432 Class 3 Slides

github.com/THOMASELOVE/432-2018

2018-01-23

Setup

```r
library(skimr)
library(simputation)
library(broom)
library(modelr)
library(tidyverse)

smartcle1 <- read.csv("data/smartcle1.csv")
```

Today's Materials

- A linear regression model using factors and quantities as predictors
- Single imputation via the `simputation` package
- Models including product terms
- Interpreting interactions, making predictions

These ideas come from Chapters 2-5, mostly.

Returning to the SMART BRFSS data (Notes Sections 2.8 - 2.11 and 5)

We're going to build `smartcle3`

We'll use a piece of the `smartcle1` data, and **simply impute** missing values.

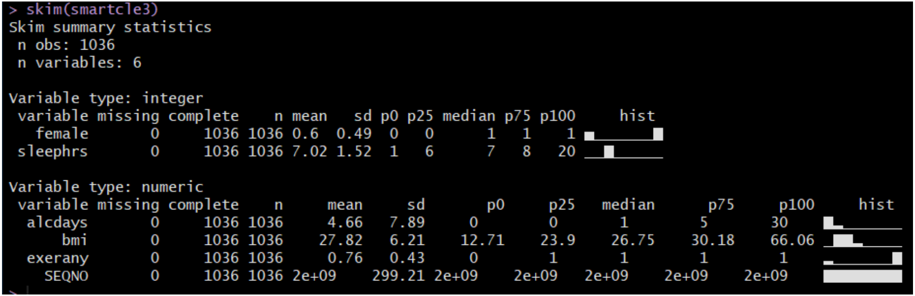
Variable	`NAs`	Description
`SEQNO`	0	respondent identification number (all begin with 2016)
`bmi`	84	Body mass index, in kg/m^2
`sleephrs`	8	On average, how many hours of sleep do you get in a 24-hour period?
`female`	0	Sex, 1 = female, 0 = male
`exerany`	3	Have you exercised at all in the past 30 days? (1 = yes, 0 = no)
`alcdays`	46	How many days during the past 30 days did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor?

smartcle3 development

```r
set.seed(20180123)

smartcle3 <- smartcle1 %>%
  select(SEQNO, bmi, sleephrs, female, alcdays, exerany) %>%
  impute_rhd(exerany ~ 1) %>%
  impute_pmm(sleephrs ~ 1) %>%
  impute_rlm(bmi ~ female + sleephrs) %>%
  impute_cart(alcdays ~ .)

colSums(is.na(smartcle3))
```

```
   SEQNO      bmi sleephrs   female  alcdays  exerany
       0        0        0        0        0        0
```

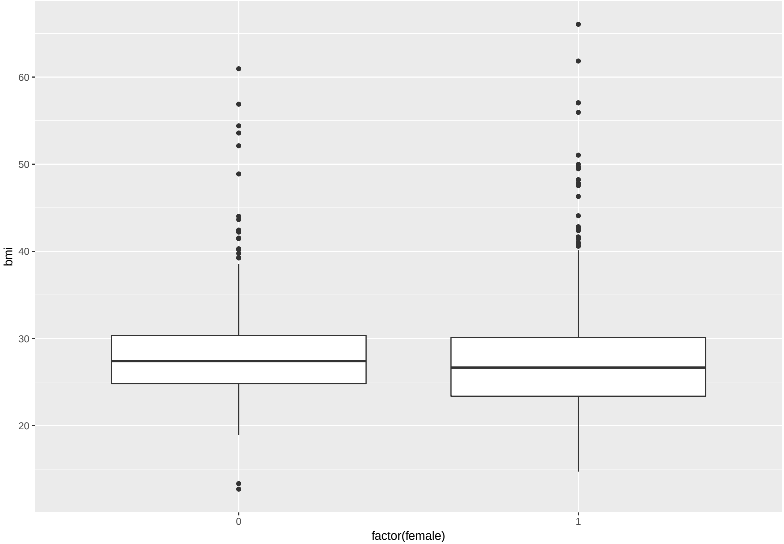
skim(smartcle3)



