432 Class 3 Slides

github.com/THOMASELOVE/432-2018

2018-01-22

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
- Centering and Rescaling predictors
- Two-Way Analysis of Variance

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.

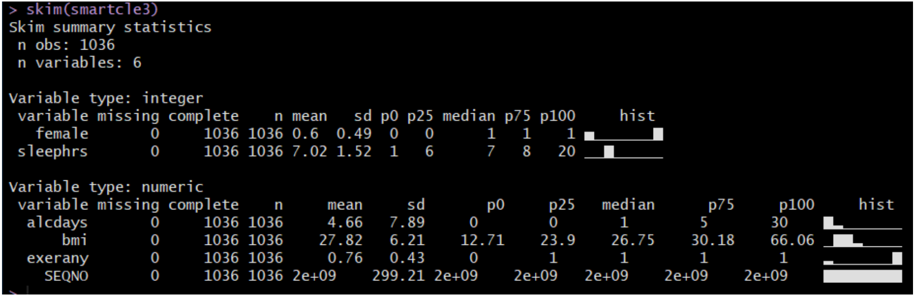
Variable	NAs	Description
SEQNO	0	respondent identification number (all begin with 2016)
bmi	84	Body mass index, in kg/m^2
sleephrs	8	On average, how many hours of sleep do you get in a 24-hour period?
female	0	Sex, 1 = female, 0 = male
exerany	6	Have you used the internet in the past 30 days? (1 = yes, 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

```
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 
```

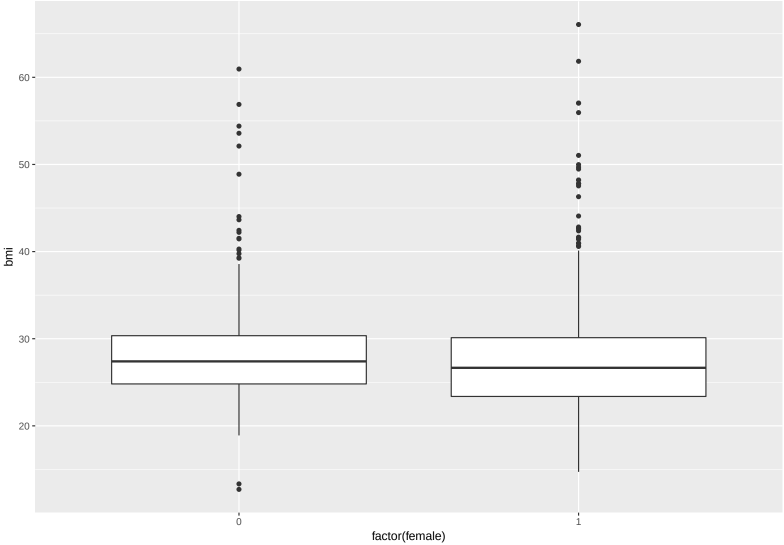
skim(smartcle3)



