

432 Class 02

Thomas E. Love, Ph.D.

2026-01-15

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

1

Today's Agenda

- Fitting linear regression models with `lm`
 - ANOVA: Incorporating an interaction between factors
 - ANCOVA: Incorporating a quantitative covariate
- Regression Diagnostics via `check_model()`
- Validating / evaluating results in a test sample

Appendix

How the `smt_im` data were created from `smart_ohio.csv`

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

2

Today's R Setup

```
1 knitr::opts_chunk$set(comment = NA)
2
3 library(janitor)
4 library(naniar)
5 library(broom)
6 library(car)
7 library(gt)
8 library(mosaic)      ## for df_stats and favstats
9 library(mice)         ## imputation of missing data
10 library(patchwork)
11 library(rsample)      ## data splitting
12 library(easystats)
13 library(tidyverse)
14
15 theme_set(theme_lucid())
```

The smt_im data

The `smt_im` data

- 894 subjects in Cleveland-Elyria with `bmi` and no history of diabetes (missing values singly imputed: assume MAR)
- All subjects have `hx_diabetes` (all 0), and are located in the `MMSA` labeled Cleveland-Elyria.
- See [Course Notes Chapter on BRFSS SMART data](#) for variable details
- Appendix provides details on data development.

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

5

The Five Variables We'll Use Today

9 variables in the data but we'll use only these 5 today.

Variable	Description
<code>ID</code>	subject identifying code
<code>bmi</code>	(outcome) Body-Mass index in kg/m^2 .
<code>exerany</code>	any exercise in past month: 1 = yes, 0 = no
<code>health</code>	self-reported overall health (5 levels)
<code>fruit_day</code>	average fruit servings consumed per day

Data Load

```
1 smt_im <- read_rds("c02/data/smt_im.Rds") |>
2   select(ID, bmi, exerany, health, fruit_day, everything())
3 
4 smt_im

# A tibble: 894 × 9
# ... with 9 variables:
#   ID     <chr> 
#   bmi    <dbl>  
#   exerany <dbl>  
#   health <fct>  
#   fruit_day <dbl> 
#   inc_imp <dbl>  
#   drinks_wk <dbl> 
#   female  <dbl>  
#   race_etc <fct>
# ... with 884 more rows
```

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

7

Checking our Data

Are there any missing values?

```
1 smt_im |> n_miss()

[1] 0
```

Does each row have a unique `ID` value?

```
1 identical(nrow(smt_im), n_distinct(smt_im$ID))

[1] TRUE
```

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

8

Range and Level Checks?

```
1 data_codebook(smt_im |> select(bmi, exerany, health, fruit_day))  
select(smt_im, bmi, exerany, health, fruit_day) (894 rows and 4 variables, 4 shown)
```

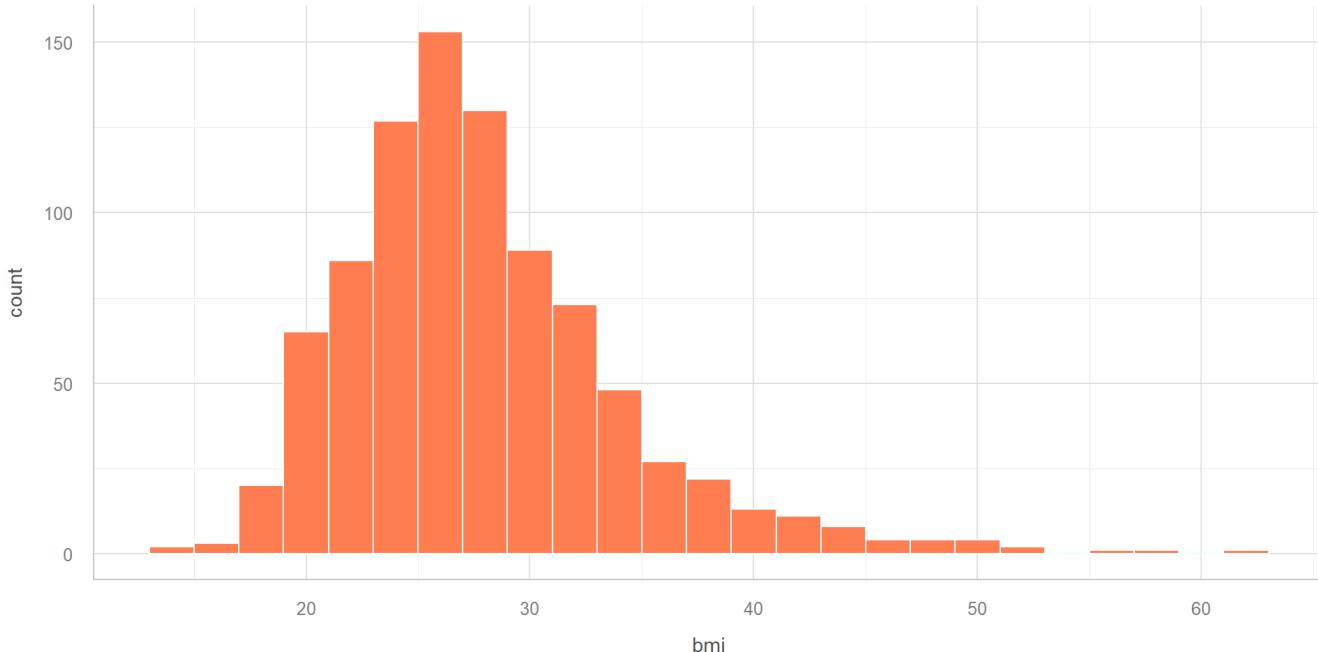
ID	Name	Type	Missing	Values	N
1	bmi	numeric	0 (0.0%)	[13.3, 63]	894
2	exerany	numeric	0 (0.0%)	0 233 (26.1%) 1 661 (73.9%)	
3	health	categorical	0 (0.0%)	E 148 (16.6%) VG 324 (36.2%) G 274 (30.6%) F 113 (12.6%) P 35 (3.9%)	

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

9

Our outcome, `bmi`

```
1 ggplot(smt_im, aes(x = bmi)) +  
2   geom_histogram(binwidth = 2, col = "azure", fill = "coral")
```



432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

10

Key predictors: exerany, health

```
1 smt_im |> tabyl(exerany, health) |>
2   adorn_totals(where = c("row", "col")) |>
3   gt() |> tab_options(table.font.size = 28)
```

exerany	E	VG	G	F	P	Total
0	24	68	76	51	14	233
1	124	256	198	62	21	661
Total	148	324	274	113	35	894

Here, it doesn't matter much whether we store the 1/0 in `exerany` as numeric or as a two-level factor in R. For binary variables, sometimes the numeric version will be more useful and sometimes a factor will be more useful.

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

11

Our covariate, `fruit_day`

We are mostly interested in whether accounting for the quantitative covariate `fruit_day` changes the modeled association of our key predictors with `bmi`.

- Sometimes we center such a covariate (subtracting its mean.)

```
1 smt_im <- smt_im |>
2   mutate(fruit_c = fruit_day - mean(fruit_day))
```

- Why? So that we can easily plug in the covariate's mean (which will now be 0) when making predictions.

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

12

Did we center `fruit_day` properly?

Here's a little “sanity check”:

```
1 df_stats(~ fruit_day + fruit_c, data = smt_im) |> gt() |>  
2   fmt_number(columns = min:sd, decimals = 3) |>  
3   tab_options(table.font.size = 24)
```

	response	min	Q1	median	Q3	max	mean	sd	n	missing
fruit_day	0.000	0.710	1.070	2.000	10.000	1.445	1.129	894	0	
fruit_c	-1.445	-0.735	-0.375	0.555	8.555	0.000	1.129	894	0	

The `df_stats()` function comes from the `mosaic` package, and can be used to apply `favstats()` to multiple variables at once.

Modeling Plan

1. Split `smt_im` into training and testing samples.
2. Predict `bmi` using `exer_any` and `health`
 - (`fit1`): without an interaction between the predictors
 - (`fit2`): and then with an interaction term
3. (`fit3`): Add in our (centered) covariate, `fruit_c` to `fit1`
4. (`fit4`): Add in our (centered) covariate, `fruit_c` to `fit2`
5. Assess all four models in training and testing samples.

Splitting the Sample

We'll partition our data set using some tools from the `rsample` package, into:

- a training sample containing 75% of the data
- a testing sample containing the remaining 25%

```
1 set.seed(432)      ## for future replication
2
3 smt_im_split <- initial_split(smt_im, prop = 3/4)
4
5 train_smt_im <- training(smt_im_split)
6 test_smt_im <- testing(smt_im_split)
7
8 c(nrow(smt_im), nrow(train_smt_im), nrow(test_smt_im))
```

[1] 894 670 224

Building Our Four Models

Modeling Plan

- Predict `bmi` using `exer_any` and `health`
 - (`fit1`): without an interaction between the predictors
 - (`fit2`): and then with an interaction term
 - (`fit3`): Add in our (centered) covariate, `fruit_c` to `fit1`
 - (`fit4`): Add in our (centered) covariate, `fruit_c` to `fit2`

