

432 Class 4 Slides

github.com/THOMASELOVE/2019-432

2019-01-31

SMART BRFSS 2017 data: A New Pull

```
library(skimr); library(broom); library(janitor)
library(simputation); library(tidyverse)

smart_oh_2017 <- readRDS("data/smart_2017_oh.rds")

smart2_raw <- smart_oh_2017 %>%
  mutate(personcode = as.character(1:nrow(smart_oh_2017))) %>%
  select(personcode, genhealth, alcdays, female,
         bmi, height_m, weight_kg, exerany,
         seatbelt = seatbelt_always, mmsaname)
```

Missingness?

```
colSums(is.na(smart2_raw))
```

personcode	genhealth	alcdays	female	bmi
0	11	46	0	323
height_m	weight_kg	exerany	seatbelt	mmsaname
84	296	13	26	0

Simple Imputation and Re-calculating of BMI

```
set.seed(20190131)
```

```
smart2 <- smart2_raw %>%  
  impute_cart(seatbelt ~ mmsaname) %>%  
  impute_pmm(exerany ~ mmsaname) %>%  
  impute_pmm(height_m ~ exerany + female) %>%  
  impute_pmm(weight_kg ~ exerany + female) %>%  
  impute_cart(genhealth ~ mmsaname + weight_kg) %>%  
  impute_pmm(alcdays ~ mmsaname + female) %>%  
  mutate(bmi = weight_kg / (height_m^2))  
  
colSums(is.na(smart2))
```

personcode	genhealth	alcdays	female	bmi
0	0	0	0	0
height_m	weight_kg	exerany	seatbelt	mmsaname
0	0	0	0	0

Saving as an R data set

```
saveRDS(smart2, "data/smart2.rds")
```

Now, we could have started with ...

```
smart2 <- readRDS("data/smart2.rds")
```

and ignored everything except for the package loading.

Using mosaic::inspect

```
mosaic::inspect(smart2)
```

categorical variables:

	name	class	levels	n	missing
1	personcode	character	6277	6277	0
2	genhealth	factor	5	6277	0
3	seatbelt	factor	2	6277	0
4	mmsaname	character	6	6277	0

distribution

1	1 (0%), 10 (0%), 100 (0%) ...
2	2_VeryGood (32.8%), 3_Good (31.7%) ...
3	Yes (88.2%), No (11.8%)
4	(%) ...

quantitative variables:

name	class	min	Q1	median	Q3
------	-------	-----	----	--------	----

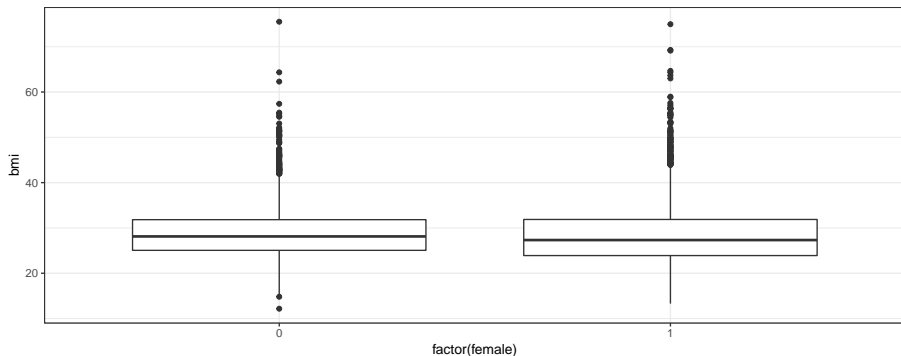
mosaic::inspect(smart2)

name <chr>	class <chr>	levels <int>	n <int>	missing <int>	distribution <chr>
personcode	character	6277	6277	0	1 (0%), 10 (0%), 100 (0%) ...
genhealth	factor	5	6277	0	2_VeryGood (32.8%), 3_Good (31.7%) ...
seatbelt	factor	2	6277	0	Yes (88.2%), No (11.8%)
mmsaname	character	6	6277	0	(%) ...
bmigroup	factor	4	6277	0	[25.0,30.0) (36.9%) ...