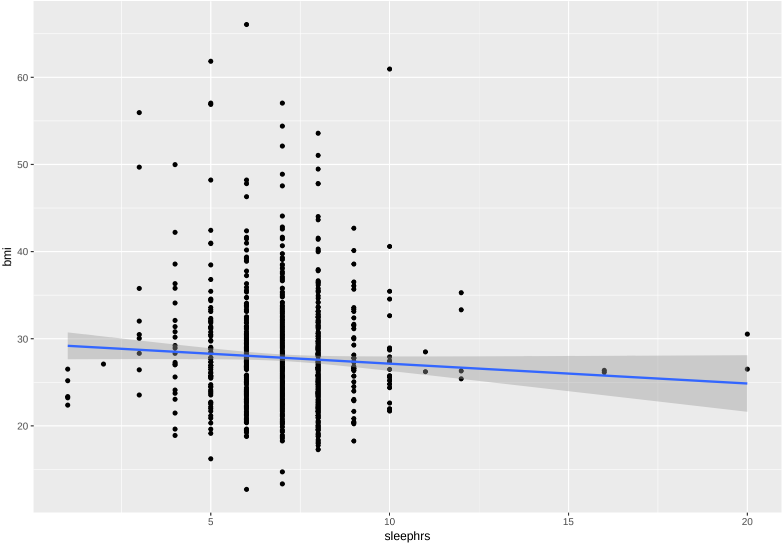
```
> skim(smartcle3)
Skim summary statistics
 n obs: 1036
 n variables: 6

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

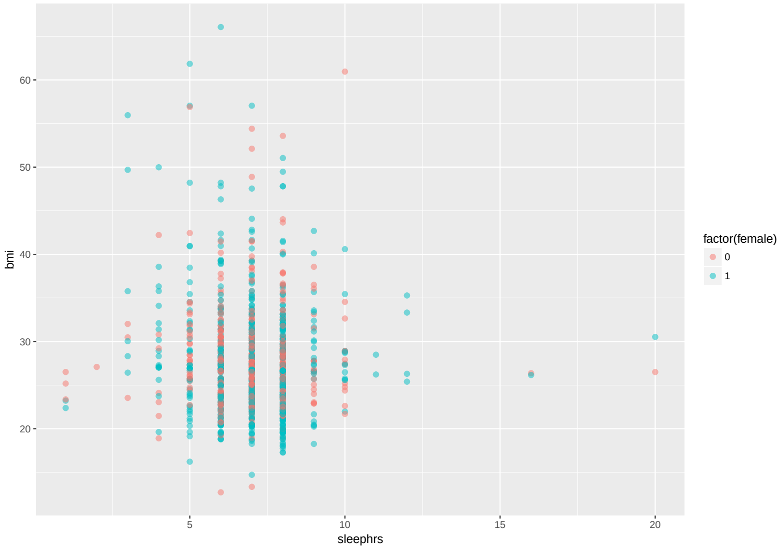
Using `female` to model `bmi`



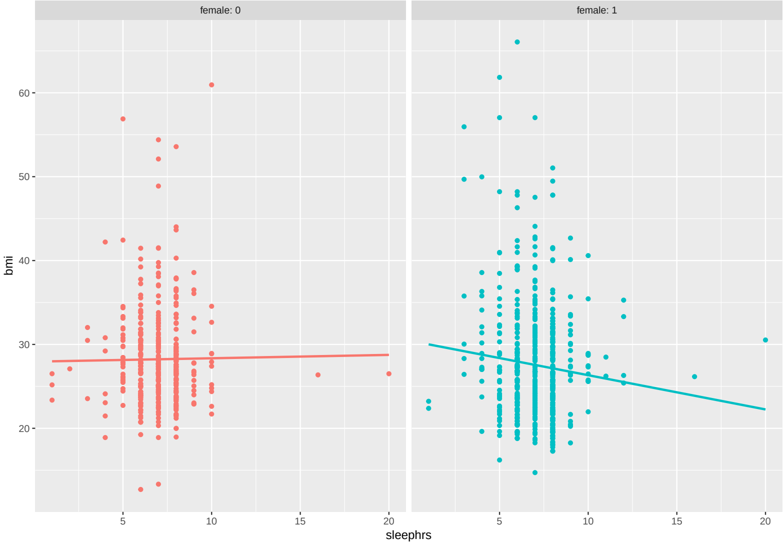
Using `sleephrs` to model `bmi`



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

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

Comparing Nested Models via glance

```
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
```

```
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

```
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

```r
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` = 25.48 - 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

Building Predictions for New Data (Individual Subjects)

What do we predict for the `bmi` of a female subject who gets 10 hours of sleep per night? What if the subject was male, instead?

```
new1 <- data_frame(female = c(1, 0), sleephrs = c(10,10))

predict(model_int, newdata = new1,
        interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 26.33333 14.13710 38.52955
2 28.35049 16.13121 40.56977
```