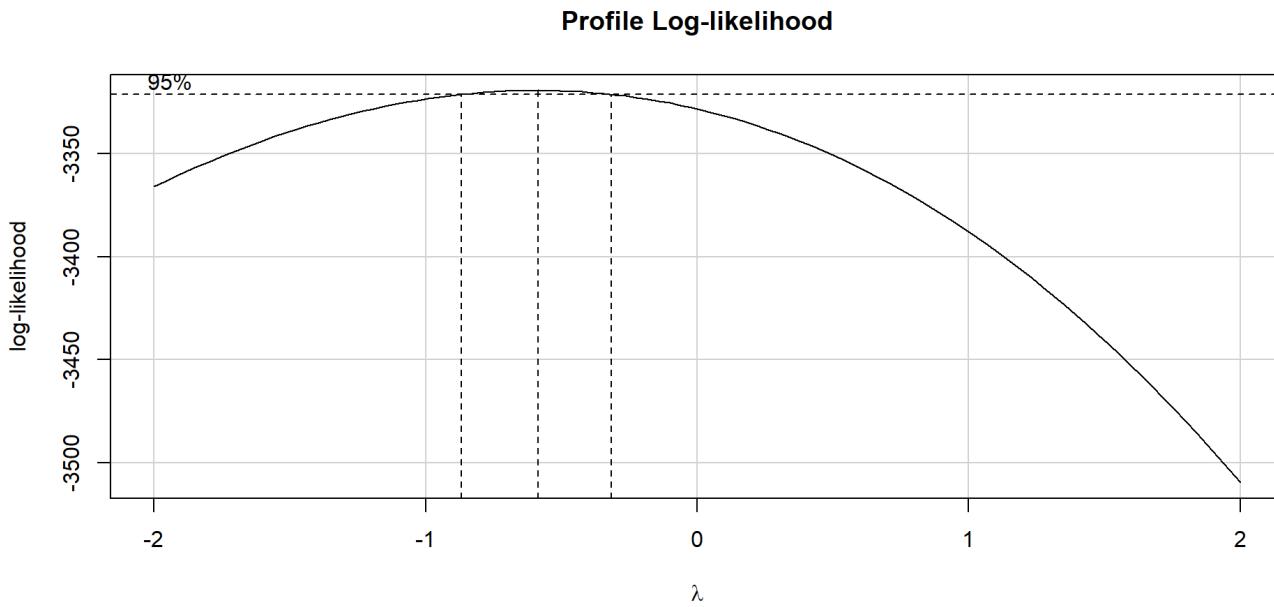
Tukey's transformation ladder

λ	2	1	0.5	0	-0.5	-1	-2
Transform	y^2	y	\sqrt{y}	$\log(y)$	$\frac{1}{\sqrt{y}}$	$\frac{1}{y}$	$\frac{1}{y^2}$

- Used in combination with a Box-Cox plot from `car`
- Requires the y variable to be **strictly positive**.
 - If desirable, you can add any constant to y or multiply y by any constant, before or after the transformation.
- Be sure you can back out of the transformation:
- $\exp(\log(y)) = y$, $\sqrt{y^2} = y$, $1/\frac{1}{y} = y$, $[1/(\frac{1}{\sqrt{y}})]^2 = y$

Consider transforming `bmi`?

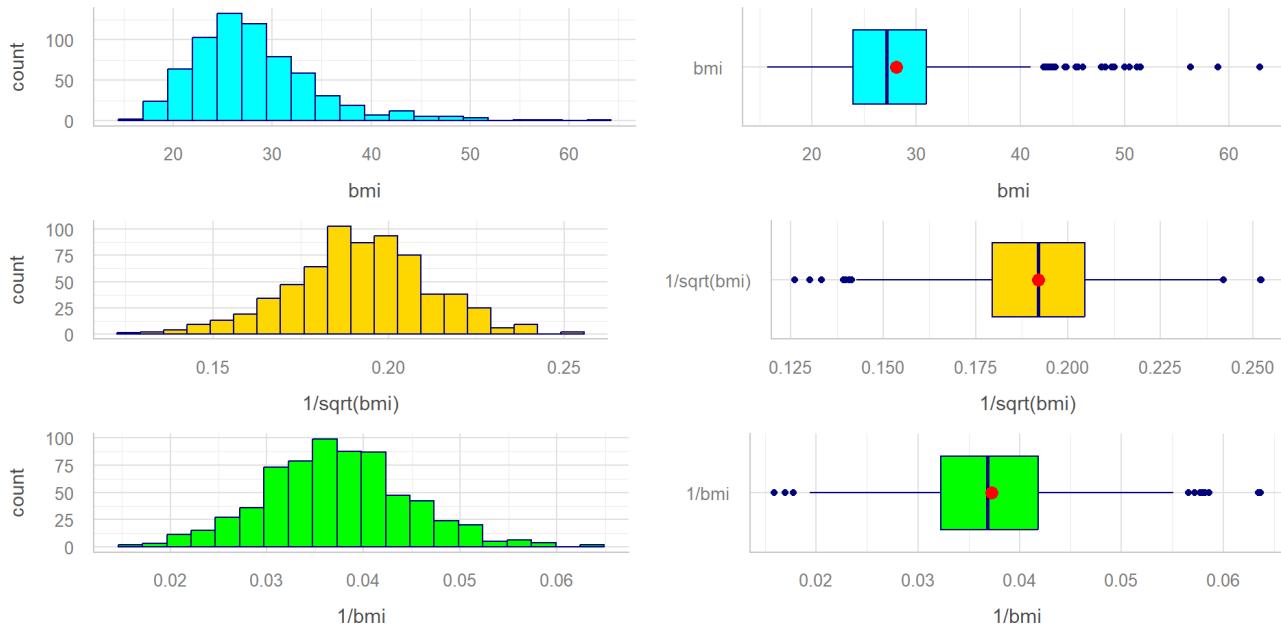
```
1 m0 <- lm(bmi ~ exerany + health + fruit_c, data = train_smt_im)
2 boxCox(m0)
```



Should we transform `bmi`? ($n = 670$)

```
1 p1a <- ggplot(train_smt_im, aes(x = bmi)) +
  geom_histogram(col = "navy", fill = "cyan", bins = 20)
3
4 p1b <- ggplot(train_smt_im, aes(x = bmi, y = "bmi")) +
  geom_boxplot(col = "navy", fill = "cyan") + labs(y = "") +
  stat_summary(geom = "point", col = "red", fun = mean, size = 3)
7
8 p2a <- ggplot(train_smt_im, aes(x = 1/sqrt(bmi))) +
  geom_histogram(col = "navy", fill = "gold", bins = 20)
10
11 p2b <- ggplot(train_smt_im, aes(x = 1/sqrt(bmi), y = "1/sqrt(bmi)")) +
  geom_boxplot(col = "navy", fill = "gold") + labs(y = "") +
  stat_summary(geom = "point", col = "red", fun = mean, size = 3)
14
15 p3a <- ggplot(train_smt_im, aes(x = 1/bmi)) +
  geom_histogram(col = "navy", fill = "green", bins = 20)
17
18 p3b <- ggplot(train_smt_im, aes(x = 1/bmi, y = "1/bmi")) +
```

Should we transform `bmi`? ($n = 670$)



432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

21

Re-scaling the transformation

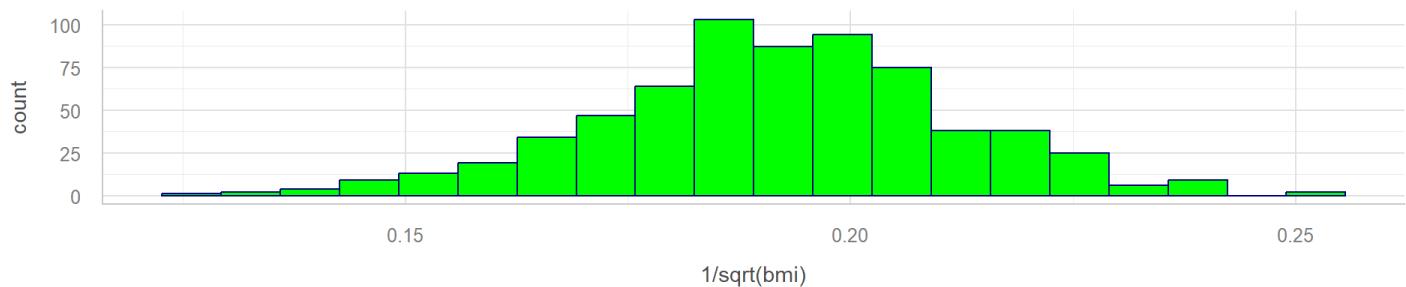
To ease interpretation of coefficients, I sometimes scale an outcome transformation so that its values fall in (10, 100), rather than between 0 and 1.

```
1 bind_rows( favstats(~ 1/sqrt(bmi), data = train_smt_im),
2             favstats(~ 100/sqrt(bmi), data = train_smt_im)) |>
3   mutate(outcome = c("1/sqrt(bmi)", "100/sqrt(bmi)")) |>
4   relocate(outcome) |>
5   gt() |> fmt_number(columns = min:sd, decimals = 3) |>
6   tab_options(table.font.size = 24)
```

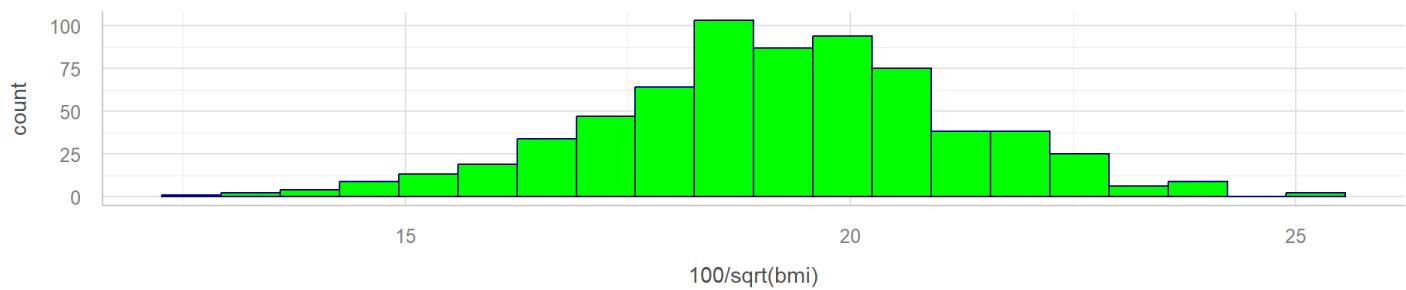
outcome	min	Q1	median	Q3	max	mean	sd	n	missing
<code>1/sqrt(bmi)</code>	0.126	0.180	0.192	0.204	0.252	0.192	0.020	670	0
<code>100/sqrt(bmi)</code>	12.599	17.958	19.194	20.447	25.230	19.195	1.987	670	0

Shape doesn't change

1/sqrt(BMI)



100/sqrt(BMI)



