432 Class 17 Slides

github.com/THOMASELOVE/432-2018

2018-03-20

Setup

```r
library(skimr)
library(rms)
library(broom)
library(tidyverse)

colscr <- read.csv("screening.csv") %>% tbl_df
colscr2 <- read.csv("screening2.csv") %>% tbl_df
```

Today's Materials

0. Start of Class Announcements
1. Logistic Regression for Aggregated Data
2. Probit Regression for a Binary Outcome
3. Project 1 Group Meetings

Logistic Regression for Aggregated Data

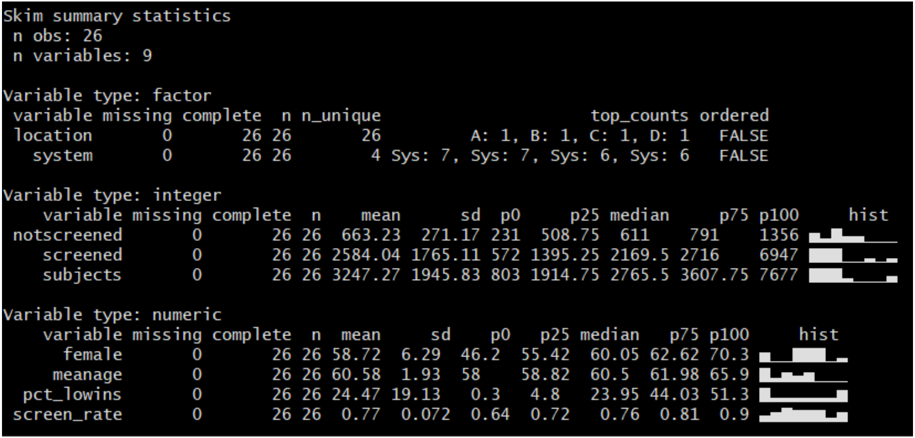
Colorectal Cancer Screening Data

The `screening.csv` data (imported into the R tibble `colscr` are simulated. They mirror a subset of the actual results from the Better Health Partnership's pilot study of colorectal cancer screening in primary care clinics in Northeast Ohio.

Available to us are the following variables

Variable	Description
location	clinic code
subjects	number of subjects reported by clinic
screen_rate	proportion of subjects who were screened
screened	number of subjects who were screened
notscreened	number of subjects not screened
meanage	mean age of clinic's subjects, years
female	% of clinic's subjects who are female
pct_lowins	% of clinic's subjects who have Medicaid or are uninsured
system	system code

Skim results



```
Skim summary statistics
 n obs: 26
 n variables: 9

Variable type: factor
 variable missing complete  n n_unique                     top_counts ordered
 location       0       26 26       26         A: 1, B: 1, C: 1, D: 1   FALSE
   system       0       26 26        4 Sys: 7, Sys: 7, Sys: 6, Sys: 6   FALSE

Variable type: integer
    variable missing complete  n    mean      sd  p0     p25 median     p75 p100     hist
 notscreened       0       26 26  663.23  271.17 231  508.75  611    791    1356
    screened       0       26 26 2584.04 1765.11 572 1395.25 2169.5 2716    6947
    subjects       0       26 26 3247.27 1945.83 803 1914.75 2765.5 3607.75 7677

Variable type: numeric
    variable missing complete  n  mean     sd    p0   p25 median   p75 p100     hist
      female       0       26 26 58.72  6.29  46.2  55.42  60.05 62.62 70.3
     meanage       0       26 26 60.58  1.93  58    58.82  60.5  61.98 65.9
  pct_lowins       0       26 26 24.47 19.13   0.3   4.8  23.95 44.03 51.3
 screen_rate       0       26 26  0.77  0.072  0.64  0.72   0.76  0.81  0.9
```

Fitting a Logistic Regression Model to Proportion Data

Here, we have a binary outcome (was the subject screened or not?) but we have aggregated results. We can use the counts of the numbers of subjects at each clinic (in `subjects`) and the proportion who were screened (in `screen_rate`) to fit a logistic regression model, as follows:

```r
m_screen1 <-  glm(screen_rate ~ meanage + female +
                    pct_lowins + system, family = binomial,
                  weights = subjects, data = colscr)
```

tidy(m_screen1)

```
         term    estimate    std.error   statistic
1 (Intercept) -1.32703925 0.5530782215  -2.3993699
2     meanage  0.06798655 0.0089754129   7.5747549
3      female -0.01931425 0.0015830906 -12.2003429
4  pct_lowins -0.01345472 0.0008585381 -15.6716603
5 systemSys_2 -0.13821891 0.0246591342  -5.6051809
6 systemSys_3 -0.04001702 0.0254505472  -1.5723443
7 systemSys_4  0.02292732 0.0294207148   0.7792918
       p.value
1 1.642331e-02
2 3.598062e-14
3 3.095177e-34
4 2.363243e-55
5 2.080376e-08
6 1.158707e-01
7 4.358078e-01
```