Building Predictions for New Data (Average Predictions)

What do we predict for the average `bmi` of a population of female subjects who sleep for 10 hours? What about the population of male subjects?

```
new1 <- data_frame(female = c(1, 0), sleephrs = c(10,10))

predict(model_int, newdata = new1,
        interval = "confidence", level = 0.95)
```

```
       fit      lwr      upr
1 26.33333 25.25921 27.40744
2 28.35049 27.04027 29.66071
```

Centering and Rescaling Predictors (See Notes sections 2.13, 2.14 and 4.7)

Centering sleephrs to ease interaction description

```
smartcle3 <- smartcle3 %>%
  mutate(sleep_c = sleephrs - mean(sleephrs))

model_int_c <- lm(bmi ~ female * sleep_c, data = smartcle3)
model_int_c
```

```

Call:
lm(formula = bmi ~ female * sleep_c, data = smartcle3)

Coefficients:
   (Intercept)          female         sleep_c
      28.23061        -0.67926         0.04019
female:sleep_c
      -0.44857
```

Interpreting Interaction: Centered sleephrs

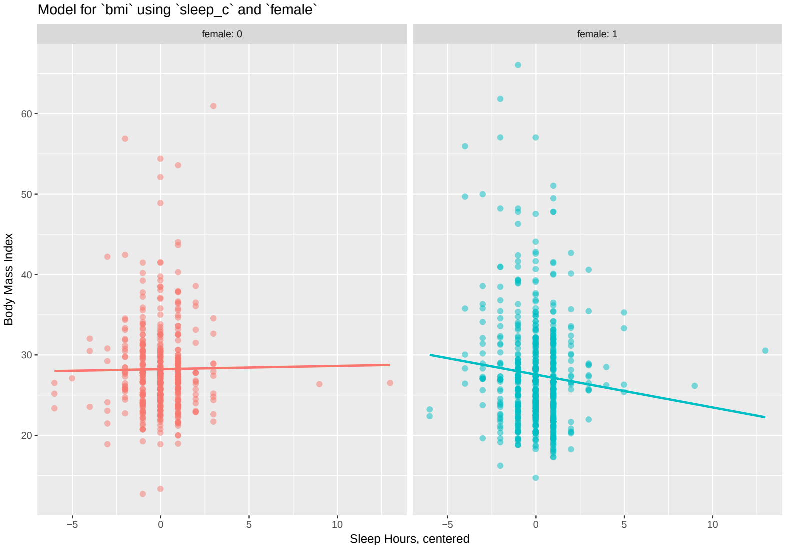
bmi = 28.23 - 0.68 female + 0.04 centered sleep_c - 0.45 female x centered sleep_c

- Now, 28.23 is the predicted bmi for a male who gets the average amount of sleep (7.02 hours)
- And 28.23 - 0.68 = 27.55 is the predicted bmi for a female who gets the average amount of sleep.
- So, the main effect of female is the predictive difference (female - male) in bmi for mean sleephrs,
- the product term is the change in the slope of centered sleephrs_c on bmi for a female rather than a male, and
- the residual standard deviation and the R-squared values remain unchanged from the model before centering.

```
glance(model_int_c) %>% round(., 3)
```

```
  r.squared adj.r.squared sigma statistic p.value df
1     0.009         0.006 6.191      3.08   0.027  4
```

Plotting bmi on centered sleep_c by female



Rescaling?

Centering helped us interpret the main effects in the regression, but it still leaves a scaling problem.