Means by exerany and health

```
1 summaries_1 <- train_smt_im |>
2   group_by(exerany, health) |>
3   summarise(n = n(), mean = mean(100/sqrt(bmi)),
4             stdev = sd(100/sqrt(bmi)))
5 summaries_1

# A tibble: 10 × 5
# Groups:   exerany [2]
  exerany health     n   mean  stdev
  <dbl> <fct> <int> <dbl> <dbl>
1     0 E         19  19.2  1.19
2     0 VG        54  19.5  1.91
3     0 G         60  18.5  2.21
4     0 F         35  17.8  2.57
5     0 P          8  17.1  2.26
6     1 E         91  19.9  1.65
7     1 VG        190 19.6  1.74
8     1 G         150 18.8  1.91
9     1 F         46  19.4  1.93
10    1 P         17  19.3  2.70
```

Code for Interaction Plot

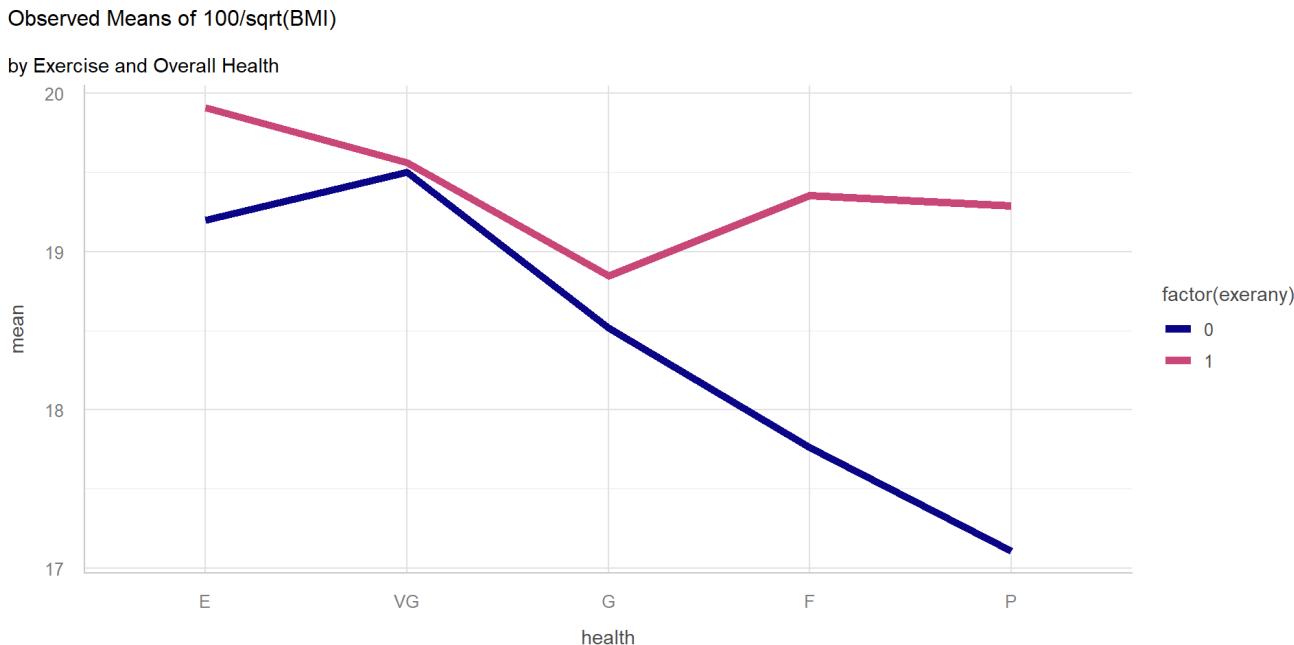
```
1 ggplot(summaries_1, aes(x = health, y = mean, col = factor(exerany))) +  
2   geom_line(aes(group = factor(exerany)), linewidth = 2) +  
3   scale_color_viridis_d(option = "C", end = 0.5) +  
4   labs(title = "Observed Means of 100/sqrt(BMI)",  
5       subtitle = "by Exercise and Overall Health")
```

- Note the use of `factor` here since the `exerany` variable is in fact numeric, although it only takes the values 1 and 0.
 - Sometimes it's helpful to treat 1/0 as a factor, and sometimes not.
- Where is the evidence of serious non-parallelism (if any) in the plot (see next slide) that results from this code?

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

25

Code for Interaction Plot



Fitting a Two-Way ANOVA model for $100/\sqrt{BMI}$

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

27

Create our transformed outcome

We'll want to do this in both our training and test samples.

```
1 train_smt_im <- train_smt_im |> mutate(bmi_tr = 100 / sqrt(bmi))
2
3 test_smt_im <- test_smt_im |> mutate(bmi_tr = 100 / sqrt(bmi))
```

Model fit1 without interaction

```
1 fit1 <- lm(bmi_tr ~ exerany + health, data = train_smt_im)
```

Using the `tidy()` function from `broom`:

```
1 tidy(fit1, conf.int = TRUE, conf.level = 0.90) |>
2   gt() |> fmt_number(columns = estimate:conf.high, decimals = 3) |>
3   tab_options(table.font.size = 24)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19.315	0.232	83.255	0.000	18.933	19.697
exerany	0.566	0.172	3.302	0.001	0.284	0.849
healthVG	-0.207	0.221	-0.938	0.349	-0.572	0.157
healthG	-0.968	0.227	-4.256	0.000	-1.342	-0.593
healthF	-0.969	0.285	-3.396	0.001	-1.439	-0.499
healthP	-1.110	0.427	-2.599	0.010	-1.814	-0.407

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

29

Model Parameters for fit1

```
1 model_parameters(fit1, ci = 0.90)
```

Parameter	Coefficient	SE	90% CI	t(664)	p
<hr/>					
(Intercept)	19.32	0.23	[18.93, 19.70]	83.26	< .001
exerany	0.57	0.17	[0.28, 0.85]	3.30	0.001
health [VG]	-0.21	0.22	[-0.57, 0.16]	-0.94	0.349
health [G]	-0.97	0.23	[-1.34, -0.59]	-4.26	< .001
health [F]	-0.97	0.29	[-1.44, -0.50]	-3.40	< .001
health [P]	-1.11	0.43	[-1.81, -0.41]	-2.60	0.010

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

30

Model Parameters for fit1

Reformatting with `gt()`...

```
1 model_parameters(fit1, ci = 0.90) |>
2   gt() |> fmt_number(columns = -c(CI, df_error), decimals = 3) |>
3   tab_options(table.font.size = 24)
```

Parameter	Coefficient	SE	CI	CI_low	CI_high	t	df_error	p
(Intercept)	19.315	0.232	0.9	18.933	19.697	83.255	664	0.000
exerany	0.566	0.172	0.9	0.284	0.849	3.302	664	0.001
healthVG	-0.207	0.221	0.9	-0.572	0.157	-0.938	664	0.349
healthG	-0.968	0.227	0.9	-1.342	-0.593	-4.256	664	0.000
healthF	-0.969	0.285	0.9	-1.439	-0.499	-3.396	664	0.001
healthP	-1.110	0.427	0.9	-1.814	-0.407	-2.599	664	0.010

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

31

The fit1 equation

fit1: $100/\sqrt{bmi} = 19.31 + .57 \text{ exerany} - .20 \text{ (VG)} - .96 \text{ (G)} - .98 \text{ (F)} - 1.08 \text{ (P)}$

Name	exerany	health	predicted $100/\sqrt{bmi}$
Harry	0	Excellent	19.31
Sally	1	Excellent	$19.31 + .57 = 19.88$
Billy	0	Fair	$19.31 - .98 = 18.33$
Meg	1	Fair	$19.31 + .57 - .98 = 18.90$

- Effect of `exerany` on $100/\sqrt{bmi}$?
- Effect of `health` = Fair instead of Excellent?

Does `fit1` fit the training data well?

```
1 n_obs(fit1)
[1] 670
1 model_performance(fit1)
# Indices of model performance

AIC      |   AICc    |     BIC   |      R2   | R2 (adj.) |     RMSE  |   Sigma
-----+-----+-----+-----+-----+-----+-----+
2786.9 | 2787.1 | 2818.5 | 0.069 | 0.062 | 1.916 | 1.925

1 glance(fit1) |>
2   select(r.squared, adj.r.squared, sigma, nobs,
3         df, df.residual, AIC, BIC) |>
4   gt() |> fmt_number(columns = r.squared:sigma, decimals = 3) |>
5   fmt_number(columns = AIC:BIC, decimals = 1) |>
6   tab_options(table.font.size = 24)
```

r.squared	adj.r.squared	sigma	nobs	df	df.residual	AIC	BIC
0.069	0.062	1.925	670	5		664	2,786.9 2,818.5

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

33

Tidied ANOVA for `fit1`

