

432 Class 4 Slides

github.com/THOMASELOVE/2020-432

2020-01-23

Today's Agenda

- ① Continuing our discussion of two-way ANOVA and ANCOVA with binary factors
- ② Building a two-factor ANOVA model with multi-categorical factors
 - again, focus on interpreting the interaction
 - add covariates, as desired

Setup

```
library(here); library(magrittr); library(janitor)
library(broom); library(simputation); library(patchwork)
library(naniar); library(visdat)
library(tidyverse)

theme_set(theme_bw())

smart1 <- readRDS(here("data/smart1.Rds"))
smart1_sh <- readRDS(here("data/smart1_sh.Rds"))
```

smart1_sh Variables, by Type

Variable	Type	Description
landline	Binary (1/0)	survey conducted by landline? (vs. cell)
healthplan	Binary (1/0)	subject has health insurance?
age_imp	Quantitative	age (imputed from groups - see Notes)
fruit_day	Quantitative	mean servings of fruit / day
drinks_wk	Quantitative	mean alcoholic drinks / week
bmi	Quantitative	body-mass index (in kg/m ²)
physhealth	Count (0-30)	of last 30 days, # in poor physical health
dm_status	Categorical	diabetes status (now 2 levels)
activity	Categorical	physical activity level (4 levels)
smoker	Categorical	tobacco use status (now 3 levels)
genhealth	Categorical	self-reported overall health (5 levels)

Section 1

Modeling with ANOVA for Binary Factors

Models we have fit (so far)

```
a1 <- smart1_sh %$% lm(bmi ~ dm_status)

a2 <- smart1_sh %$% lm(bmi ~ dm_status * healthplan)

a2_noint <- smart1_sh %$% lm(bmi ~ dm_status + healthplan)
```

Is the interaction term important here?

- ① Does the interaction plot display important non-parallelism?
- ② Does the interaction term account for a substantial fraction of the variation in our outcome?
- ③ Does the interaction term's estimate/standard error/uncertainty interval meet usual standards for statistical significance?

If **all** of these things are true, then it's easy to conclude that the interaction is important, and we cannot interpret the main effects of `dm_status` and `healthplan` without thinking first about the interaction of those two factors.

- So let's walk through the decision. I've repeated the interaction plot on the next slide.

Interaction Plot

We'll plot the means of the bmi in the four combinations:

- two levels of dm_status combined with
- two levels of healthplan

```
summaries1 <- smart1_sh %>%
  group_by(dm_status, healthplan) %>%
  summarize(n = n(), mean = mean(bmi), stdev = sd(bmi))

summaries1 %>% knitr::kable(digits = 2)
```

dm_status	healthplan	n	mean	stdev
Yes	0	35	31.02	6.93
Yes	1	1067	32.01	7.31
No	0	364	28.01	6.41
No	1	5946	28.01	6.01

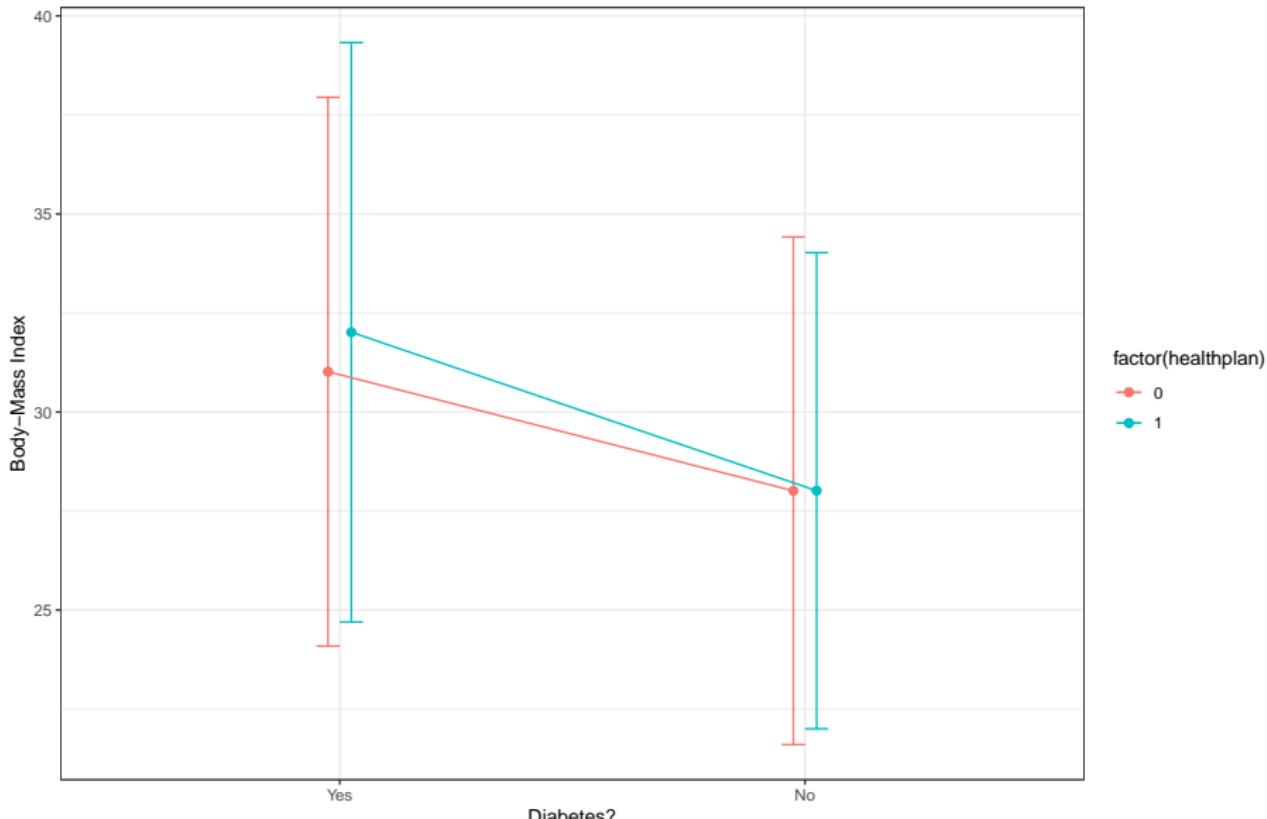
Interaction Plot for Two-Way ANOVA (code)

```
pd <- position_dodge(0.1)
ggplot(summaries1, aes(x = dm_status, y = mean,
                      col = factor(healthplan))) +
  geom_errorbar(aes(ymin = mean - stdev,
                     ymax = mean + stdev),
                 width = 0.1, position = pd) +
  geom_point(size = 2, position = pd) +
  geom_line(aes(group = healthplan), position = pd) +
  labs(y = "Body-Mass Index",
       x = "Diabetes?",
       title = "Observed Means (+/- SD) for BMI",
       subtitle = "by Diabetes Status and Insurance")
```