name <chr>	class <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
alcdays	numeric	0.00000	0.0000	0.00000	4.00000	30.00000	4.3506452	7.6882082	6277
female	numeric	0.00000	0.0000	1.00000	1.00000	1.00000	0.5830811	0.4930885	6277
bmi	numeric	12.17018	24.3372	27.60355	31.82532	75.52133	28.7198241	6.5132396	6277
height_m	numeric	1.35000	1.6300	1.68000	1.78000	2.06000	1.6942345	0.1040645	6277
weight_kg	numeric	31.75000	68.0400	79.38000	92.99000	208.65000	82.7418225	21.2671513	6277
exerany	numeric	0.00000	0.0000	1.00000	1.00000	1.00000	0.6899793	0.4625386	6277

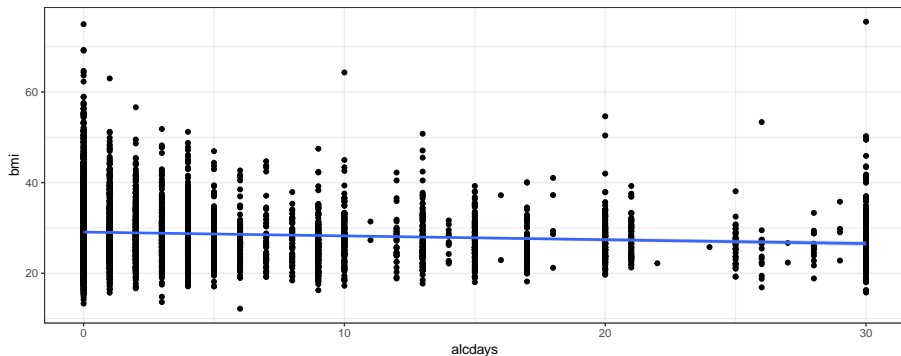
Using female to model bmi

```
ggplot(smart2, aes(x = factor(female), y = bmi)) +  
  geom_boxplot() + theme_bw()
```



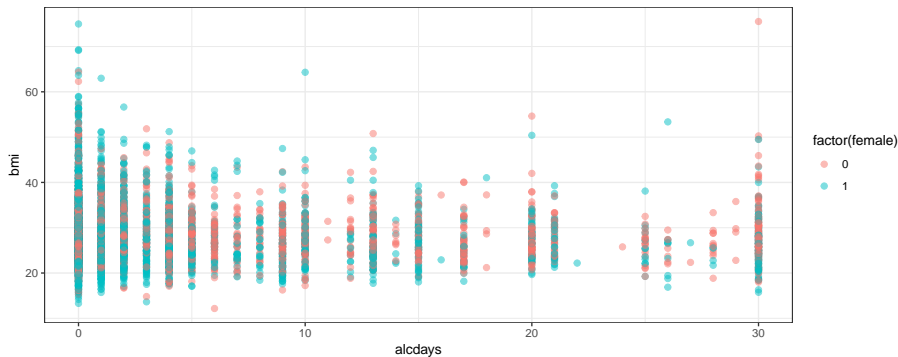
Using alcdays to model bmi

```
ggplot(smart2, aes(x = alcdays, y = bmi)) +  
  geom_point() + geom_smooth(method = "lm") + theme_bw()
```



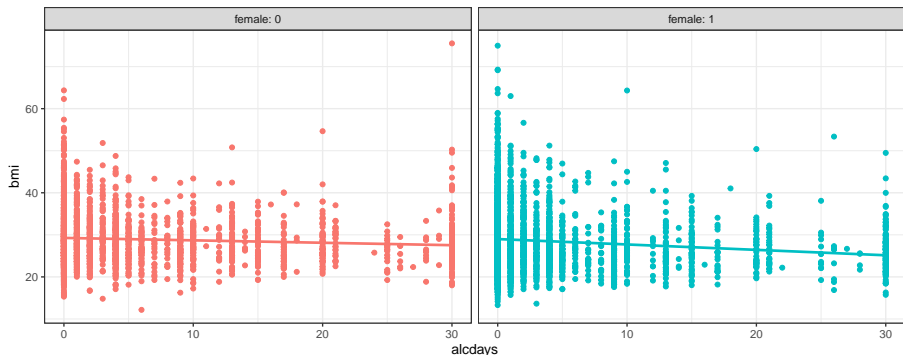
Using alcdays to model bmi, stratified by female

```
ggplot(smart2, aes(x = alcdays, y = bmi,  
                   color = factor(female))) +  
  geom_point(alpha = 0.5, size = 2) + theme_bw()
```