- The female coefficient estimate is much larger than that of sleephrs, but this is misleading, considering that we are comparing the complete change in one variable (sex = female or not) to a 1-hour change in average sleep.
- Gelman and Hill (2007) recommend all continuous predictors be scaled by dividing by 2 standard deviations
    - A 1-unit change in the rescaled predictor corresponds to a change from 1 standard deviation below the mean, to 1 standard deviation above.
    - An unscaled binary (1/0) predictor with 50% probability of occurring will be exactly comparable

Rescaling to sleep_z and re-fitting the model

```
smartcle3 <- smartcle3 %>%
    mutate(sleep_z = (sleephrs - mean(sleephrs)) /
             (2*sd(sleephrs)))

model_int_z <- lm(bmi ~ female * sleep_z, data = smartcle3)

model_int_z
```

```
Call:
lm(formula = bmi ~ female * sleep_z, data = smartcle3)

Coefficients:
   (Intercept)          female         sleep_z
       28.2306         -0.6793          0.1224
female:sleep_z
       -1.3660
```

Comparing our Interaction Models

Original Model

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

Centered Model

- bmi = 28.23 - 0.68 `female` + 0.04 `sleep_c` - 0.45 `female` x `sleep_c`

Centered, Rescaled Model

- bmi = 28.23 - 0.68 `female` + 0.12 `sleep_z` - 1.37 `female` x `sleep_z`

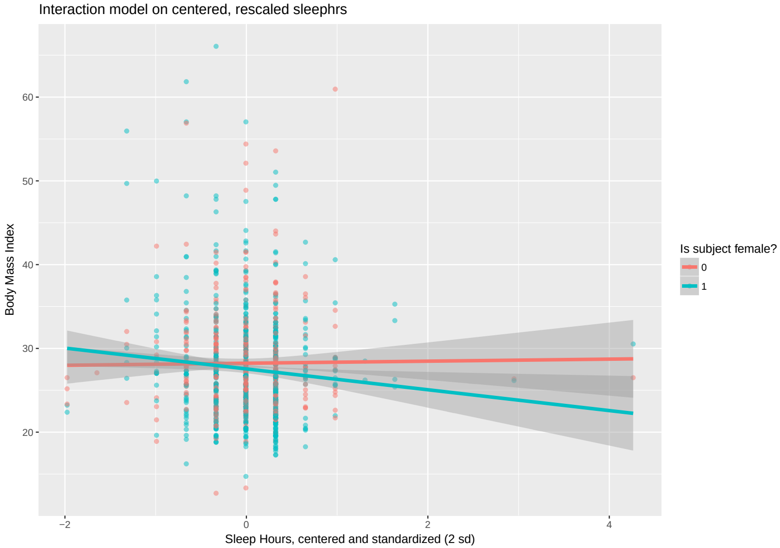
Interpreting the Centered, Rescaled Model

- Main effect of `female`, -0.68, is still the predictive difference (female - male) in `bmi` with `sleephrs` at its mean, 7.02 hours,
- Intercept (28.23) is still the predicted `bmi` for a male who sleeps the mean number of hours, and
- the residual standard deviation and the R-squared values remain unchanged

but now we also have:

- the coefficient of `sleep_z` is the predictive difference in bmi associated with a change in `sleephrs` of 2 standard deviations (from one standard deviation below the mean of 7.02 to one standard deviation above 7.02.)
  - Since sd(sleephrs) is 1.52, this corresponds to a change from 5.50 hours per night to 8.54 hours per night.
- the coefficient of the product term (-1.37) corresponds to the change in the coefficient of `sleep_z` for females as compared to males.

Plotting the Rescaled, Centered Model



Two-Factor Analysis of Variance (see Notes Chapter 3)

How do female and exerany relate to bmi?

```
smart3_sum <- smartcle3 %>%
  group_by(female, exerany) %>%
  summarize(mean.bmi = mean(bmi), sd.bmi = sd(bmi))
```