Interaction Plot for Two-Way ANOVA

Observed Means (+/- SD) for BMI

by Diabetes Status and Insurance



Evaluation in our Two-Way ANOVA of Interaction

- ① Does the interaction plot display important non-parallelism?
 - No, I don't think so.
- ② Does the interaction term account for a substantial fraction of the variation in our outcome?

```
anova(a2) %>% knitr::kable(digits = 0)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dm_status	1	14775	14775	380	0
healthplan	1	3	3	0	1
dm_status:healthplan	1	30	30	1	0
Residuals	7408	288338	39	NA	NA

- $SS(\text{total}) = 288,338 + 30 + 3 + 14,775 = 303,146$.
- $SS(\text{interaction}) = 30$
- $\eta^2(\text{interaction}) = \frac{30}{303146} = .000099$, or about 0.01% of `bmi` variation.

Is the interaction term important here?

- ① Does the interaction plot display important non-parallelism?
 - No.
- ② Does the interaction term account for a substantial fraction of the variation in our outcome?
 - It accounts for just under 0.01% of variation, so no.
- ③ Does the interaction term's estimate/standard error/uncertainty interval meet usual standards for statistical significance?

```
tidy(a2, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  knitr::kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	31.017	1.055	29.282	32.752
dm_statusNo	-3.006	1.104	-4.823	-1.190
healthplan	0.996	1.072	-0.767	2.759
dm_statusNo:healthplan	-0.994	1.123	-2.842	0.855

Is the interaction term important here?

- ① Does the interaction plot display important non-parallelism?
 - No.
- ② Does the interaction term account for a substantial fraction of the variation in our outcome?
 - No.
- ③ Does the interaction term's estimate/standard error/uncertainty interval meet usual standards for statistical significance?
 - No.

It's clearly easier to ignore the interaction term (and fit the no-interaction model) if none of these three things are true.

Interpreting the “No Interaction” Model

```
tidy(a2_noint, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate, std.error, conf.low, conf.high) %>%  
  knitr::kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	31.893	0.364	31.295	32.491
dm_statusNo	-3.966	0.204	-4.301	-3.631
healthplan	0.091	0.321	-0.437	0.620

- If Harry and Sally have the same healthplan status, but only Harry has diabetes, then Harry's BMI is estimated to be 3.97 kg/m^2 higher than Sally's. (90% uncertainty interval: 3.63, 4.30).
- If Harry and Sally have the same dm_status but Harry has a health plan and Sally doesn't, our model will estimate Harry's BMI as 0.09 kg/m^2 higher than Sally's (90% interval: -0.44, 0.62).

Adding a covariate

We saw that the no-interaction model might well be sufficient for BMI as a function of dm_status and healthplan. Would this still be true if we first adjusted for the impact of a continuous covariate, like physhealth, that is meaningfully correlated with BMI?

```
a3 <- smart1_sh %$%
  lm(bmi ~ physhealth + dm_status * healthplan)

anova(a3) %>% knitr::kable(digits = 1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
physhealth	1	4986.2	4986.2	129.2	0.0
dm_status	1	12185.9	12185.9	315.7	0.0
healthplan	1	0.3	0.3	0.0	0.9
dm_status:healthplan	1	22.1	22.1	0.6	0.4
Residuals	7407	285952.2	38.6	NA	NA

Model without the Covariate

Compare that ANOVA table to this one for our interaction model without the covariate. What changes?

```
anova(a2) %>% knitr::kable(digits = 1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dm_status	1	14774.8	14774.8	379.6	0.0
healthplan	1	3.1	3.1	0.1	0.8
dm_status:healthplan	1	30.4	30.4	0.8	0.4
Residuals	7408	288338.2	38.9	NA	NA

a3 covariate model without interaction term

```
a3_noint <- smart1_sh %$%
  lm(bmi ~ physhealth + dm_status + healthplan)

anova(a3_noint) %>% knitr::kable(digits = 1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
physhealth	1	4986.2	4986.2	129.2	0.0
dm_status	1	12185.9	12185.9	315.7	0.0
healthplan	1	0.3	0.3	0.0	0.9
Residuals	7408	285974.3	38.6	NA	NA

Interpreting “No Interaction” Model + Covariate

```
tidy(a3_noint, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate, std.error, conf.low, conf.high) %>%  
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	31.39	0.37	30.79	32.00
physhealth	0.06	0.01	0.05	0.07
dm_statusNo	-3.67	0.21	-4.01	-3.33
healthplan	0.03	0.32	-0.50	0.56

- If Harry and Sally have the same healthplan status and the same physhealth, but only Harry has diabetes, then Harry's BMI is estimated to be 3.67 kg/m^2 higher than Sally's. (90% uncertainty interval: 3.33, 4.01).
- See next slide, too.

Interpreting “No Interaction” Model + Covariate

```
tidy(a3_noint, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate, std.error, conf.low, conf.high) %>%  
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	31.39	0.37	30.79	32.00
physhealth	0.06	0.01	0.05	0.07
dm_statusNo	-3.67	0.21	-4.01	-3.33
healthplan	0.03	0.32	-0.50	0.56

- If Harry and Sally have the same dm_status and the same physhealth, but Harry has a health plan and Sally doesn't, our model will estimate Harry's BMI as 0.03 kg/m² higher than Sally's (90% uncertainty interval: -0.50, 0.56).
- Why aren't I talking here about the covariate's effect?