alcdays, female and interaction to model bmi

```
ggplot(smart2, aes(x = alcdays, y = bmi,  
                   color = factor(female))) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE) +  
  guides(col = FALSE) + theme_bw() +  
  facet_wrap(~ female, labeller = label_both)
```



Building Two Models

We'll predict bmi using female and alcdays...

- and their interaction

```
model_2i <- lm(bmi ~ female * alcdays, data = smart2)
```

- without their interaction

```
model_2no <- lm(bmi ~ female + alcdays, data = smart2)
```

ANOVA comparison for Nested Models

```
anova(model_2i, model_2no)
```

Analysis of Variance Table

Model 1: bmi ~ female * alcdays

Model 2: bmi ~ female + alcdays

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	6273	262619				
2	6274	263070	-1	-451.24	10.778	0.001032 **

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comparing Nested Models via glance

```
glance(model_2i) %>% round(., 2) %>% print.data.frame
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df
1	0.01	0.01	6.47	28.85	0	4
	logLik	AIC	BIC	deviance	df.residual	
1	-20625.25	41260.49	41294.21	262618.9		6273

```
glance(model_2no) %>% round(., 2) %>% print.data.frame
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df
1	0.01	0.01	6.48	37.83	0	3
	logLik	AIC	BIC	deviance	df.residual	
1	-20630.63	41269.27	41296.25	263070.1		6274

Predictions with model_2i

```
tidy(model_2i) %>% print.data.frame
```

	term	estimate	std.error	statistic
1	(Intercept)	29.27018647	0.15145149	193.264430
2	female	-0.27412942	0.19301062	-1.420282
3	alcdays	-0.05764733	0.01461732	-3.943768
4	female:alcdays	-0.07076169	0.02155357	-3.283062

	p.value
1	0.000000e+00
2	1.555754e-01
3	8.108538e-05
4	1.032485e-03

Interpreting the Interaction Model

With interaction, the model is

$$\text{bmi} = 29.26 - 0.27 \text{ female} - 0.06 \text{ alcdays} - 0.07 \text{ female} \times \text{alcdays}$$

- 1 What is the predicted bmi for a male who used alcohol on 10 of the last 30 days?
- 2 What is the predicted bmi for a female who used alcohol on 10 of the last 30 days?

Interpreting the Interaction Model

With interaction, the model is

$$\text{bmi} = 29.26 - 0.27 \text{ female} - 0.06 \text{ alcdays} - 0.07 \text{ female} \times \text{alcdays}$$

So, for males, the model is:

$$\text{bmi} = 29.26 - 0.06 \text{ alcdays}$$

And, for females, the model is:

$$\text{bmi} = (29.26 - 0.27) + (-0.06 - 0.07) \text{ alcdays, or } 28.99 - 0.13 \text{ alcdays}$$

Both the slope and the intercept of the bmi - alcdays model depend on female.

Predictions with the Main Effects Model

```
tidy(model_2no)
```

```
# A tibble: 3 x 5
```

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	29.5	0.141	209.	0.
2	female	-0.589	0.168	-3.51	4.44e- 4
3	alcdays	-0.0902	0.0108	-8.39	5.96e-17

$\text{bmi} = 29.46 - 0.59 \text{ female} - 0.09 \text{ alcdays}$

- 1 What is the predicted bmi for a male who used alcohol on 10 of the last 30 days?
- 2 What is the predicted bmi for a female who used alcohol on 10 of the last 30 days?