Resulting tibble for smart3_sum

```
smart3_sum
```

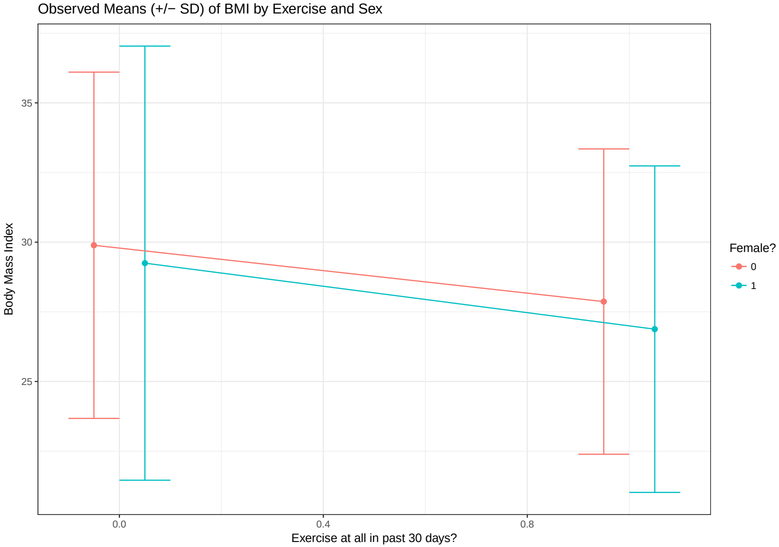
```
# A tibble: 4 x 4
# Groups: female [?]
  female exerany mean.bmi sd.bmi
   <int>   <dbl>    <dbl>  <dbl>
1      0    0        29.9   6.21
2      0    1.00     27.9   5.48
3      1    0        29.2   7.79
4      1    1.00     26.9   5.86
```

This would be more useful as a plot.

Building a Means Plot (result on next slide)

```
pd <- position_dodge(0.2)

ggplot(smart3_sum, aes(x = exerany, y = mean.bmi, col = factor
  geom_errorbar(aes(ymin = mean.bmi - sd.bmi,
                    ymax = mean.bmi + sd.bmi),
                width = 0.2, position = pd) +
  geom_point(size = 2, position = pd) +
  geom_line(aes(group = female), position = pd) +
  scale_color_discrete(name = "Female?") +
  theme_bw() +
  labs(y = "Body Mass Index", x = "Exercise at all in past 30
       title = "Observed Means (+/- SD) of BMI by Exercise and
```

Means Plot (Do we have a strong interaction effect?)