Fitting Counts of Successes and Failures

```
m_screen2 <-  glm(cbind(screened, notscreened) ~
                    meanage + female + pct_lowins + system,
           family = binomial, data = colscr)
```

tidy(m_screen2)

```
         term    estimate    std.error   statistic
1 (Intercept) -1.32703925 0.5530782214  -2.3993699
2     meanage  0.06798655 0.0089754129   7.5747549
3      female -0.01931425 0.0015830906 -12.2003430
4  pct_lowins -0.01345472 0.0008585381 -15.6716604
5 systemSys_2 -0.13821891 0.0246591342  -5.6051809
6 systemSys_3 -0.04001702 0.0254505472  -1.5723443
7 systemSys_4  0.02292732 0.0294207148   0.7792918
       p.value
1 1.642331e-02
2 3.598062e-14
3 3.095174e-34
4 2.363242e-55
5 2.080375e-08
6 1.158707e-01
7 4.358078e-01
```

How does one address this problem in `rms`?

We can use `Glm`.

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

mod_screen_1 <-  Glm(screen_rate ~ meanage + female +
                         pct_lowins + system,
                     family = binomial, weights = subjects,
                     data = colscr, x = T, y = T)
```

mod_screen_1

```
General Linear Model

 Glm(formula = screen_rate ~ meanage + female + pct_lowins + s
     family = binomial, data = colscr, weights = subjects, x =
     y = T)

                      Model Likelihood
                         Ratio Test
 Obs       26         LR chi2    2008.90
 Residual d.f.19      d.f.             6
 g 0.4614539          Pr(> chi2) <0.0001

              Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept    -1.3270 0.5531  -2.40 0.0164
 meanage       0.0680 0.0090   7.57 <0.0001
 female       -0.0193 0.0016 -12.20 <0.0001
 pct_lowins   -0.0135 0.0009 -15.67 <0.0001
```

Probit Regression

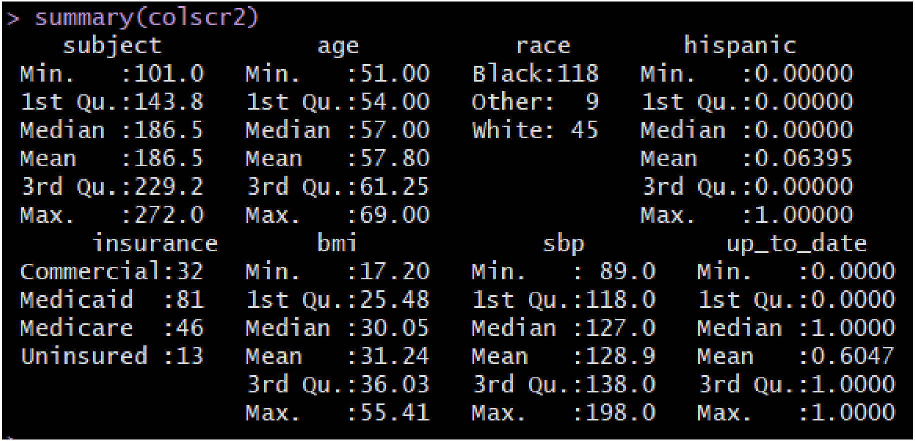
Colorectal Cancer Screening Data on Individuals

The data in the `colscr2` data frame describe (disguised) data on the status of 172 adults who were eligible for colon cancer screening. The goal is to use the other variables (besides subject ID) to predict whether or not a subject is up to date.

colscr2 contents

Variable	Description
subject	subject ID code
age	subject's age (years)
race	subject's race (White/Black/Other)
hispanic	subject of Hispanic ethnicity (1 = yes / 0 = no)
insurance	Commercial, Medicaid, Medicare, Uninsured
bmi	body mass index at most recent visit
sbp	systolic blood pressure at most recent visit
up_to_date	meets colon cancer screening standards

summary(colscr2)



```
> summary(colscr2)
    subject           age             race        hispanic      
 Min.   :101.0   Min.   :51.00   Black:118   Min.   :0.00000  
 1st Qu.:143.8   1st Qu.:54.00   Other:  9   1st Qu.:0.00000  
 Median :186.5   Median :57.00   White: 45   Median :0.00000  
 Mean   :186.5   Mean   :57.80               Mean   :0.06395  
 3rd Qu.:229.2   3rd Qu.:61.25               3rd Qu.:0.00000  
 Max.   :272.0   Max.   :69.00               Max.   :1.00000  
      insurance       bmi             sbp          up_to_date    
 Commercial:32   Min.   :17.20   Min.   : 89.0   Min.   :0.0000  
 Medicaid  :81   1st Qu.:25.48   1st Qu.:118.0   1st Qu.:0.0000  
 Medicare  :46   Median :30.05   Median :127.0   Median :1.0000  
 Uninsured :13   Mean   :31.24   Mean   :128.9   Mean   :0.6047  
                 3rd Qu.:36.03   3rd Qu.:138.0   3rd Qu.:1.0000  
                 Max.   :55.41   Max.   :198.0   Max.   :1.0000  
```

A logistic regression model