Interpreting the Main Effects Model

Without the interaction, the model is

$$\text{bmi} = 29.46 - 0.59 \text{ female} - 0.09 \text{ alcdays}$$

So, for males, the model is:

$$\text{bmi} = 9.46 - 0.59 \text{ female} - 0.09 \text{ alcdays}$$

And, for females, the model is:

$$\text{bmi} = (29.46 - 0.59) - 0.09 \text{ alcdays, or } 28.87 - 0.09 \text{ alcdays}$$

Only the intercept of the `bmi - alcdays` model depends on `female`.

- The change in `bmi` per additional day of alcohol use does not depend on `sex`.

What if we had a multi-categorical factor?

Suppose we want to study the impact of both exerany and genhealth on BMI.

```
smart2 %>% count(genhealth)
```

```
# A tibble: 5 x 2
  genhealth      n
  <fct>      <int>
1 1_Excellent    872
2 2_VeryGood   2057
3 3_Good        1987
4 4_Fair         991
5 5_Poor         370
```

Does it seem like we need to collapse any levels here?

Collapsing?

```
smart2 %>% count(genhealth, exerany)
```

```
# A tibble: 10 x 3
  genhealth    exerany      n
  <fct>        <dbl> <int>
1 1_Excellent      0    124
2 1_Excellent      1    748
3 2_VeryGood       0    474
4 2_VeryGood       1   1583
5 3_Good           0    651
6 3_Good           1   1336
7 4_Fair           0    464
8 4_Fair           1    527
9 5_Poor           0    233
10 5_Poor          1    137
```

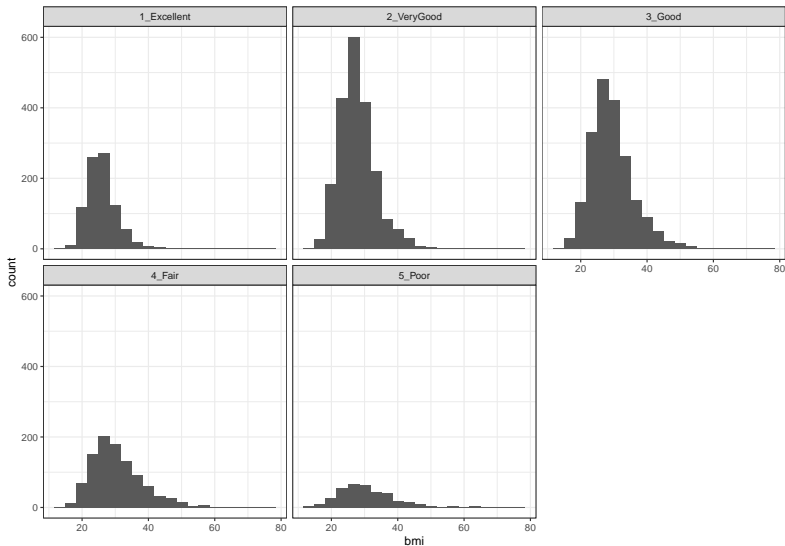
Now does it seem like we need to collapse any levels here?

Cross-Tabulation?

```
smart2 %>% tabyl(genhealth, exerany)
```

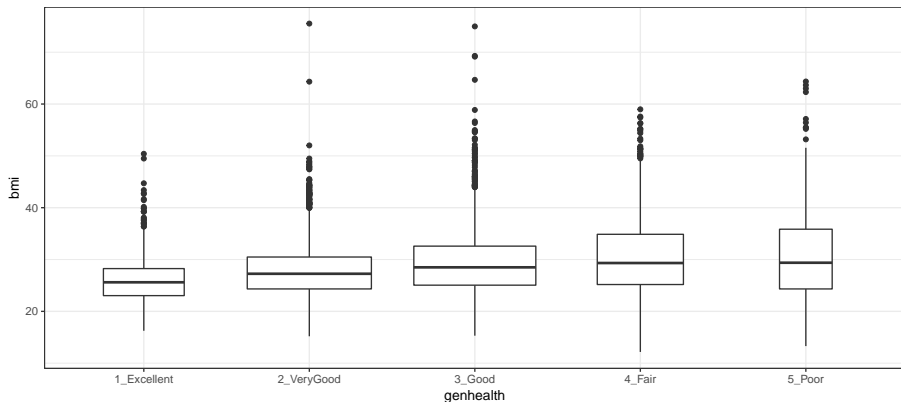
genhealth	0	1
1_Excellent	124	748
2_VeryGood	474	1583
3_Good	651	1336
4_Fair	464	527
5_Poor	233	137