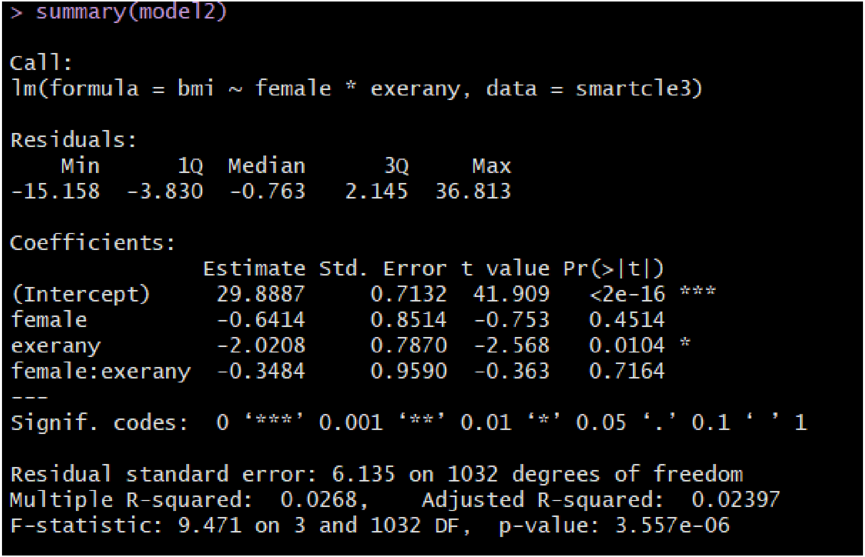
Two-Way ANOVA model with Interaction

```
model2 <- lm(bmi ~ female * exerany, data = smartcle3)

anova(model2)
```

```
Analysis of Variance Table

Response: bmi
                 Df Sum Sq Mean Sq F value    Pr(>F)    
female            1    118  117.76  3.1288   0.07722 .  
exerany           1    947  946.71 25.1530 6.231e-07 ***
female:exerany    1      5    4.97  0.1320   0.71642    
Residuals      1032  38843   37.64                      
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Does it seem like we need the interaction term in this case?

Summary of Two-Factor ANOVA with Interaction



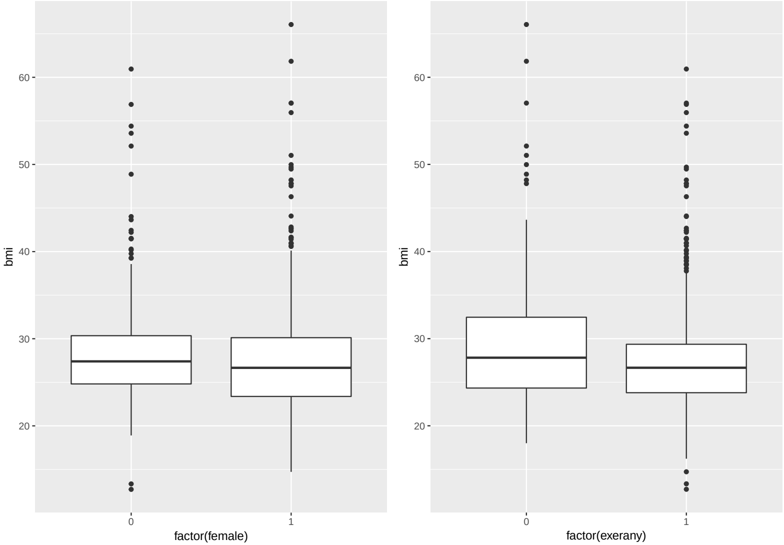
```
> summary(model2)

Call:
lm(formula = bmi ~ female * exerany, data = smartcle3)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.158  -3.830  -0.763   2.145  36.813 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     29.8887     0.7132  41.909   <2e-16 ***
female          -0.6414     0.8514  -0.753   0.4514    
exerany         -2.0208     0.7870  -2.568   0.0104 *  
female:exerany  -0.3484     0.9590  -0.363   0.7164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.135 on 1032 degrees of freedom
Multiple R-squared:  0.0268,	Adjusted R-squared:  0.02397 
F-statistic: 9.471 on 3 and 1032 DF,  p-value: 3.557e-06
```

What if we wanted the model with no interaction?

Here's the key plot, then…

```r
p1 <- ggplot(smartcle3, aes(x = factor(female), y = bmi)) +
    geom_boxplot()
p2 <- ggplot(smartcle3, aes(x = factor(exerany), y = bmi)) +
    geom_boxplot()

gridExtra::grid.arrange(p1, p2, nrow = 1)
```

Key Plot for Two-Way ANOVA, no interaction

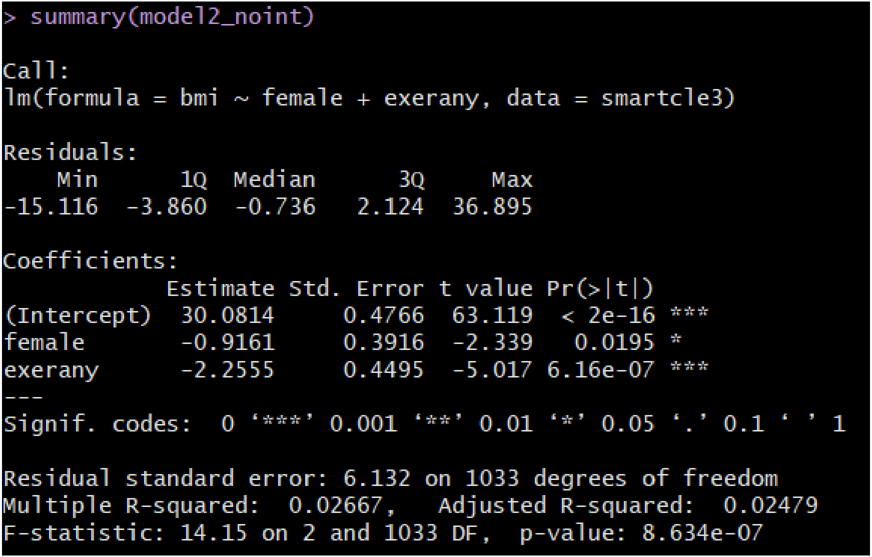


Two-Way ANOVA model without Interaction