Does the model fit the data well?

We have the usual strategies applicable in any linear model:

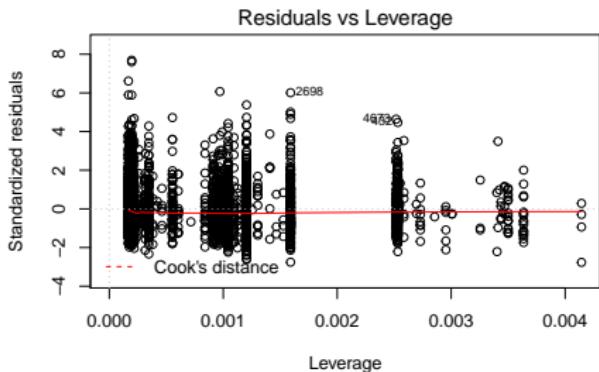
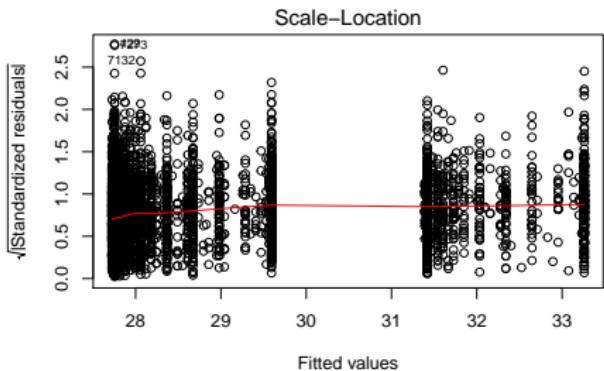
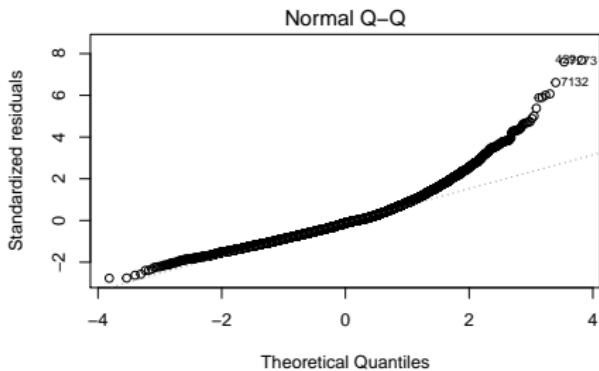
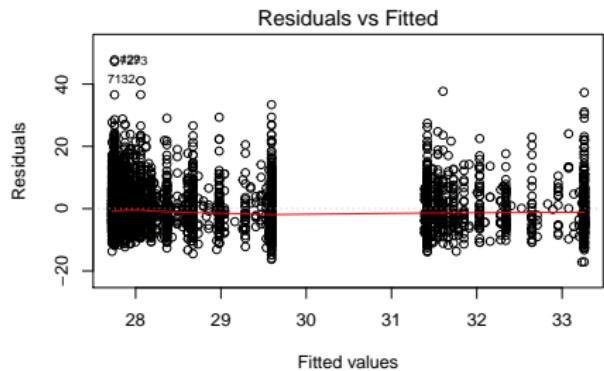
- evaluate the R^2 and other summary statistics, especially in comparison to alternative specifications of models for the same outcome.
- evaluate the fit of the model to regression assumptions, mostly through diagnostics based on residuals
- cross-validate our model selection process, perhaps by partitioning the sample into a training sample (where candidate models are developed) and a holdout / test sample (where we choose between the candidates)

Summary Statistics (Whole Sample)

```
bind_rows(glance(a1), glance(a2_noint), glance(a3_noint)) %>%  
  mutate(model =  
    c("dm_status", "+ healthplan", "+ physhealth")) %>%  
  select(model, r.squared, sigma, AIC, BIC, adj.r.squared) %>%  
  knitr::kable(digits = 3)
```

model	r.squared	sigma	AIC	BIC	adj.r.squared
dm_status	0.049	6.238	48176.79	48197.52	0.049
+ healthplan	0.049	6.239	48178.71	48206.35	0.048
+ physhealth	0.057	6.213	48118.91	48153.46	0.056

plot(a3_noint)



Section 2

ANOVA and ANCOVA with Multi-Categorical Predictors in Linear Models

New Questions

- ① How does a subject's self-reported general health and their tobacco status combine when predicting their body mass index?
- ② Does adjusting for the number of alcoholic drinks consumed per week affect our assessment?

Addressing Question 1: Simple Summary

- ① How does a subject's genhealth and smoker status combine when predicting their body mass index?

```
smart1_sh %$%
```