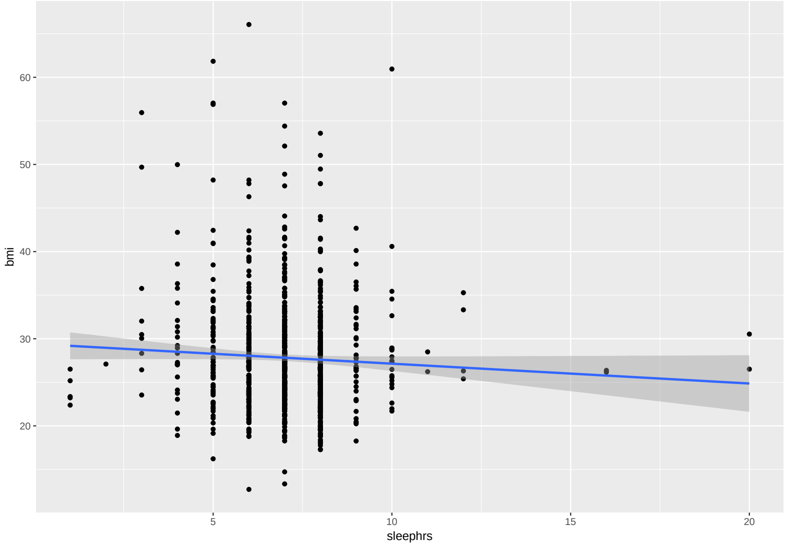
```
> skim(smartcle3)
Skim summary statistics
 n obs: 1036
 n variables: 6

Variable type: integer
 variable missing complete    n mean   sd p0 p25 median p75 p100     hist
   female       0     1036 1036 0.6  0.49  0   0      1   1    1
 sleephrs       0     1036 1036 7.02 1.52  1   6      7   8   20

Variable type: numeric
 variable missing complete    n    mean     sd     p0    p25  median    p75   p100     hist
  alcdays       0     1036 1036    4.66   7.89      0      0       1      5     30
      bmi       0     1036 1036   27.82   6.21  12.71   23.9   26.75  30.18  66.06
  exerany       0     1036 1036    0.76   0.43      0      1       1      1      1
    SEQNO       0     1036 1036 2e+09   299.21 2e+09  2e+09   2e+09   2e+09  2e+09
```

Plot, early and often

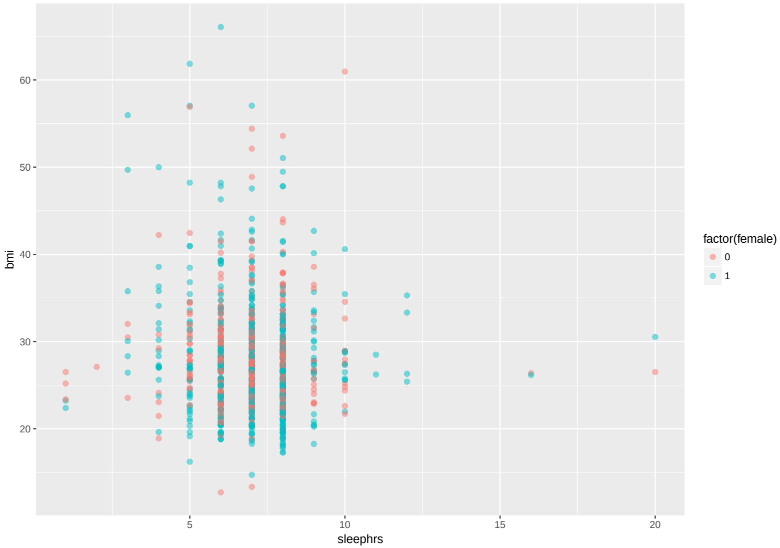
Using `female` to model `bmi`



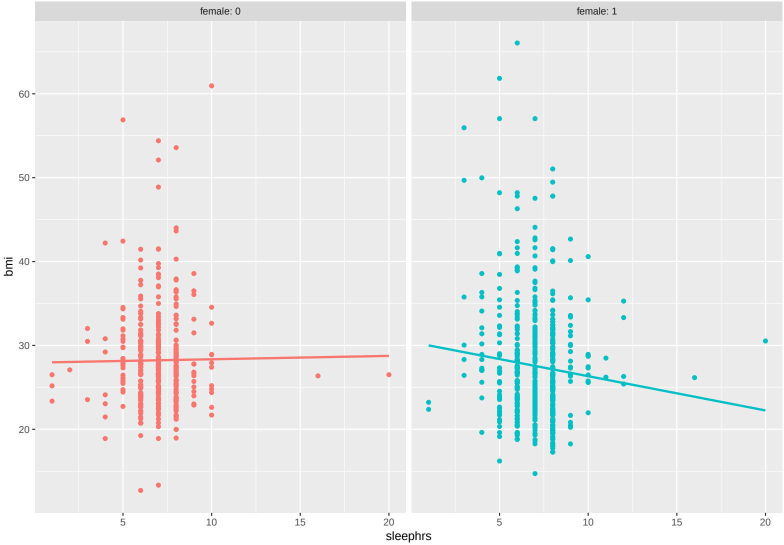
Using `sleephrs` to model `bmi`



Using `sleephrs` to model `bmi`, stratified by `female`



Using `female` and `sleephrs` and their interaction to model `bmi`



Incorporating a categorical-quantitative product term in a regression model (See Sections 2.11 - 2.12 and 4)