```
1 tidy(anova(fit1)) |> gt() |>
2   fmt_number(columns = sumsq:statistic, decimals = 2) |>
3   fmt_number(columns = p.value, decimals = 4) |>
4   tab_options(table.font.size = 24)
```

term	df	sumsq	meansq	statistic	p.value
exerany	1	63.76	63.76	17.21	0.0000
health	4	118.17	29.54	7.97	0.0000
Residuals	664	2,460.51	3.71	NA	NA

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

34

Model Checks

We'll be checking assumptions related to:

- linearity
- homoscedasticity (constant variance)
- influential observations (outliers, leverage and influence)
- whether the residuals follow a Normal distribution
- collinearity (variance inflation factor)
- and a posterior predictive check of our predictions

My slides and `check_model()`

When building a regular HTML file, I would just use:

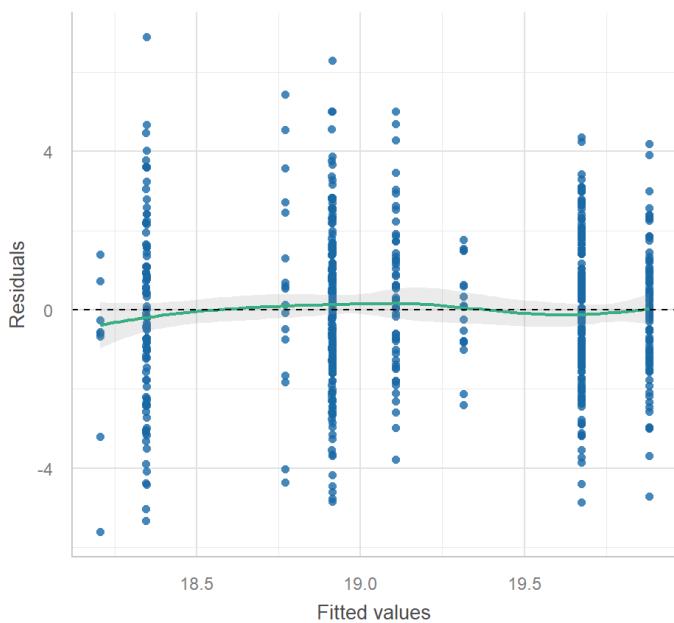
```
1 check_model(fit1, detrend = FALSE)
```

with `#| fig-height: 9` at the start of the code chunk so that the plots are taller than the default height (thus easier to read) but I will split out the plots for slides.

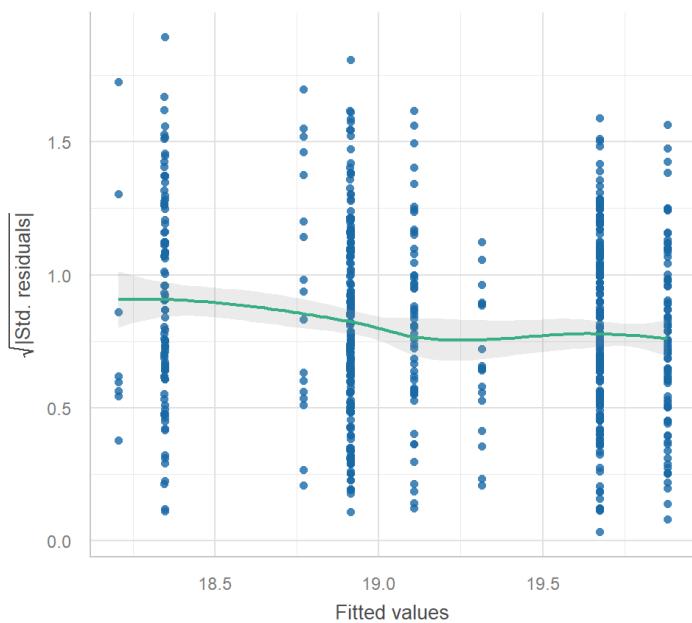
Checking model `fit1` ($n = 670$)

```
1 check_model(fit1, check = c("linearity", "homogeneity"))
```

Linearity
Reference line should be flat and horizontal



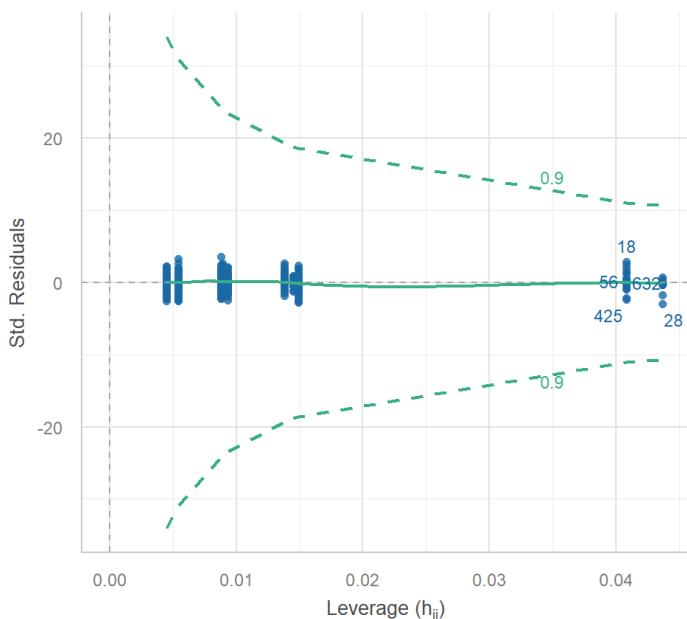
Homogeneity of Variance
Reference line should be flat and horizontal



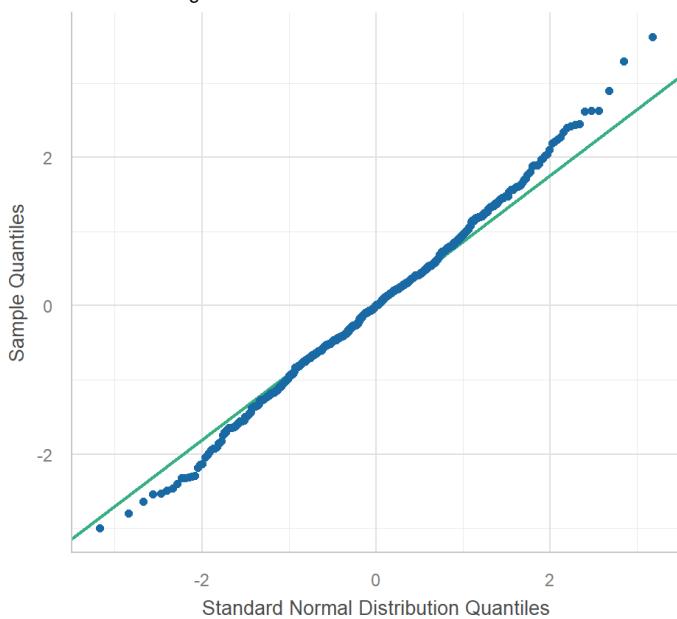
Checking model `fit1` ($n = 670$)

```
1 check_model(fit1, check = c("outliers", "qq"), detrend = FALSE)
```

Influential Observations
Points should be inside the contour lines



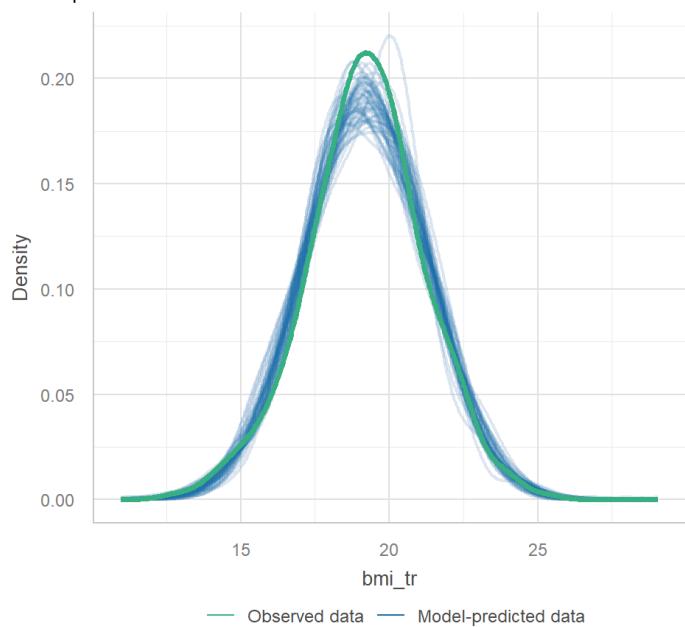
Normality of Residuals
Dots should fall along the line



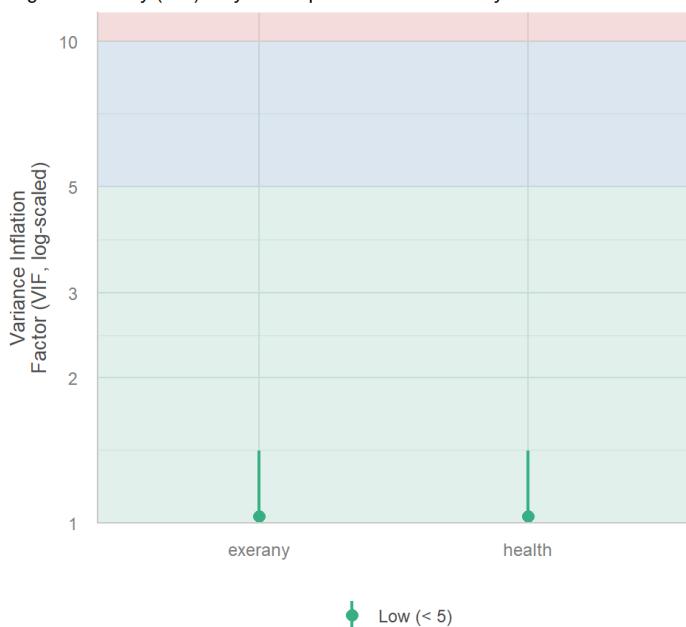
Checking model `fit1` ($n = 670$)

```
1 check_model(fit1, check = c("pp_check", "vif"))
```

Posterior Predictive Check
Model-predicted lines should resemble observed data line



Collinearity
High collinearity (VIF) may inflate parameter uncertainty



Standing Break

Fitting ANOVA model fit2 including interaction

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

41

Adding the interaction term to fit1

```
1 fit2 <- lm(bmi_tr ~ exerany * health, data = train_smt_im)
```

- How do our models compare on fit to the training data?