Distribution of bmi by genhealth?



Boxplots with variable widths?

```
ggplot(smart2, aes(x = genhealth, y = bmi)) +  
  geom_boxplot(varwidth = TRUE) + theme_bw()
```



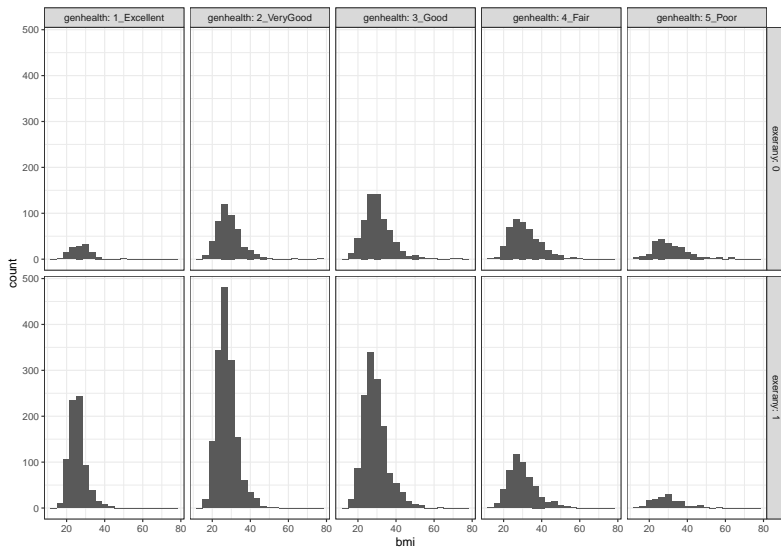
Summary Statistics?

```
mosaic::favstats(bmi ~ genhealth, data = smart2)
```

	genhealth	min	Q1	median	Q3	max
1	1_Excellent	16.25310	23.03495	25.60879	28.25260	50.40000
2	2_VeryGood	15.19440	24.32872	27.25224	30.48356	75.52133
3	3_Good	15.30622	25.05365	28.49379	32.58449	74.97521
4	4_Fair	12.17018	25.17792	29.32571	34.86687	58.97888
5	5_Poor	13.29938	24.32673	29.38125	35.84917	64.34948

	mean	sd	n	missing
1	26.04942	4.496603	872	0
2	27.83312	5.442950	2057	0
3	29.50860	6.659790	1987	0
4	30.57877	7.581389	991	0
5	30.72796	8.826062	370	0

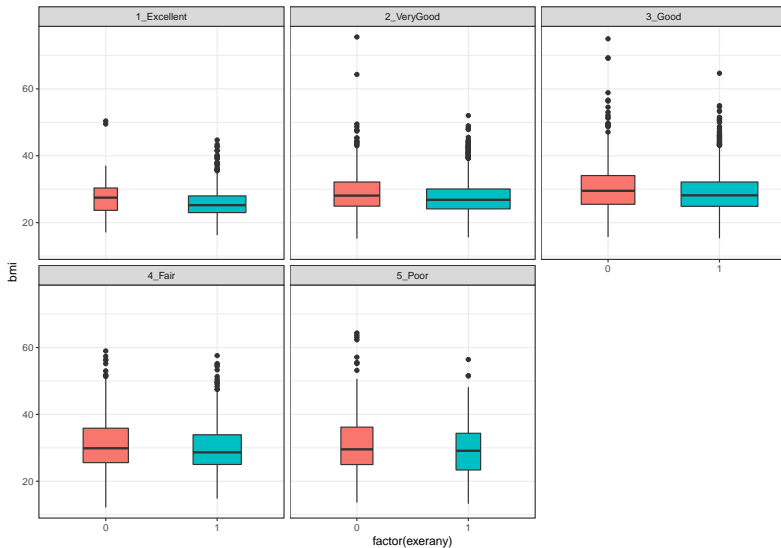
bmi by genhealth and exerany?



