly, no cell should contain a hard return or tab. Although individual empty cells are acceptable in a table, be sure there are no empty columns.

Place each row of data in a separate row of cells:

Table 1. Title

Treatment	Group A	Group B
Medical	500	510
Surgical	500	490

Note that numbers and percentages are presented in the same cell, and measures of variability are in the same cell as their corresponding statistic:

Table 2. Title

Characteristics	Group A (n = 50)	Group B (n = 50)	Relative Risk (95% CI)
Women, No. (%)	25 (50)	20 (40)	1.25 (1.11-1.57)
Age, mean (SD), y	35 (8)	37 (7)	0.98 (0.92-1.05)

To present data that span more than 1 row, do not merge the cells vertically. Instead, put the data in a cell near the middle of the rows. In Table 3, the final column lists the *P* value for the overall age comparison:

Table 3. Title

Age, y	Blood Pressure, mm Hg	*P* Value
18-34	120/75	
35-50	110/80	.08
51-80	125/82	

The table should be constructed such that comparisons between groups read horizontally (see Tables 1 and 2).

Do not draw lines or rules—the table grid feature will display the outlines of each cell.

Data Presentation

When presenting percentages, include numbers (numerator, and denominator if necessary). Include variability where applicable (eg, mean [SD] or median [interquartile range]).

All *P* values should be reported as exact numbers to 2 digits past the decimal point, regardless of significance, unless they are lower than .01, in which case they should be presented to 3 digits. Express any *P* values lower than .001 as $P<.001$. *P* values can never equal 0 or 1.

Footnotes

Be sure to explain empty cells. Also, if necessary add a footnote to explain why numbers may not sum to group totals or percentages do not total 100. List abbreviations for the table in a footnote and use superscript letters to mark each footnote (a,b,c, etc).

Questions

For questions on table construction or formatting, contact Stacy Christiansen, director of manuscript editing, at stacy.christiansen @jama-archives.org

A Data Set

The `bradley.csv` data set on our web site is simulated, but consists of 1,374 observations (687 Cases and 687 Controls) containing:

- a subject identification code, in `subject`
- `status` (case or control)
- age (in years)
- sex (Male or Female)
- race/ethnicity (white or non-white)
- married (1 = yes or 0 = no)
- location (ICU, bed, other)

The `bradley.csv` data closely match the summary statistics provided in Table 1 of the Bradley et al. article. Our job is to recreate that part of Table 1, as best as we can.

The `bradley.csv` data (first 5 rows)

- The `bradley_sim.md` file on our web site shows you how I simulated the data.

	A	B	C	D	E	F	G
1	subject	status	age	sex	race_eth	married	location
2	1	Control	64	Male	white	1	Bed
3	2	Case	70	Male	white	1	ICU
4	3	Control	68	Male	white	0	Bed
5	4	Control	76	Male	white	1	Bed
6	5	Control	70	Male	white	1	Bed

To "Live" Coding

On our web site (Data and Code + Class 12 materials)

- In the `data` folder:
  - `bradley.csv` data file
- `bradley_table1.Rmd` R Markdown script
- `bradley_table1.md` Results of running R Markdown
- `bradley_table1_result.csv` is the table generated by that R Markdown script

Section 3

To The "Live Code"

Opening `bradley_table1_result.csv` in Excel

	A	B	C	D	E
1		Case	Control	p	test
2	n	687	687		
3	age (mean (SD))	73.78 (10.24)	72.60 (10.50)	0.035	
4	sex = Male (%)	677 (98.5)	666 (96.9)	0.069	
5	race_eth = white (%)	546 (79.5)	527 (76.7)	0.24	
6	marital = yes (%)	356 (51.8)	310 (45.1)	0.015	
7	loc (%)			<0.001	
8	ICU	186 (27.1)	65 (9.5)		
9	Bed	446 (64.9)	580 (84.4)		
10	Other	55 (8.0)	42 (6.1)		
11					

Learning More About Table 1

Chapter 18 of the Course Notes covers two larger examples, and more details, like...

- specifying factors, and re-ordering them when necessary
- using non-normal summaries or exact categorical tests
- dealing with warning messages and with missing data
- producing Table 1 in R so you can cut and paste it into Excel or Word

FYI: Lab 05 (due 2023-03-06) requires you to build a Table 1 from data.

Next Time

Thinking About Power: Retrospective Design

Good luck on the Quiz!