```r
model2_noint <- lm(bmi ~ female + exerany, data = smartcle3)

anova(model2_noint)
```

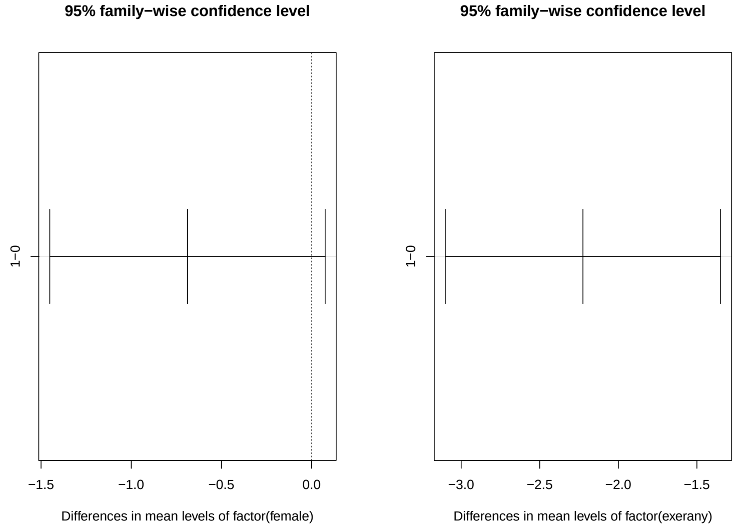
```
Analysis of Variance Table

Response: bmi
            Df Sum Sq Mean Sq F value    Pr(>F)    
female       1    118  117.76  3.1314   0.07709 .  
exerany      1    947  946.71 25.1742 6.164e-07 ***
Residuals 1033  38848   37.61                      
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary of Two-Factor No Interaction ANOVA



```
> summary(model2_noint)

Call:
lm(formula = bmi ~ female + exerany, data = smartcle3)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.116  -3.860  -0.736   2.124  36.895 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  30.0814     0.4766  63.119  < 2e-16 ***
female       -0.9161     0.3916  -2.339   0.0195 *  
exerany      -2.2555     0.4495  -5.017 6.16e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.132 on 1033 degrees of freedom
Multiple R-squared:  0.02667,	Adjusted R-squared:  0.02479 
F-statistic: 14.15 on 2 and 1033 DF,  p-value: 8.634e-07
```

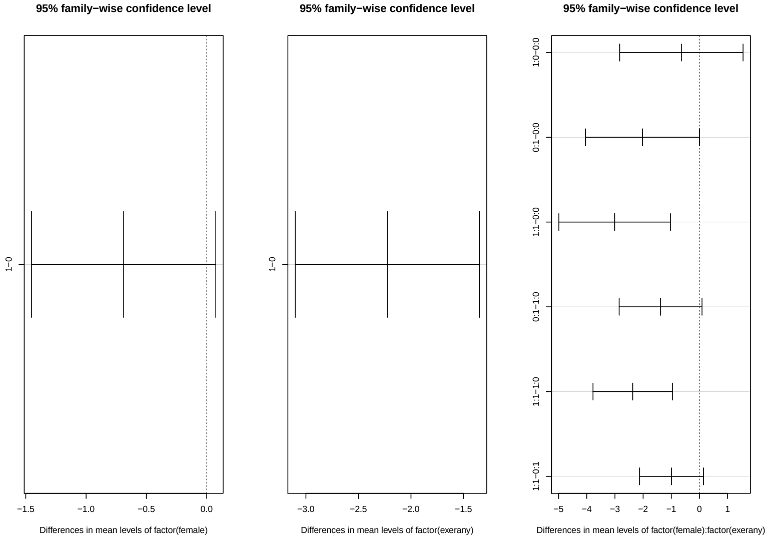
Tukey HSD Comparisons (no interaction)