Code for Previous Slide

```
ggplot(smart2, aes(x = bmi)) +  
  geom_histogram(bins = 20) +  
  theme_bw() +  
  facet_grid(exerany ~ genhealth, labeller = "label_both")
```

Boxplots instead?



Code for previous plot

```
ggplot(smart2, aes(x = factor(exerany), y = bmi)) +  
  geom_boxplot(aes(fill = factor(exerany)),  
               varwidth = TRUE) +  
  theme_bw() + guides(fill = FALSE) +  
  facet_wrap(~ genhealth)
```

Can we use favstats for two factors at once?

```
mosaic::favstats(bmi ~ genhealth + exerany,  
                  data = smart2)[c("genhealth.exerany",  
                                   "mean", "sd", "n", "missing")]
```

	genhealth.exerany	mean	sd	n	missing
1	1_Excellent.0	27.64487	5.320510	124	0
2	2_VeryGood.0	28.94985	6.479965	474	0
3	3_Good.0	30.33703	7.438949	651	0
4	4_Fair.0	31.26841	7.859391	464	0
5	5_Poor.0	31.18231	9.226123	233	0
6	1_Excellent.1	25.78493	4.292097	748	0
7	2_VeryGood.1	27.49874	5.046002	1583	0
8	3_Good.1	29.10493	6.208035	1336	0
9	4_Fair.1	29.97157	7.281440	527	0
10	5_Poor.1	29.95524	8.074475	137	0

Table of Means and Standard Deviations

```
smart2 %>% group_by(genhealth, exerany) %>%  
  summarize(mean.bmi = mean(bmi), sd.bmi = sd(bmi))
```

```
# A tibble: 10 x 4
```

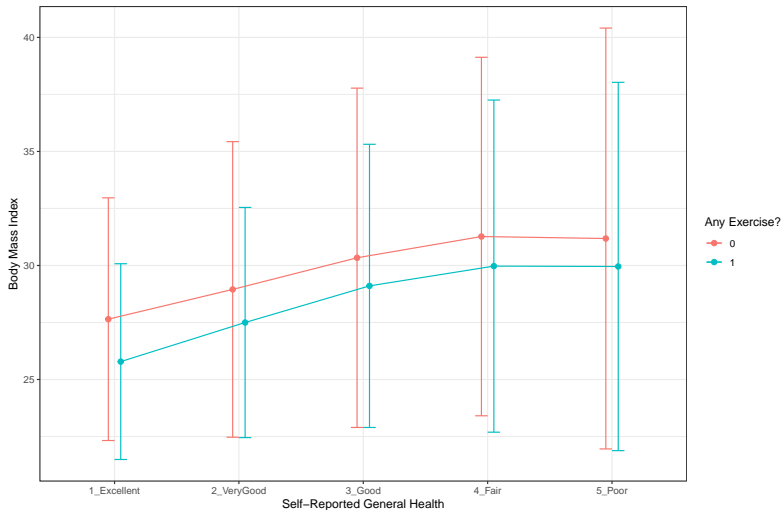
```
# Groups:   genhealth [?]
```

	genhealth	exerany	mean.bmi	sd.bmi
	<fct>	<dbl>	<dbl>	<dbl>
1	1_Excellent	0	27.6	5.32
2	1_Excellent	1	25.8	4.29
3	2_VeryGood	0	28.9	6.48
4	2_VeryGood	1	27.5	5.05
5	3_Good	0	30.3	7.44
6	3_Good	1	29.1	6.21
7	4_Fair	0	31.3	7.86
8	4_Fair	1	30.0	7.28
9	5_Poor	0	31.2	9.23
10	5_Poor	1	30.0	8.07

Check interaction first with means plot?