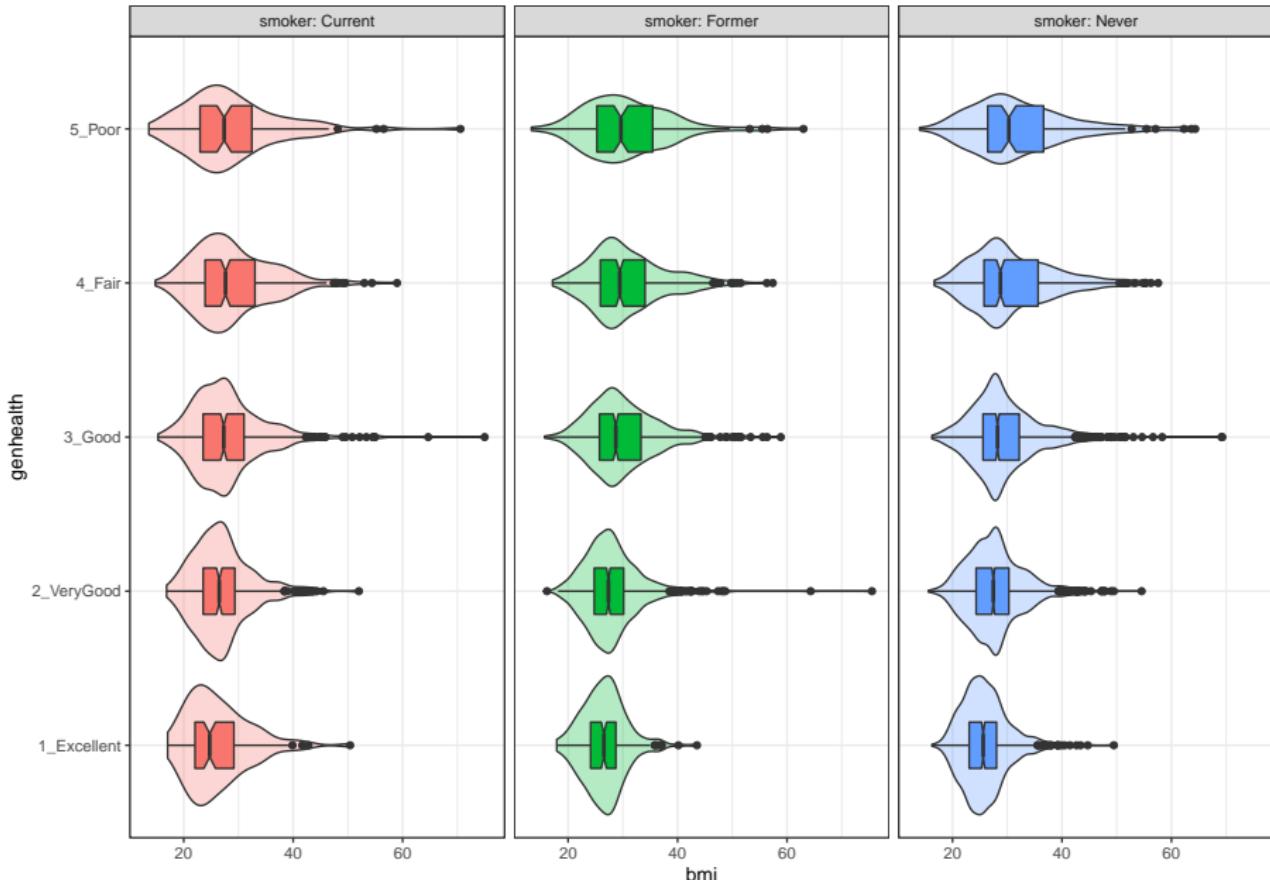
```
mosaic::favstats(bmi ~ smoker + genhealth) %>%
  rename(smoke.health = smoker.genhealth) %>%
  knitr::kable(digits = 1)
```

smoke.health	min	Q1	median	Q3	max	mean	sd
Current.1_Excellent	17.1	22.1	24.8	29.2	50.4	26.2	5.9
Former.1_Excellent	17.9	24.1	26.5	28.7	43.5	26.6	4.0
Never.1_Excellent	16.2	23.1	25.6	28.0	49.5	26.0	4.3
Current.2_VeryGood	16.9	23.6	26.5	29.4	52.0	27.0	5.4
Former.2_VeryGood	16.1	24.8	27.4	30.1	75.5	28.0	5.5
Never.2_VeryGood	15.6	24.3	27.5	30.3	54.6	27.8	5.3
Current.3_Good	15.3	23.5	27.3	31.0	75.0	28.1	7.1
Former.3_Good	15.7	25.7	28.7	33.3	58.9	30.0	6.4
Never.3_Good	16.2	25.6	28.2	32.2	69.3	29.4	6.4

Visualize Three Variables (Code)

```
ggplot(smart1_sh, aes(x = genhealth, y = bmi,  
                      fill = smoker)) +  
  geom_violin(alpha = 0.3) +  
  geom_boxplot(width = 0.3, notch = TRUE) +  
  facet_wrap(~ smoker, labeller = label_both) +  
  coord_flip() +  
  guides(fill = FALSE)
```

Visualize Three Variables



Interaction Plot

We'll plot the means of the bmi in the fifteen combinations:

- three levels of smoker combined with
- five levels of genhealth

```
summaries4 <- smart1_sh %>%
  group_by(genhealth, smoker) %>%
  summarize(n = n(), mean = mean(bmi), stdev = sd(bmi))

summaries4 %>% knitr::kable(digits = 2)
```

genhealth	smoker	n	mean	stdev
1_Excellent	Current	111	26.25	5.89
1_Excellent	Former	232	26.64	3.99
1_Excellent	Never	714	25.95	4.30
2_VeryGood	Current	332	26.99	5.42
2_VeryGood	Former	635	28.04	5.52
2_VeryGood	Never	1453	27.79	5.25
2_Good	Current	112	28.19	5.19
2_Good	Former	232	28.50	5.19
2_Good	Never	714	28.19	5.19
2_Fair	Current	112	28.19	5.19
2_Fair	Former	232	28.50	5.19
2_Fair	Never	714	28.19	5.19
2_Poor	Current	112	28.19	5.19
2_Poor	Former	232	28.50	5.19
2_Poor	Never	714	28.19	5.19