```
m_scr2_logistic <- glm(up_to_date ~ age + race + hispanic +
                    insurance + bmi + sbp,
                  family = binomial, data = colscr2)
```

Results

```
                 term      estimate   std.error
1         (Intercept)  2.7040470104 2.741862469
2                 age  0.0204900528 0.039692006
3           raceOther -1.9722351207 1.002323683
4           raceWhite -0.3210458270 0.400174430
5            hispanic  0.0005854686 0.795348176
6   insuranceMedicaid -1.0151859843 0.494516885
7   insuranceMedicare -0.5216005528 0.562993549
8  insuranceUninsured  0.1099966224 0.790619608
9                 bmi  0.0155894129 0.021354689
10                sbp -0.0241776892 0.009913777
       statistic    p.value
1   0.9862081126 0.32403100
2   0.5162261820 0.60569645
3  -1.9676628955 0.04910684
4  -0.8022647189 0.42239985
5   0.0007361161 0.99941266
```

Predicting status for Harry and Sally

- Harry is age 65, White, non-Hispanic, with Medicare insurance, a BMI of 28 and SBP of 135.
- Sally is age 60, Black, Hispanic, with Medicaid insurance, a BMI of 22 and SBP of 148.

```
newdat_s2 <- data_frame(subject = c("Harry", "Sally"),
                        age = c(65, 60),
                        race = c("White", "Black"),
                        hispanic = c(0, 1),
                        insurance = c("Medicare", "Medicaid"),
                        bmi = c(28, 22),
                        sbp = c(135, 148))
```

Predicting Harry and Sally's status

```r
predict(m_scr2_logistic, newdata = newdat_s2,
        type = "response")
```

```
        1         2 
0.5904364 0.4215335 
```

The prediction for Harry is 0.59, and for Sally, 0.42, by this logistic regression model.

A probit regression model

Now, consider a probit regression, fit by changing the default link for the binomial family as follows:

```r
m_scr2_probit <- glm(up_to_date ~ age + race + hispanic +
                         insurance + bmi + sbp,
                     family = binomial(link = "probit"),
                     data = colscr2)
```

tidy(m_scr2_probit)

```
                term     estimate   std.error  statistic
1        (Intercept)  1.584603569 1.658488821  0.9554503
2                age  0.013461338 0.024106778  0.5584047
3          raceOther -1.238445198 0.587092981 -2.1094533
4          raceWhite -0.199260184 0.243505258 -0.8182993
5           hispanic  0.029483051 0.484818945  0.0608125
6   insuranceMedicaid -0.619276718 0.293205189 -2.1120933
7   insuranceMedicare -0.322880519 0.333548759 -0.9680160
8  insuranceUninsured  0.052775722 0.463797571  0.1137904
9                bmi  0.009652339 0.012886845  0.7490071
10               sbp -0.014695526 0.005944435 -2.4721484
      p.value
1  0.33935005
2  0.57656807
3  0.03490548
4  0.41318630
5  0.95150854
```

Interpreting the Probit Model's Coefficients

```
      (Intercept)                 age           raceOther 
      1.584603569         0.013461338        -1.238445198 
        raceWhite            hispanic   insuranceMedicaid 
     -0.199260184         0.029483051        -0.619276718 
insuranceMedicare insuranceUninsured                  bmi 
     -0.322880519         0.052775722         0.009652339 
              sbp 
     -0.014695526 
```

The probit regression coefficients give the change in the z-score of the outcome of interest (here, `up_to_date`) for a one-unit change in the target predictor, holding all other predictors constant.

- So, for a one-year increase in age, holding all other predictors constant, the z-score for `up_to_date` increases by 0.013
- And for a Medicaid subject as compared to a Commercial subject of the same age, race, ethnicity, bmi and sbp, the z-score for the Medicaid subject is predicted to be -0.619 lower, according to this model.

What about Harry and Sally?

Do the predictions for Harry and Sally change much with this probit model, as compared to the logistic regression?

```
predict(m_scr2_probit, newdata = newdat_s2, type = "response")
```

```
        1         2 
0.5885511 0.4364027 
```

Project 1 Groups

Project 1 Groups

Group	Names
1	Laura Baldassari, Jenny Feng, Maher Kazimi, Satyakam Mishra, Vinh Trinh
2	Zainab (Albar) Albar, Dongze (Zaza) He, Nik Krieger, Andrew Shan
3	Andrew Tang, Sneha Vakamudi, Ruipeng Wei, Peter Wilkinson
4	Gwen Donley, Carli Lehr, Connor Swingle, Frances Wang
5	Ryan Honomichl, JJ Huang, Xin Xin Yu, Bilal Zonjy
6	Khaled Alayed, Kedar Mahajan, Preeti Pathak, Sarah Planchon Pope
7	Estee Cramer, Laura Cremer, Hyun Jo Kim, Roberto Martinez
8	Abhishek Deshpande, Jack McDonnell, Grace Park, Gabby Rieth
9	Haimeng Bai, Sophia Cao, Kate Dobbs, Elina Misicka
10	Vaishali (Vee) Deo, Caroline El Sanadi, Kaylee Sarna, Sandra Silva Camargo