BMI by General Health, Exercise

Means \pm Standard Deviations



Means Plot code

```
pd <- position_dodge(0.2)
smart_sum <- smart2 %>%
  group_by(genhealth, exerany) %>%
  summarize(mean.bmi = mean(bmi), sd.bmi = sd(bmi))
ggplot(smart_sum, aes(x = genhealth, y = mean.bmi,
                      col = factor(exerany))) +
  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 = factor(exerany)), position = pd) +
  scale_color_discrete(name = "Any Exercise?") +
  theme_bw() +
  labs(y = "Body Mass Index",
       x = "Self-Reported General Health",
       title = "BMI by General Health, Exercise",
       subtitle = "Means +/- Standard Deviations")
```

ANOVA with and without interaction term

```
model_3no <- lm(bmi ~ genhealth + exerany, data = smart2)
model_3i <- lm(bmi ~ genhealth * exerany, data = smart2)

anova(model_3i)
```

Analysis of Variance Table

Response: bmi

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	13988	3497.1	87.6981	< 2.2e-16 ***
exerany	1	2306	2306.2	57.8321	3.274e-14 ***
genhealth:exerany	4	39	9.8	0.2462	0.9121
Residuals	6267	249908	39.9		

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Does the interaction have a meaningful impact?

- Means plot is essentially parallel: no clear interaction.
- $SS(\text{interaction}) = 39$, $SS(\text{Total}) = 249908$, so $\eta^2 = .00016$ or 0.016%
- p value for interaction term is 0.91

What does this imply about which model might be more helpful?

Making Predictions with model_3no

- Anna exercises and is in very good health.
- Brad doesn't exercise and is in poor health.

```
round(coef(model_3no),2)
```

(Intercept)	genhealth2_VeryGood	genhealth3_Good
27.22	1.66	3.21
genhealth4_Fair	genhealth5_Poor	exerany
4.09	4.01	-1.36

Making Predictions with model_3i

- Anna exercises and is in very good health.
- Brad doesn't exercise and is in poor health.

```
round(coef(model_3i),2)
```

(Intercept)	genhealth2_VeryGood
27.64	1.30
genhealth3_Good	genhealth4_Fair
2.69	3.62
genhealth5_Poor	exerany
3.54	-1.86
genhealth2_VeryGood:exerany	genhealth3_Good:exerany
0.41	0.63
genhealth4_Fair:exerany	genhealth5_Poor:exerany
0.56	0.63

Predictions

```
newpeople <- tibble(  
  name = c("Anna", "Brad"),  
  genhealth = c("2_VeryGood", "5_Poor"),  
  exerany = c(1, 0))  
  
predict(model_3no, newdata = newpeople)
```

1	2
27.51911	31.23254

```
predict(model_3i, newdata = newpeople)
```

1	2
27.49874	31.18231

What if we add in alcdays?

```
model4 <- lm(bmi ~ alcdays + genhealth * exerany,  
             data = smart2)
```

```
anova(model4)
```

Analysis of Variance Table

Response: bmi

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
alcdays	1	2654	2654.4	66.8526	3.511e-16	***
genhealth	4	12655	3163.7	79.6799	< 2.2e-16	***
exerany	1	2109	2109.4	53.1260	3.514e-13	***
genhealth:exerany	4	33	8.2	0.2067	0.9348	
Residuals	6266	248791	39.7			

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coming Up

Building Linear Regression Models

- Using Stepwise Regression to Select Variables (review)
- Using Best Subsets approaches to Select Variables (new)
 - Mallows' C_p , Adjusted R^2 , Corrected AIC, BIC
- Box-Cox plots to motivate outcome transformation (review)
- Spearman ρ^2 Plot to help motivate non-linearity via transformations and interaction terms in Linear Regression (new)
- Cross-Validation of Linear Regression Models (old and new)

to be followed by ...

- Logistic Regression Models for Binary Outcomes