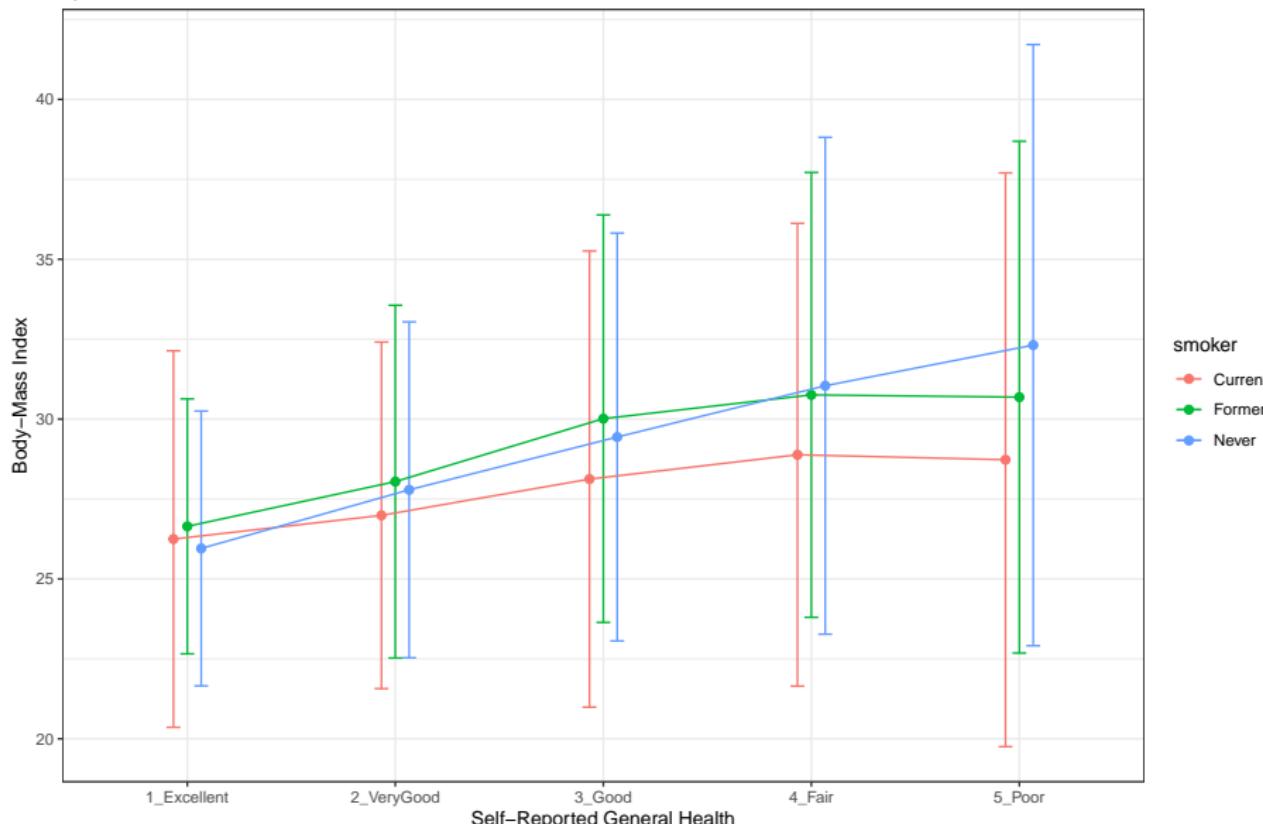
Interaction Plot for Two-Way ANOVA (code)

```
pd <- position_dodge(0.2)
ggplot(summaries4, aes(x = genhealth, y = mean,
                       col = smoker)) +
  geom_errorbar(aes(ymin = mean - stdev,
                     ymax = mean + stdev),
                width = 0.2, position = pd) +
  geom_point(size = 2, position = pd) +
  geom_line(aes(group = smoker), position = pd) +
  labs(y = "Body-Mass Index",
       x = "Self-Reported General Health",
       title = "Observed Means (+/- SD) for BMI",
       subtitle = "by General Health and Tobacco Status")
```

Interaction Plot for Two-Way ANOVA

Observed Means (+/- SD) for BMI

by General Health and Tobacco Status



Two-Way Analysis of Variance

```
a4 <- smart1_sh %$% lm(bmi ~ genhealth * smoker)
```

```
anova(a4) %>% knitr::kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	14911.707	3727.927	96.724	0.000
smoker	2	2198.101	1099.050	28.516	0.000
genhealth:smoker	8	943.286	117.911	3.059	0.002
Residuals	7397	285093.553	38.542	NA	NA

Model a4 tidied coefficients

```
tidy(a4, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high, p.value)
knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	26.25	0.59	25.28	27.22
genhealth2_VeryGood	0.74	0.68	-0.38	1.86
genhealth3_Good	1.88	0.66	0.79	2.97
genhealth4_Fair	2.64	0.69	1.50	3.78
genhealth5_Poor	2.48	0.80	1.16	3.80
smokerFormer	0.40	0.72	-0.78	1.58
smokerNever	-0.29	0.63	-1.34	0.76
genhealth2_VeryGood:smokerFormer	0.66	0.83	-0.71	2.00
genhealth3_Good:smokerFormer	1.49	0.81	0.16	2.82
genhealth4_Fair:smokerFormer	1.48	0.86	0.06	2.90
genhealth5_Poor:smokerFormer	1.56	1.03	-0.14	2.72
genhealth2_VeryGood:smokerNever	1.09	0.74	-0.12	2.22

The Equations

The model with the interaction term is

$$\begin{aligned} \text{BMI} = & 26.25 + 0.74 (\text{genhealth} = \text{Very Good}) \\ & + 1.88 (\text{genhealth} = \text{Good}) \\ & + \dots \\ & + 2.48 (\text{genhealth} = \text{Poor}) \\ & + 0.40 (\text{smoker} = \text{Former}) \\ & - 0.29 (\text{smoker} = \text{Never}) \\ & + 0.66 (\text{genhealth} = \text{Very Good})(\text{smoker} = \text{Former}) \\ & + \dots \\ & + 1.09 (\text{genhealth} = \text{Very Good})(\text{smoker} = \text{Never}) \\ & + \dots \\ & + 3.88 (\text{genhealth} = \text{Poor})(\text{smoker} = \text{Never}) \end{aligned}$$

- Predict Harry, who's in Excellent health but a Current smoker
- Predict Sally, who's in Very Good Health and a Former smoker