```
1 bind_rows(glance(fit1), glance(fit2)) |>
2   mutate(mod = c("fit1", "fit2")) |>
3   select(mod, r.sq = r.squared, adj.r.sq = adj.r.squared,
4         sigma, nobs, df, df.res = df.residual, AIC, BIC) |>
5   gt() |> fmt_number(columns = r.sq:sigma, decimals = 3) |>
6   fmt_number(columns = AIC:BIC, decimals = 1) |>
7   tab_options(table.font.size = 24)
```

mod	r.sq	adj.r.sq	sigma	nobs	df	df.res	AIC	BIC
fit1	0.069	0.062	1.925	670	5	664	2,786.9	2,818.5
fit2	0.087	0.075	1.912	670	9	660	2,781.6	2,831.2

ANOVA for the `fit2` model

```
1 tidy(anova(fit2)) |> gt() |>
2   fmt_number(columns = sumsq:statistic, decimals = 2) |>
3   fmt_number(columns = p.value, decimals = 4) |>
4   tab_options(table.font.size = 20)
```

term	df	sumsq	meansq	statistic	p.value
exerany	1	63.76	63.76	17.45	0.0000
health	4	118.17	29.54	8.08	0.0000
exerany:health	4	48.44	12.11	3.31	0.0106
Residuals	660	2,412.06	3.65	NA	NA

ANOVA test comparing `fit1` to `fit2`

```
1 anova(fit1, fit2)
```

Analysis of Variance Table

```
Model 1: bmi_tr ~ exerany + health
Model 2: bmi_tr ~ exerany * health
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     664 2460.5
2     660 2412.1  4    48.444 3.3139 0.01059 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fit2 coefficients

```
1 tidy(fit2, conf.int = TRUE, conf.level = 0.90) |>
2   gt() |> fmt_number(columns = estimate:conf.high, decimals = 3) |>
3   tab_options(table.font.size = 20)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19.196	0.439	43.770	0.000	18.474	19.919
exerany	0.710	0.482	1.472	0.141	-0.084	1.504
healthVG	0.303	0.510	0.593	0.553	-0.537	1.142
healthG	-0.678	0.503	-1.347	0.179	-1.507	0.151
healthF	-1.432	0.545	-2.629	0.009	-2.330	-0.535
healthP	-2.089	0.806	-2.593	0.010	-3.416	-0.762
exerany:healthVG	-0.646	0.565	-1.143	0.253	-1.577	0.285
exerany:healthG	-0.383	0.564	-0.680	0.497	-1.312	0.545
exerany:healthF	0.881	0.645	1.366	0.173	-0.182	1.944
exerany:healthP	1.471	0.951	1.546	0.122	-0.096	3.037

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

45

Interpreting the fit2 model

Name	exerany	health	predicted $100/\sqrt{bmi}$
Harry	0	Excellent	19.19
Sally	1	Excellent	$19.19 + .72 = 19.91$
Billy	0	Fair	$19.19 - 1.73 = 17.46$
Meg	1	Fair	$19.19 + .72 - 1.73 + 1.29 = 19.47$

- How do we interpret effect sizes here? It depends...

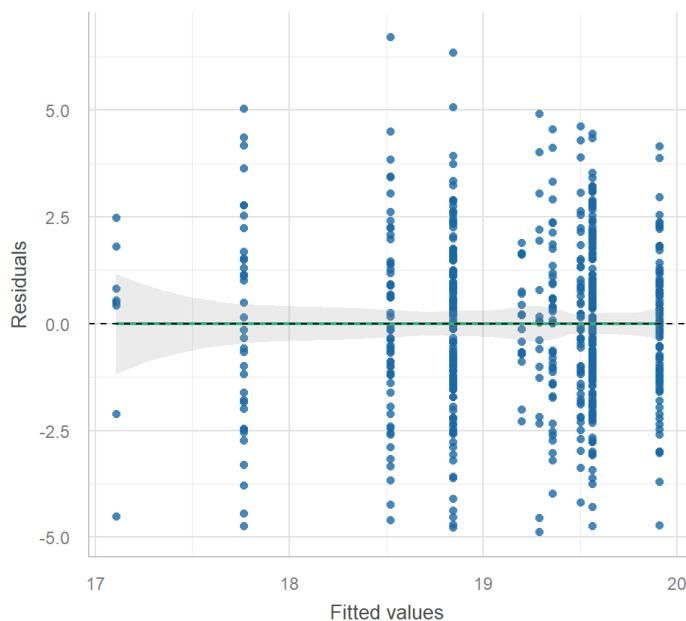
Interpreting the `fit2` model

- Effect of `exerany` on predicted $100/\sqrt{bmi}$?
 - If `health` = Excellent, effect is +0.72
 - If `health` = Fair, effect is $(0.72 + 1.29) = +2.01$
- Effect of `health` = Fair instead of Excellent?
 - If `exerany` = 0 (no), effect is -1.73
 - If `exerany` = 1 (yes), effect is $(-1.73 + 1.29) = -0.44$

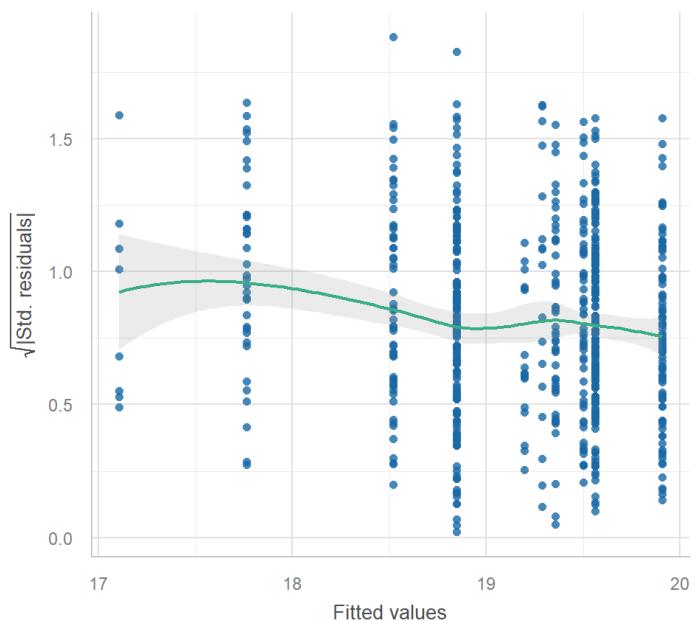
Checking model `fit2` ($n = 670$)

```
1 check_model(fit2, check = c("linearity", "homogeneity"))
```

Linearity
Reference line should be flat and horizontal



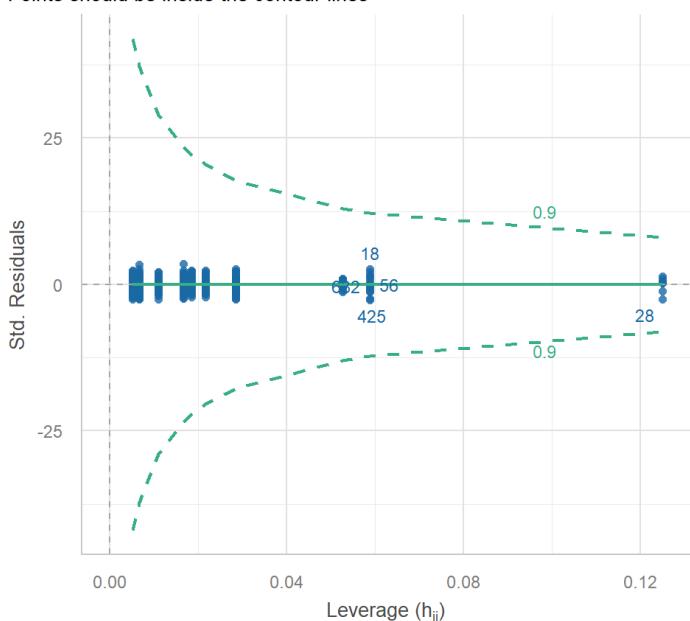
Homogeneity of Variance
Reference line should be flat and horizontal



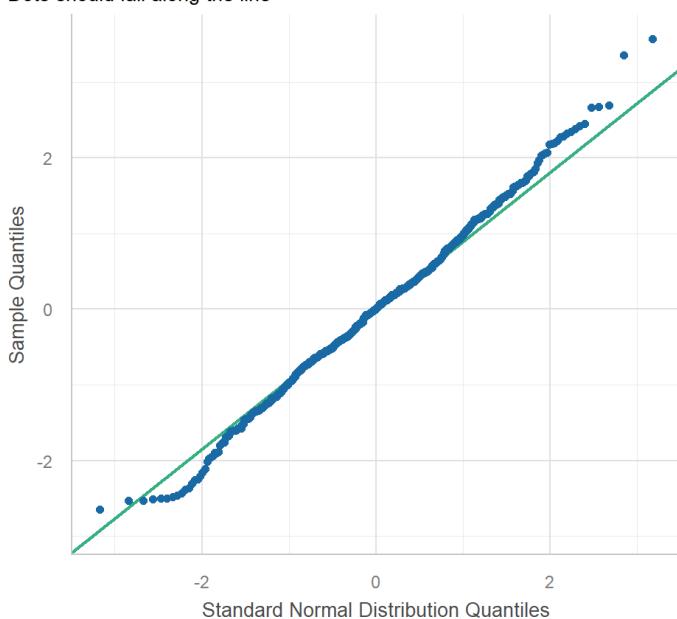
Checking model `fit2` ($n = 670$)

```
1 check_model(fit2, check = c("outliers", "qq"), detrend = FALSE)
```

Influential Observations
Points should be inside the contour lines



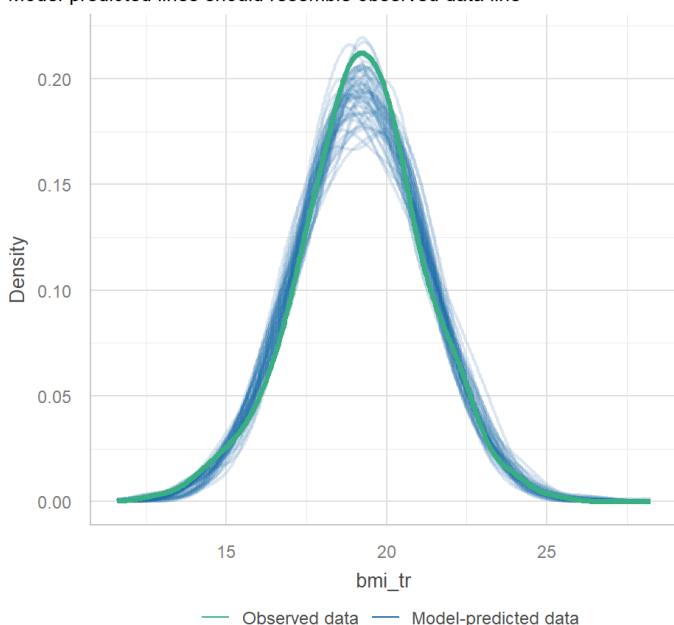
Normality of Residuals
Dots should fall along the line



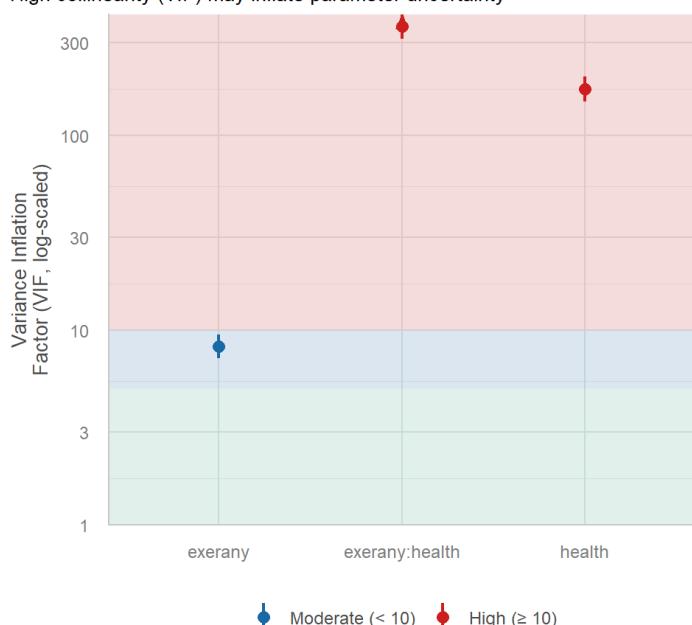
Checking model `fit2` ($n = 670$)

```
1 check_model(fit2, check = c("pp_check", "vif"))
```

Posterior Predictive Check
Model-predicted lines should resemble observed data line



Collinearity
High collinearity (VIF) may inflate parameter uncertainty



Incorporating a Covariate into our two-way ANOVA models

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

51

Add `fruit_c` to `fit1`

```
1 fit3 <- lm(bmi_tr ~ fruit_c + exerany + health, data = train_smt_im)
```

- How well does this model fit the training data?