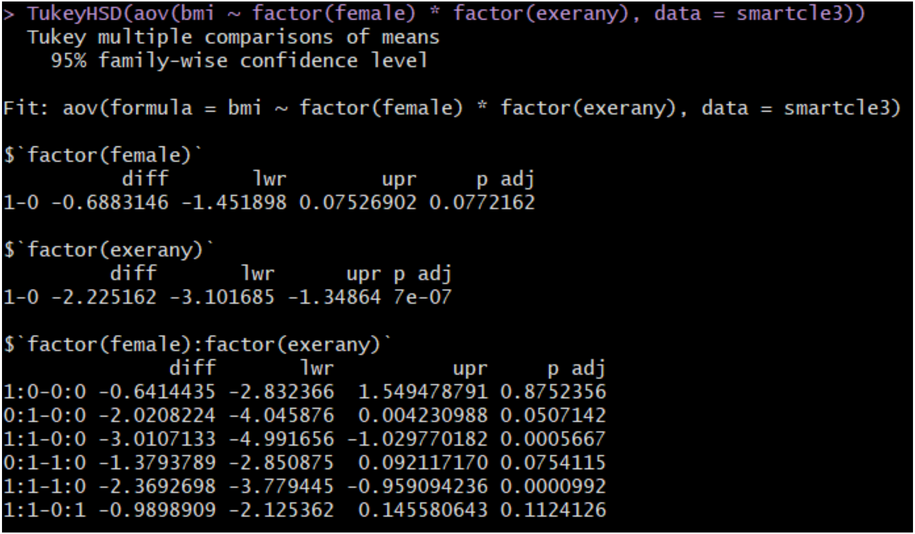
Tukey HSD Comparisons (without interaction)

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bmi ~ factor(female) + factor(exerany), dat

$`factor(female)`
          diff       lwr        upr     p adj
1-0 -0.6883146 -1.451577 0.07494728 0.0770918

$`factor(exerany)`
         diff       lwr       upr p adj
1-0 -2.225162 -3.101315 -1.349009 7e-07
```

Tukey HSD comparisons WITH interaction



Tukey HSD comparisons WITH interaction



```
> TukeyHSD(aov(bmi ~ factor(female) * factor(exerany), data = smartcle3))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bmi ~ factor(female) * factor(exerany), data = smartcle3)

$`factor(female)`
          diff       lwr        upr     p adj
1-0 -0.6883146 -1.451898 0.07526902 0.0772162

$`factor(exerany)`
         diff       lwr      upr p adj
1-0 -2.225162 -3.101685 -1.34864 7e-07

$`factor(female):factor(exerany)`
              diff       lwr          upr     p adj
1:0-0:0 -0.6414435 -2.832366  1.549478791 0.8752356
0:1-0:0 -2.0208224 -4.045876  0.004230988 0.0507142
1:1-0:0 -3.0107133 -4.991656 -1.029770182 0.0005667
0:1-1:0 -1.3793789 -2.850875  0.092117170 0.0754115
1:1-1:0 -2.3692698 -3.779445 -0.959094236 0.0000992
1:1-0:1 -0.9898909 -2.125362  0.145580643 0.1124126
```

Next Time

- Building a linear regression model
- 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 6-8, mostly.