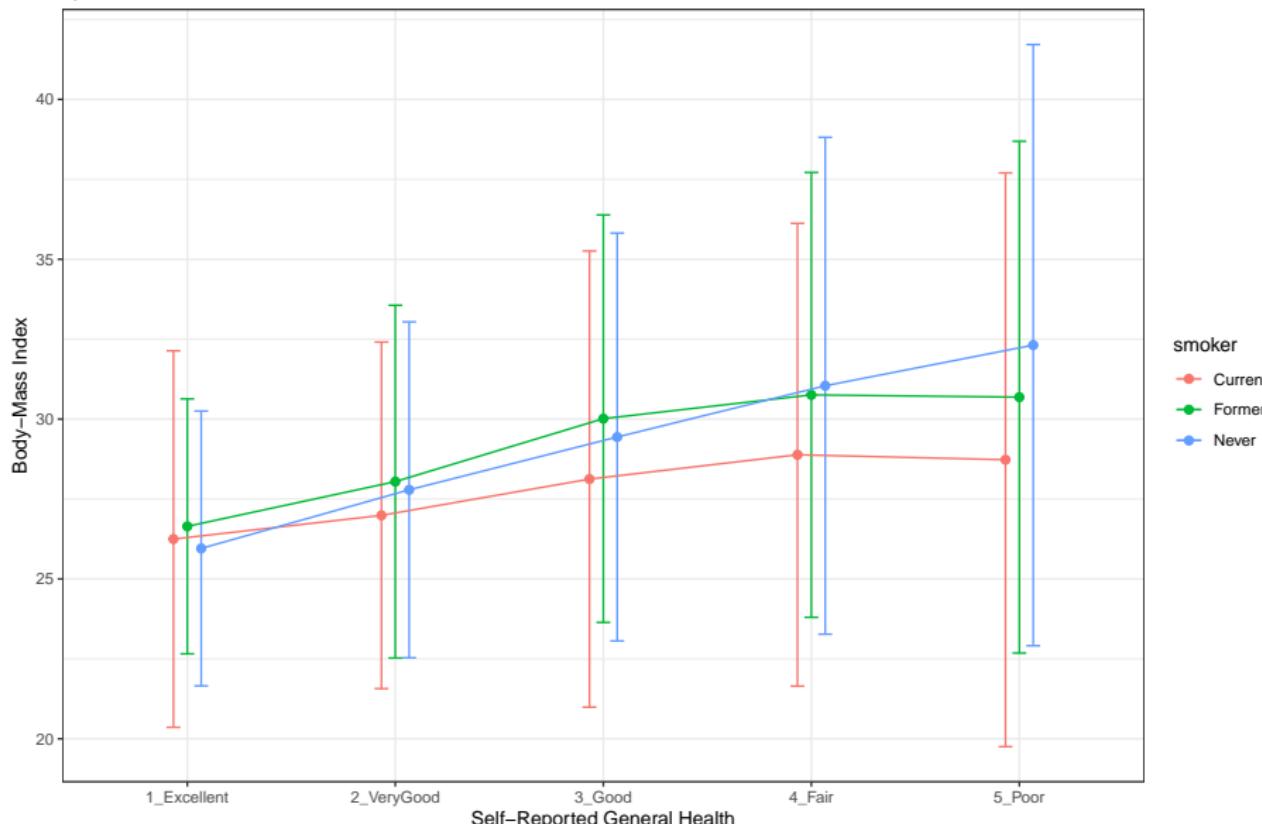
Is the interaction term important here?

- ① Does the interaction plot display important non-parallelism?
- ② Does the interaction term account for a substantial fraction of the variation in our outcome?
- ③ Does the interaction term's estimate/standard error/uncertainty interval meet usual standards for statistical significance?
- See the next 3 slides for the answers...

Interaction Plot, again...

Observed Means (+/- SD) for BMI

by General Health and Tobacco Status



Fraction of Variation accounted for by Interaction

```
anova(a4) %>% knitr::kable(digits = 0)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	14912	3728	97	0
smoker	2	2198	1099	29	0
genhealth:smoker	8	943	118	3	0
Residuals	7397	285094	39	NA	NA

- SS(total) = 14,912 + 2,198 + 943 + 285,094 = 303,147.
- SS(interaction) = 943
- $\eta^2(\text{interaction}) = \frac{943}{303147} = .0031$, or about 0.31% of bmi variation.

Are the interaction terms statistically significant?

```
a4 <- smart1_sh %$% lm(bmi ~ genhealth * smoker)  
a4_noint <- smart1_sh %$% lm(bmi ~ genhealth + smoker)  
  