```
1 bind_rows(glance(fit1), glance(fit3)) |>
2   mutate(mod = c("fit1", "fit3")) |>
3   select(mod, r.sq = r.squared, adj.r.sq = adj.r.squared,
4          sigma, df, df.res = df.residual, AIC, BIC) |>
5   gt() |> fmt_number(columns = r.sq:sigma, decimals = 3) |>
6   fmt_number(columns = AIC:BIC, decimals = 1) |>
7   tab_options(table.font.size = 24)
```

	mod	r.sq	adj.r.sq	sigma	df	df.res	AIC	BIC
	fit1	0.069	0.062	1.925	5	664	2,786.9	2,818.5
	fit3	0.079	0.071	1.916	6	663	2,781.4	2,817.4

ANOVA for the fit3 model

```
1 tidy(anova(fit3)) |> gt() |>
2   fmt_number(columns = sumsq:statistic, decimals = 2) |>
3   fmt_number(columns = p.value, decimals = 4) |>
4   tab_options(table.font.size = 24)
```

term	df	sumsq	meansq	statistic	p.value
fruit_c	1	48.95	48.95	13.34	0.0003
exerany	1	51.09	51.09	13.92	0.0002
health	4	109.59	27.40	7.47	0.0000
Residuals	663	2,432.81	3.67	NA	NA

fit3 coefficients

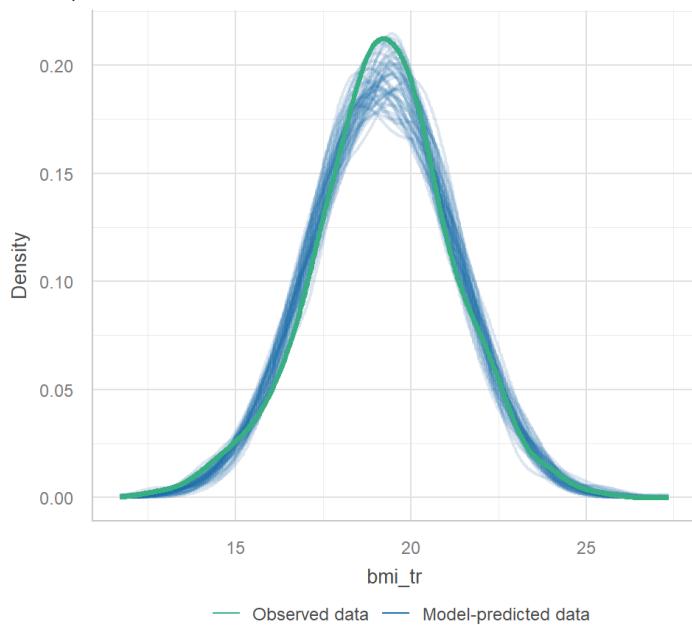
```
1 tidy(fit3, conf.int = TRUE, conf.level = 0.90) |>
2   gt() |> fmt_number(columns = estimate:conf.high, decimals = 3) |>
3   tab_options(table.font.size = 24)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19.325	0.231	83.697	0.000	18.944	19.705
fruit_c	0.178	0.065	2.747	0.006	0.071	0.285
exerany	0.512	0.172	2.982	0.003	0.229	0.795
healthVG	-0.191	0.220	-0.870	0.385	-0.554	0.171
healthG	-0.926	0.227	-4.085	0.000	-1.300	-0.553
healthF	-0.914	0.285	-3.213	0.001	-1.383	-0.446
healthP	-1.119	0.425	-2.632	0.009	-1.819	-0.419

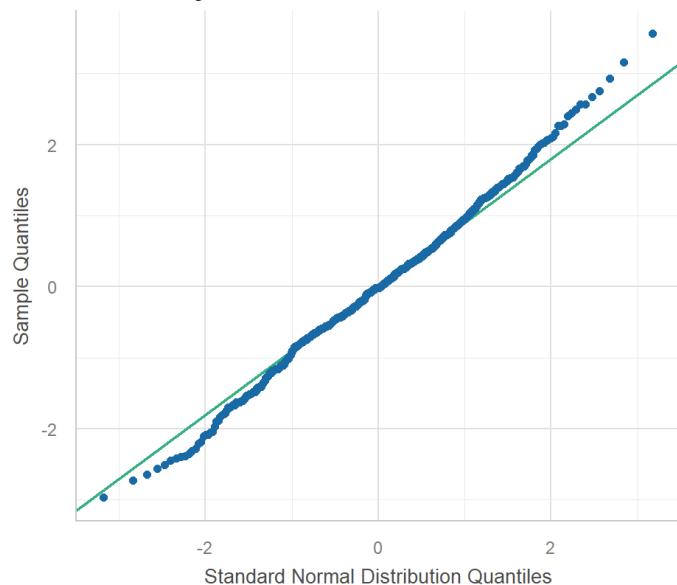
Checking model **fit3** ($n = 670$)

```
1 check_model(fit3, detrend = FALSE, check = c("pp_check", "qq"))
```

Posterior Predictive Check
Model-predicted lines should resemble observed data line



Normality of Residuals
Dots should fall along the line



Include the interaction term?

```
1 fit4 <- lm(bmi_tr ~ fruit_c + exerany * health,  
2               data = train_smt_im)
```

ANOVA for the **fit4** model

```
1 tidy(anova(fit4)) |> gt() |>  
2   fmt_number(columns = sumsq:statistic, decimals = 2) |>  
3   fmt_number(columns = p.value, decimals = 4) |>  
4   tab_options(table.font.size = 20)
```

term	df	sumsq	meansq	statistic	p.value
fruit_c	1	48.95	48.95	13.55	0.0003
exerany	1	51.09	51.09	14.15	0.0002
health	4	109.59	27.40	7.59	0.0000
exerany:health	4	53.09	13.27	3.68	0.0057
Residuals	659	2,379.72	3.61	NA	NA

fit4 coefficients

```
1 tidy(fit4, conf.int = TRUE, conf.level = 0.90) |>
2   gt() |> fmt_number(columns = estimate:conf.high, decimals = 3) |>
3   tab_options(table.font.size = 18)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19.219	0.436	44.079	0.000	18.501	19.938
fruit_c	0.193	0.065	2.993	0.003	0.087	0.300
exerany	0.636	0.480	1.325	0.186	-0.155	1.427
healthVG	0.325	0.507	0.641	0.522	-0.510	1.160
healthG	-0.643	0.500	-1.285	0.199	-1.467	0.181
healthF	-1.403	0.542	-2.590	0.010	-2.295	-0.511
healthP	-2.207	0.802	-2.753	0.006	-3.528	-0.887
exerany:healthVG	-0.653	0.562	-1.163	0.245	-1.579	0.272
exerany:healthG	-0.372	0.560	-0.663	0.507	-1.295	0.551
exerany:healthF	0.927	0.642	1.444	0.149	-0.130	1.984
exerany:healthP	1.627	0.947	1.719	0.086	0.068	3.187

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

57

ANOVA: Compare fit3 & fit4

```
1 anova(fit3, fit4)
```

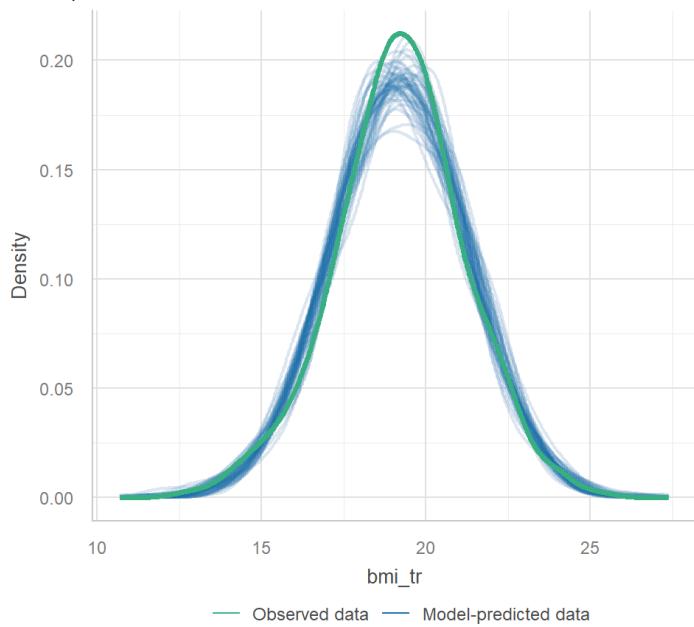
Analysis of Variance Table

```
Model 1: bmi_tr ~ fruit_c + exerany + health
Model 2: bmi_tr ~ fruit_c + exerany * health
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     663 2432.8
2     659 2379.7  4    53.092 3.6756 0.005699 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

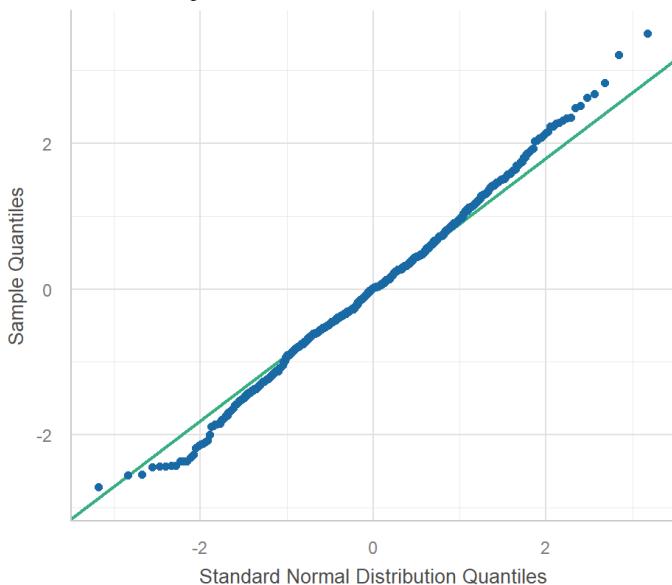
Checking model `fit4` ($n = 670$)

```
1 check_model(fit4, detrend = FALSE, check = c("pp_check", "qq"))
```

Posterior Predictive Check
Model-predicted lines should resemble observed data line



Normality of Residuals
Dots should fall along the line



Comparing Our Models

Which of the four models fits best?