Building Two Models

We'll predict bmi using female and sleephrs

- and their interaction
- without their interaction

```
model_int <- lm(bmi ~ female * sleephrs, data = smartcle3)
model_noint <- lm(bmi ~ female + sleephrs, data = smartcle3)
```

Comparing Nested Models via glance

```r
glance(model_int) %>% round(., 3)
```

```
  r.squared adj.r.squared sigma statistic p.value df
1     0.009         0.006 6.191      3.08   0.027  4
     logLik      AIC      BIC deviance df.residual
1 -3356.783 6723.566 6748.281 39557.91        1032
```

```r
glance(model_noint) %>% round(., 3)
```

```
  r.squared adj.r.squared sigma statistic p.value df
1     0.006         0.004 6.197     3.087   0.046  3
     logLik      AIC      BIC deviance df.residual
1 -3358.313 6724.626 6744.398 39674.92        1033
```

ANOVA comparison for nested models

```r
anova(model_int, model_noint)
```

```
Analysis of Variance Table

Model 1: bmi ~ female * sleephrs
Model 2: bmi ~ female + sleephrs
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   1032 39558
2   1033 39675 -1   -117.01 3.0526 0.0809 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predictions with `model_int`

```
tidy(model_int)
```

```
             term    estimate std.error statistic
1     (Intercept) 27.94857162 1.4058797 19.879775
2          female  2.46850949 1.8408850  1.340936
3        sleephrs  0.04019189 0.1966903  0.204341
4 female:sleephrs -0.44856728 0.2567379 -1.747180
       p.value
1 1.038557e-74
2 1.802362e-01
3 8.381273e-01
4 8.090355e-02
```

Interpreting the Interaction model

With interaction, we have...

`bmi` = 27.95 + 2.47 `female` + 0.04 `sleephrs` - 0.45 `female` x `sleephrs`

- What is the predicted `bmi` for a male who sleeps 10 hours?
- What is the predicted `bmi` for a female who sleeps 10 hours?

Interpreting the Interaction model

`bmi` = 27.95 + 2.47 `female` + 0.04 `sleephrs` - 0.45 `female` x `sleephrs`

- so for males, our model is: `bmi` = 27.95 + 0.04 `sleephrs`, and
- for females, our model is: `bmi` = 30.42 - 0.41 `sleephrs`

Both the slope and the intercept of the `bmi-sleephrs` model **depend** on `sex`

Predictions with model_noint

```
tidy(model_noint)
```

```
         term   estimate std.error statistic       p.value
1 (Intercept) 29.7857897 0.9340801 31.887831 6.710149e-156
2      female -0.6737812 0.3931768 -1.713685  8.688661e-02
3    sleephrs -0.2230855 0.1265395 -1.762971  7.820099e-02
```

Interpreting the NO Interaction model

Without interaction, we have…

`bmi` = 29.79 - 0.67 `female` - 0.22 `sleephrs`

- Now, what is the predicted `bmi` for a male who sleeps 10 hours?
- What is the predicted `bmi` for a female who sleeps 10 hours?

Interpreting the NO Interaction model

`bmi` = 29.79 - 0.67 `female` - 0.22 `sleephrs`

- so for males, our model is: `bmi` = 29.79 - 0.22 `sleephrs`,
- and for females, our model is: `bmi` = 29.12 - 0.22 `sleephrs`

Only the **intercept** of the `bmi-sleephrs` model depends on `sex`

- Change in `bmi` per additional hour of sleep **does not depend** on sex

Next Time

- Centering and Rescaling Predictors
- Analysis of Variance
- Cross-validation of a linear model
- Sequential Variable Selection (Stepwise Regression)
- Forward Selection, Backward Elimination, Allen-Cady approaches
- Best Subsets Variable Selection
- Adjusted R^2, bias-corrected AIC, BIC and C_p

These ideas come from Chapters 2-8, mostly.