anova(a4_noint, a4)
```

Analysis of Variance Table

Model 1: bmi ~ genhealth + smoker

Model 2: bmi ~ genhealth * smoker

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	7405	286037				
2	7397	285094	8	943.29	3.0593	0.00193 **

Signif. codes:

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

So which model should we use? (Interaction or No Interaction)

Equation for the Interaction Model (a4)

bmi = 26.248
+ 0.741 (genhealth = Very Good)
+ 1.875 (genhealth = Good)
+ 2.637 (genhealth = Fair)
+ 2.481 (genhealth = Poor)
+ 0.397 (smoker = Former)
- 0.293 (smoker = Never)
+ 0.658 (Very Good)(Former)
+ 1.493 (Good)(Former)
+ 1.475 (Fair)(Former)
+ 1.560 (Poor)(Former)
+ 1.093 (Very Good)(Never)
+ 1.610 (Good)(Never)
+ 2.452 (Fair)(Never)
+ 3.878 (Poor)(Never)

Comparing Harry and Sally (interaction model)

Scenario	Subject	genhealth	smoker
1	Harry	Very Good	Current
1	Sally	Very Good	Never

- Harry's predicted BMI is $26.248 + 0.741 = 26.989$
- Sally's predicted BMI is $26.248 + 0.741 - 0.293 + 1.093 = 27.789$
- If genhealth Very Good, effect of Never vs. Current is 0.800

Scenario	Subject	genhealth	smoker
2	Harry	Poor	Current
2	Sally	Poor	Never

- Harry's predicted BMI is $26.248 + 2.481 = 28.729$
- Sally's predicted BMI is $26.248 + 2.481 - 0.293 + 3.878 = 32.314$
- If genhealth Poor, effect of Never vs. Current is 3.585

Comparing Harry and Sally (interaction model)

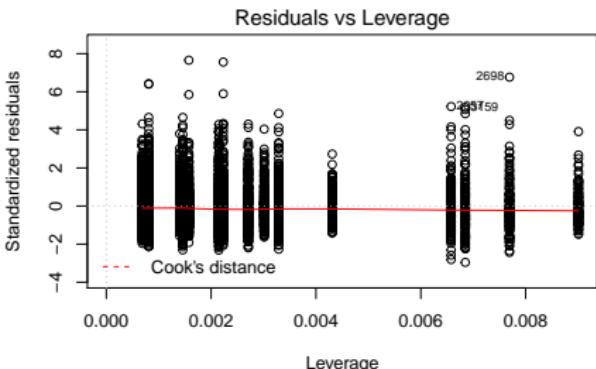
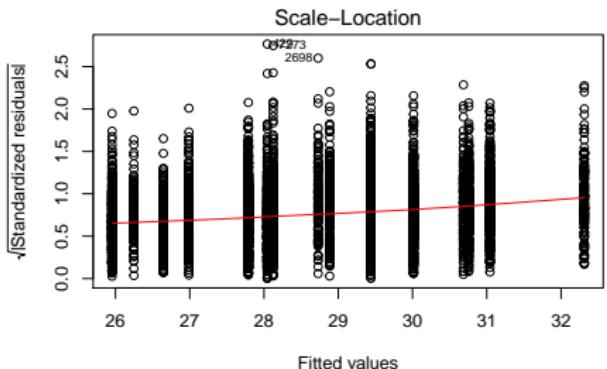
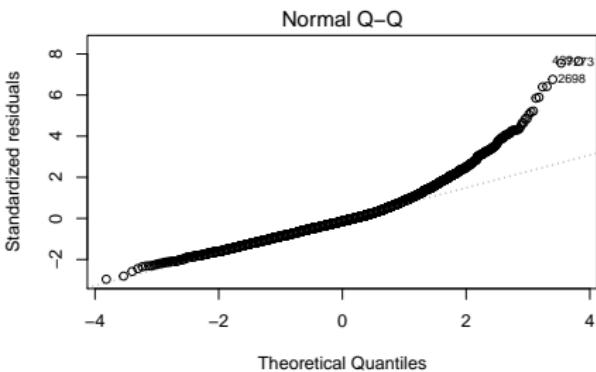
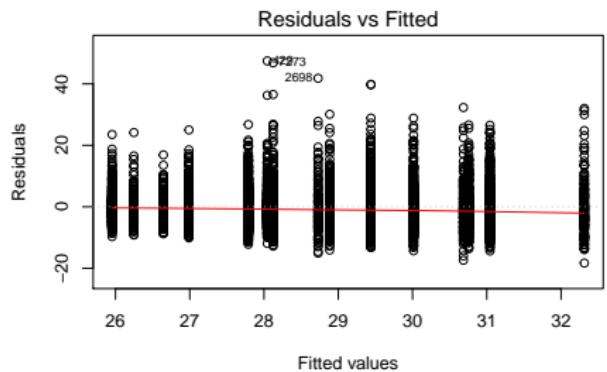
Scenario	Subject	genhealth	smoker
3	Harry	Very Good	Current
3	Sally	Poor	Current

- Harry's predicted BMI is $26.248 + 0.741 = 26.989$
- Sally's predicted BMI is $26.248 + 2.481 = 28.729$
- If Current smoker, effect of Poor vs. Very Good is 1.740

Scenario	Subject	genhealth	smoker
4	Harry	Very Good	Never
4	Sally	Poor	Never

- Harry's predicted BMI is $26.248 + 0.741 - 0.293 + 1.093 = 27.789$
- Sally's predicted BMI is $26.248 + 2.481 - 0.293 + 3.878 = 32.314$
- If Never smoker, effect of Poor vs. Very Good is 4.525

Residual Plots for model a4

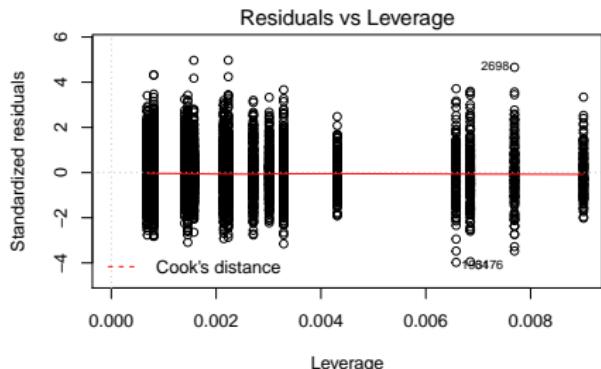
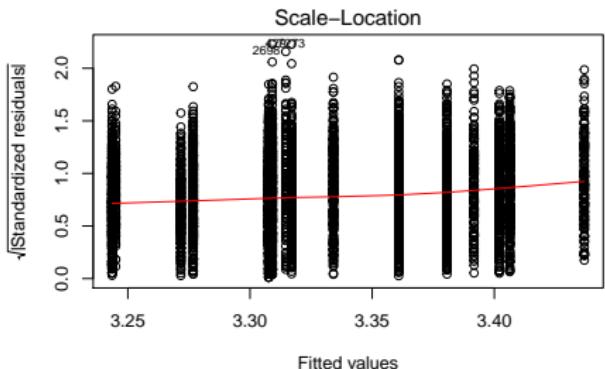
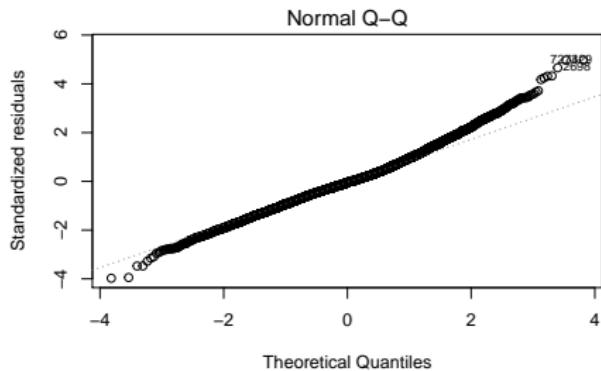
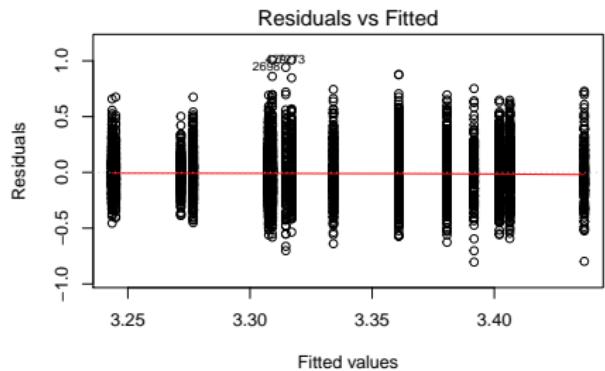


Would using log(BMI) make the difference?