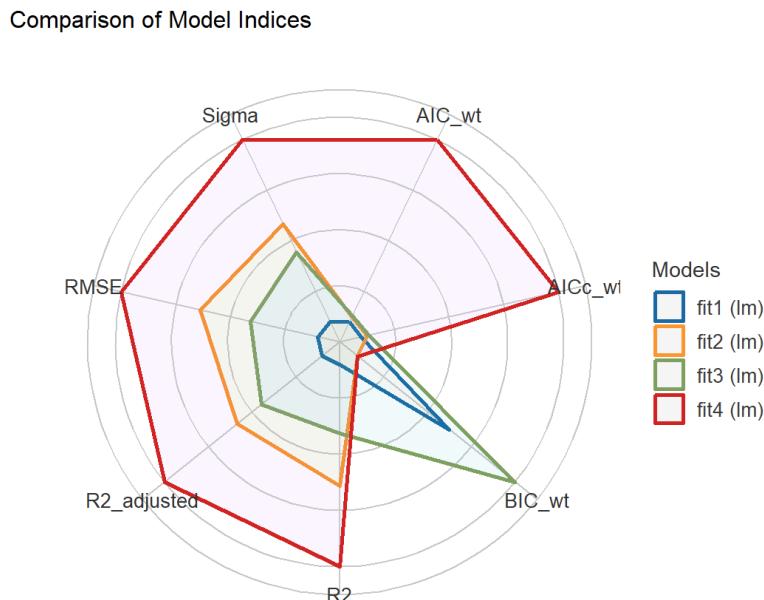
In the training sample, our model results (note ordering):

mod	r.sq	adj.r.sq	sigma	df	df.res	AIC	BIC
fit1	0.069	0.062	1.925	5	664	2,786.9	2,818.5
fit3	0.079	0.071	1.916	6	663	2,781.4	2,817.4
fit2	0.087	0.075	1.912	9	660	2,781.6	2,831.2
fit4	0.099	0.086	1.900	10	659	2,774.6	2,828.7

- Adjusted R^2 , σ and AIC all improve as we move down this table. BIC likes `fit1` and `fit3`.
- The training sample is the data our models have *already seen*, so we should be cautious.

Comparison Plot: Training Sample

```
1 plot(compare_performance(fit1, fit2, fit3, fit4))
```



What does `augment()` give us?

```
1 fit1_test_aug <- augment(fit1, newdata = test_smt_im)
2 fit1_test_aug |> select(ID, bmi_tr, bmi, .fitted, .resid, health, exerany) |>
3   slice(198:202) |> gt() |>
4   fmt_number(columns = bmi_tr:.resid, decimals = 2) |>
5   tab_options(table.font.size = 20)
```

ID	bmi_tr	bmi	.fitted	.resid	health	exerany
1016	18.75	28.44	18.20	0.55	P	0
1018	19.36	26.68	19.88	-0.52	E	1
1019	19.71	25.74	18.91	0.80	F	1
1020	22.05	20.57	19.67	2.37	VG	1
1024	20.19	24.52	18.35	1.85	G	0

Here, `.fitted` = predicted `bmi_tr` and `.resid` = `bmi_tr - .fitted`.

Back-Transformation of `bmi_tr`

Our models predict $\text{bmi_tr} = 100/\sqrt{\text{bmi}}$, but we want to predict `bmi`. How do we convert predicted $100/\sqrt{\text{bmi}}$ to predicted `bmi`?

$$1/\left(\frac{100}{\sqrt{\text{bmi}}}\right) = \sqrt{\text{bmi}}/100,$$

$$\text{so } 100/\left(\frac{100}{\sqrt{\text{bmi}}}\right) = \sqrt{\text{bmi}},$$

$$\text{and so } \left[100/\left(\frac{100}{\sqrt{\text{bmi}}}\right)\right]^2 = \text{bmi}$$

augment() with results for bmi

We use $(\frac{100}{.fitted})^2$ for predicted bmi, then errors are bmi_res = observed bmi - predicted bmi.

```
1 fit1_test_aug <- augment(fit1, newdata = test_smt_im) |>
2   mutate(bmi_fit = (100/.fitted)^2, bmi_res = bmi - bmi_fit)
3 
4 fit1_test_aug |> select(ID, bmi, bmi_fit, bmi_res,
5   bmi_tr, .fitted, .resid, exerany, health, fruit_c) |>
6   slice(5:6) |> gt() |>
7   fmt_number(columns = c(bmi:.resid, fruit_c), decimals = 2) |>
8   tab_options(table.font.size = 24)
```

ID	bmi	bmi_fit	bmi_res	bmi_tr	.fitted	.resid	exerany	health	fruit_c
21	18.83	25.84	-7.01	23.04	19.67	3.37	1	VG	-0.45
29	44.66	29.71	14.95	14.96	18.35	-3.38	0	G	-0.74

Augment all four models so far...

```
1 fit1_test_aug <- augment(fit1, newdata = test_smt_im) |>
2   mutate(bmi_fit = (100/.fitted)^2, bmi_res = bmi - bmi_fit)
3 
4 fit2_test_aug <- augment(fit2, newdata = test_smt_im) |>
5   mutate(bmi_fit = (100/.fitted)^2, bmi_res = bmi - bmi_fit)
6 
7 fit3_test_aug <- augment(fit3, newdata = test_smt_im) |>
8   mutate(bmi_fit = (100/.fitted)^2, bmi_res = bmi - bmi_fit)
9 
10 fit4_test_aug <- augment(fit4, newdata = test_smt_im) |>
11   mutate(bmi_fit = (100/.fitted)^2, bmi_res = bmi - bmi_fit)
```

Four Key Error Summaries

We'll look at all four of these summaries when we do linear regression, usually.

- Mean absolute prediction error (MAPE)
- Maximum absolute prediction error (Max. Error)
- Square root of mean squared prediction error (RMSPE)
- Squared correlation of observed and predicted `bmi` (validated R^2)

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

67

Key Summaries for `fit1`

```
1 fit1_esum <- fit1_test_aug |>
2   summarise(MAPE = mean(abs(bmi_res)),
3             Max_E = max(abs(bmi_res)),
4             RMSPE = sqrt(mean(bmi_res^2)),
5             Val_R2 = cor(bmi, bmi_fit)^2) |>
6   mutate(Model = "fit1")
7
8 fit1_esum
```



```
# A tibble: 1 × 5
  MAPE Max_E RMSPE Val_R2 Model
  <dbl> <dbl> <dbl>   <dbl> <chr>
1  4.30  21.1  5.62  0.0777 fit1
```

- I built the key summaries for `fit2`, `fit3` and `fit4` in the same way (included in code, not shown in slides.)

Compare Models in Test Sample

```
1 bind_rows(fit1_esum, fit2_esum, fit3_esum, fit4_esum) |>
2   relocate(Model) |> gt() |> fmt_number(decimals = 3) |>
3   tab_options(table.font.size = 24)
```

Model	MAPE	Max_E	RMSPE	Val_R2
fit1	4.300	21.140	5.621	0.078
fit2	4.430	20.530	5.792	0.042
fit3	4.325	21.107	5.650	0.068
fit4	4.465	21.162	5.844	0.036

Our Four Models

- `fit1`: `exerany` and `health` main effects; `fit2`: add interaction
- `fit3`: add `fruit_c` to `fit1`; `fit4`: add `fruit_c` to `fit2`

What's coming up next?

Basics of logistic regression fitting and evaluation

- What if we have a binary (yes/no or 1/0) outcome?
- Predict “whether or not BMI < 30”, rather than BMI?
 - A linear probability model as a first idea
 - Using `glm()` rather than `lm()` to get a logistic model
 - Coefficients as log(odds ratios)
 - Changes in how we measure the model’s performance
 - Changes in the assumptions we make

What to do before Class 3?

1. [Lab 1](#) due to [Canvas](#) at noon Wednesday 2026-01-21.
2. Get started reading *How To Be a Modern Scientist* so you finish the little book by the end of January.
3. Be sure you have everything working properly in RStudio. Visit TA office hours (starting tomorrow, no hours this Monday) for help, or email **431-help** at case dot edu.
4. Review the [Project A instructions](#).
5. The [Calendar](#) is the most important thing to look at.

Appendix: Creating Today's Data

Creating Today's Data Set

```
1 url1 <- "https://raw.githubusercontent.com/THOMASELOVE/432-data/master/data/smart_"
2
3 smart_ohio <- read_csv(url1)
4
5 smt <- smart_ohio |>
6   filter(hx_diabetes == 0, mmsa == "Cleveland-Elyria",
7         complete.cases(bmi)) |>
8   select(bmi, inc_imp, fruit_day, drinks_wk,
9         female, exerany, genhealth, race_eth,
10        hx_diabetes, mmsa, SEQNO) |>
11  mutate(across(where(is.character), as_factor)) |>
12  mutate(ID = as.character(SEQNO - 2017000000)) |>
13  relocate(ID)
```

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

73

Codebook for useful smt variables (1)

- 894 subjects in Cleveland-Elyria with `bmi` and no history of diabetes

Variable	Description
<code>bmi</code>	(outcome) Body-Mass index in kg/m^2 .
<code>inc_imp</code>	income (imputed from grouped values) in \$
<code>fruit_day</code>	average fruit servings consumed per day
<code>drinks_wk</code>	average weekly alcoholic drinks consumed
<code>female</code>	sex: 1 = female, 0 = male

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

74

Codebook for useful `smt` variables (2)

- 894 subjects in Cleveland-Elyria without diabetes

Variable	Description
<code>exerany</code>	any exercise in past month: 1 = yes, 0 = no
<code>genhealth</code>	self-reported overall health (5 levels)
<code>race_eth</code>	race and Hispanic/Latinx ethnicity (5 levels)

- plus `ID`, `SEQNO`, `hx_diabetes` (all 0), `MMSA`
- See [Course Notes Chapter 6](#) on BRFSS SMART data

Basic Data Summaries

Available approaches include:

- `data_codebook()` from `datawizard` in `easystats`
- `Hmisc` package's `describe()`, or
- `summary()`

all of which can work nicely in an HTML presentation, but none of them fit well on a slide.

Histogram of each quantity

Note

I used `#| warning: false` in this code chunk to avoid warnings about missing values, like this one for `inc_imp`:

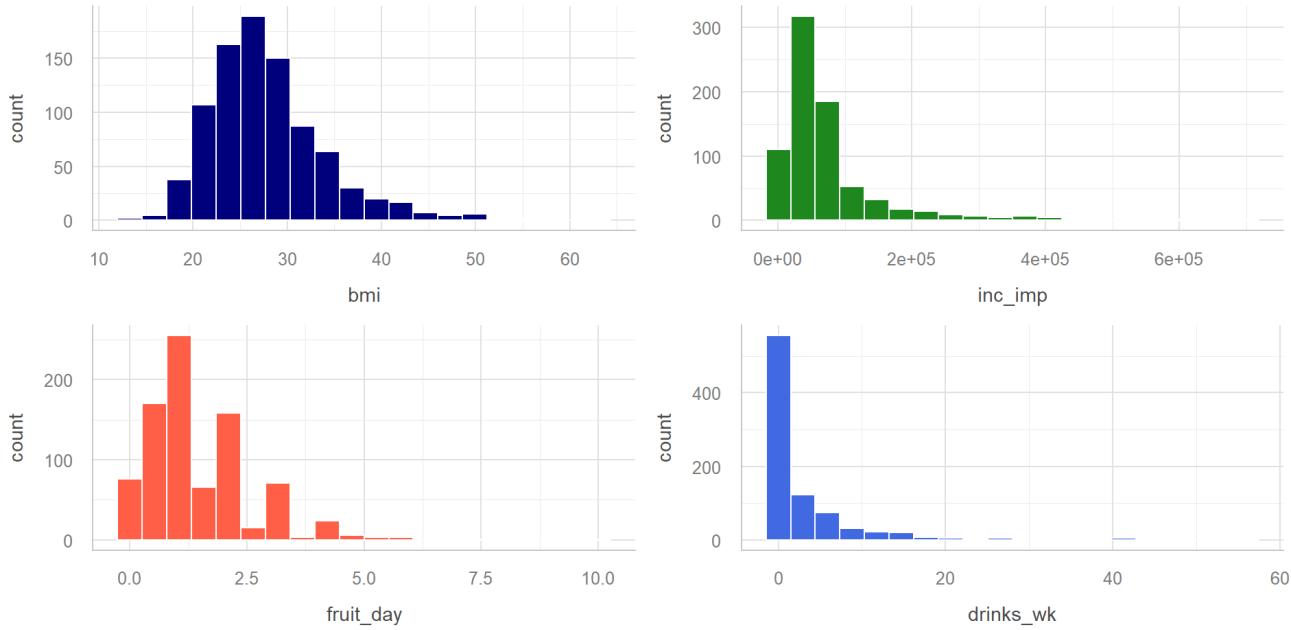
```
Warning: Removed 120 rows containing non-finite values
```

```
1 p1 <- ggplot(smt, aes(x = bmi)) +
2   geom_histogram(fill = "navy", col = "white", bins = 20)
3 p2 <- ggplot(smt, aes(x = inc_imp)) +
4   geom_histogram(fill = "forestgreen", col = "white", bins = 20)
5 p3 <- ggplot(smt, aes(x = fruit_day)) +
6   geom_histogram(fill = "tomato", col = "white", bins = 20)
7 p4 <- ggplot(smt, aes(x = drinks_wk)) +
8   geom_histogram(fill = "royalblue", col = "white", bins = 20)
9
10 (p1 + p2) / (p3 + p4)
```

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

77

Histogram of each quantity



432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

78

Binary variables in raw smt

```
1 smt |> tabyl(female, exerany) |> adorn_title()
```

	exerany			
female	0	1	NA_	
0	95	268	20	
1	128	361	22	

- `female` is based on biological sex (1 = female, 0 = male)
- `exerany` comes from a response to “During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?” (1 = yes, 0 = no, don’t know and refused = missing)
- Any signs of trouble here?

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

79

Multicategorical genhealth in raw smt

```
1 smt |> tabyl(genhealth)
```

	genhealth	n	percent	valid_percent
1_Excellent	148	0.165548098	0.16573348	
3_Good	274	0.306487696	0.30683091	
2_VeryGood	324	0.362416107	0.36282195	
4_Fair	112	0.125279642	0.12541993	
5_Poor	35	0.039149888	0.03919373	
<NA>	1	0.001118568		NA

- The variable is based on “Would you say that in general your health is ...” using the five specified categories (Excellent -> Poor), numbered for convenience after data collection.
- Don’t know / not sure / refused treated as missing.
- How might we manage this variable?

Changing the levels for genhealth

```
1 smt <- smt |>
2   mutate(health =
3     fct_recode(genhealth,
4       E = "1_Excellent",
5       VG = "2_VeryGood",
6       G = "3_Good",
7       F = "4_Fair",
8       P = "5_Poor"),
9     health = fct_relevel(health, "E", "VG", "G", "F", "P"))
```

Might want to run a sanity check here, just to be sure...

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

81

Checking health vs. genhealth

```
1 smt |> tabyl(genhealth, health) |> adorn_title()
```

health						
genhealth	E	VG	G	F	P	NA_
1_Excellent	148	0	0	0	0	0
3_Good	0	0	274	0	0	0
2_VeryGood	0	324	0	0	0	0
4_Fair	0	0	0	112	0	0
5_Poor	0	0	0	0	35	0
<NA>	0	0	0	0	0	1

- OK. We've adjusted the order to something more sensible, retained the missing value, and we have much shorter labels.

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

82

Multicategorical `race_eth` in raw smt

```
1 smt |> count(race_eth)

# A tibble: 6 × 2
  race_eth                n
  <fct>                 <int>
1 White non-Hispanic     646
2 Other race non-Hispanic    22
3 Black non-Hispanic      167
4 Multiracial non-Hispanic   19
5 Hispanic                  27
6 <NA>                      13
```

“Don’t know”, “Not sure”, and “Refused” were treated as missing.

- What is this variable actually about?
- What is the most common thing people do here?

What is the question you are asking?

Collapsing `race_eth` levels *might* be rational for some questions.

- We have lots of data from two categories, but only two.
- Systemic racism affects people of color in different ways across these categories, but also *within* them.

Is combining race and Hispanic/Latinx ethnicity helpful?

It's hard to see the justice in collecting this information and not using it in as granular a form as possible, though this leaves some small sample sizes. There is no magic number for "too small a sample size."

- Most people identified themselves in one category.
- These data are not ordered, and (I'd argue) ordering them isn't helpful.
- Regression models are easier to interpret, though, if the "baseline" category is a common one.

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

85

Resorting the factor for `race_eth`

Let's sort all five levels, from most observations to least...

```
1 smt <- smt |>  
2     mutate(race_eth = fct_infreq(race_eth))  
3  
4 smt |> tabyl(race_eth)
```

	race_eth	n	percent	valid_percent
White	non-Hispanic	646	0.72259508	0.73325766
Black	non-Hispanic	167	0.18680089	0.18955732
	Hispanic	27	0.03020134	0.03064699
Other	race non-Hispanic	22	0.02460850	0.02497162
Multiracial	non-Hispanic	19	0.02125280	0.02156640
	<NA>	13	0.01454139	NA

- Not a perfect solution, certainly, but we'll try it out.

“Cleaned” Data and Missing Values

```
1 smt <- smt |>
2   select(ID, bmi, inc_imp, fruit_day, drinks_wk,
3         female, exerany, health, race_eth)
4
5 miss_var_summary(smt)

# A tibble: 9 × 3
  variable n_miss pct_miss
  <chr>     <int>    <num>
1 inc_imp      120    13.4
2 exerany       42     4.70
3 fruit_day     41     4.59
4 drinks_wk     39     4.36
5 race_eth      13     1.45
6 health        1     0.112
7 ID            0     0
8 bmi           0     0
9 female        0     0
```

432 Class 02 | 2026-01-15 | <https://thomaselove.github.io/432-2026/>

87

Single Imputation with `mice`

```
1 smt_im <- mice(smt, m = 1, seed = 20260115, print = FALSE) |>
2   complete() |>
3   tibble()
```

Note

You may get a logged event for the ID variable expressed as a character, and that can be ignored.

```
1 prop_miss_case(smt_im)
```

```
[1] 0
```

```
1 dim(smt_im)
```

```
[1] 894   9
```

Saving the tidied data

Let's save both the unimputed and the imputed tidy data as R data sets.

```
1 write_rds(smt, "c02/data/smt.Rds")
2
3 write_rds(smt_im, "c02/data/smt_im.Rds")
```

To reload these files, we'll use `read_rds()`.

- The main advantage here is that we've saved the whole R object, including all characteristics that we've added since the original download.