```
a4_log <- smart1_sh %$% lm(log(bmi) ~ genhealth * smoker)  
  
anova(a4_log) %>% knitr::kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	14.651	3.663	88.951	0.000
smoker	2	3.149	1.575	38.242	0.000
genhealth:smoker	8	0.933	0.117	2.833	0.004
Residuals	7397	304.582	0.041	NA	NA

Residual Plots for model a4_log



What if we add a covariate?

```
smart1_sh <- smart1_sh %>%  
  mutate(drinks_c = drinks_wk - mean(drinks_wk))  
  
a5_log <- smart1_sh %$%  
  lm(log(bmi) ~ drinks_c + genhealth * smoker)  
  
anova(a5_log) %>% knitr::kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
drinks_c	1	1.239	1.239	30.148	0.000
genhealth	4	14.229	3.557	86.545	0.000
smoker	2	2.895	1.448	35.220	0.000
genhealth:smoker	8	0.958	0.120	2.914	0.003
Residuals	7396	303.994	0.041	NA	NA

- ① Can we make predictions?
- ② Why center the `drinks_wk`?

Equation for model a5_log

log(BMI) = 3.248
- 0.0014 drinks_c
+ 0.033 (genhealth = Very Good)
+ 0.063 (genhealth = Good)
+ 0.087 (genhealth = Fair)
+ 0.067 (genhealth = Poor)
+ 0.025 (smoker = Former)
- 0.005 (smoker = Never)
+ 0.013 (Very Good)(Former)
+ 0.045 (Good)(Former)
+ 0.041 (Fair)(Former)
+ 0.052 (Poor)(Former)
+ 0.031 (Very Good)(Never)
+ 0.053 (Good)(Never)
+ 0.074 (Fair)(Never)
+ 0.125 (Poor)(Never)

Comparing Models

- Does the addition of the covariate add statistically detectable predictive value?

```
anova(a4_log, a5_log)
```

Analysis of Variance Table

Model 1: log(bmi) ~ genhealth * smoker

Model 2: log(bmi) ~ drinks_c + genhealth * smoker

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	7397	304.58				
2	7396	303.99	1	0.58817	14.31	0.0001563 ***

Signif. codes:

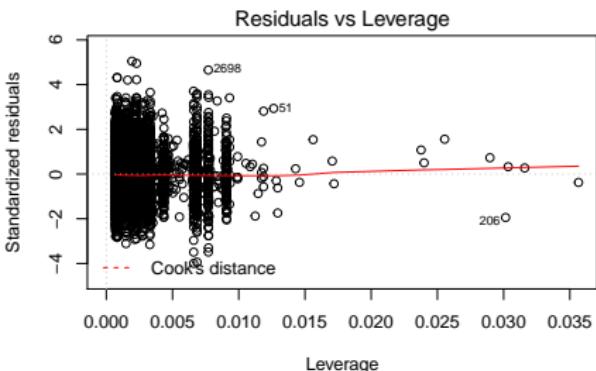
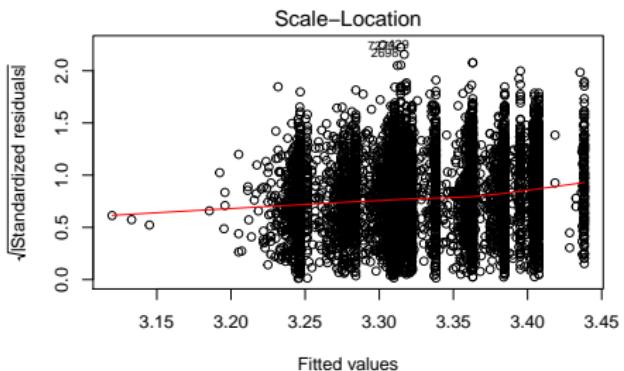
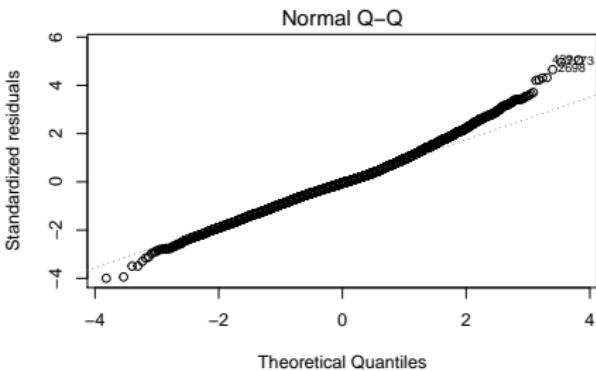
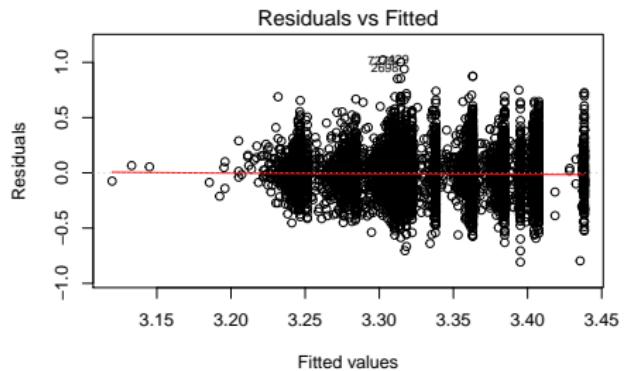
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comparing Models

```
bind_rows(glance(a4_log), glance(a5_log)) %>%
  mutate(model = c("ANOVA", "+ drinks_c")) %>%
  select(model, r2 = r.squared, sigma, AIC, BIC,
         adjr2 = adj.r.squared) %>%
knitr::kable(digits = c(0, 3, 3, 0, 0, 3))
```

model	r2	sigma	AIC	BIC	adjr2
ANOVA	0.058	0.203	-2592	-2482	0.056
+ drinks_c	0.060	0.203	-2604	-2487	0.058

Residual Plots for model a5_log



Next up

What if we have a binary outcome?