432 Class 4 Slides

github.com/THOMASELOVE/2020-432

2020-01-23

Today's Agenda

1. Building a two-factor ANOVA model with multi-categorical factors
   - again, focus on interpreting the interaction
   - add covariates, as desired
2. Building similar models for a binary outcome using linear probability models.
3. Building similar models for a binary outcome with generalized linear models (specifically logistic regression).

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)

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?

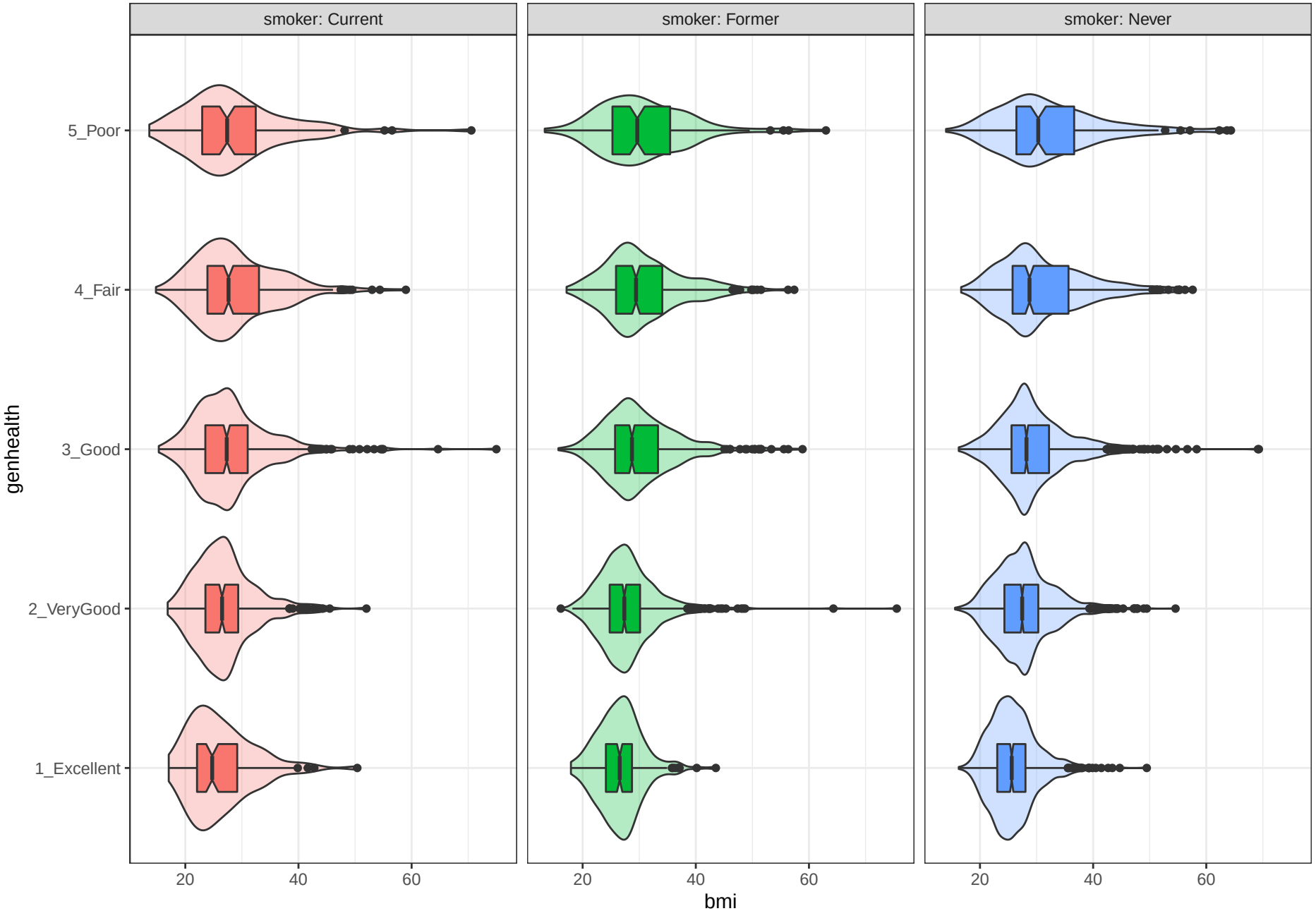
```
smart1_sh %$%
  mosaic::favstats(bmi ~ smoker + genhealth) %>%
  rename(smoke.health = smoker.genhealth) %>%
  knitr::kable(digits = 1)
```

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	1
Former.1_Excellent	17.9	24.1	26.5	28.7	43.5	26.6	4.0	2
Never.1_Excellent	16.2	23.1	25.6	28.0	49.5	26.0	4.3	7
Current.2_VeryGood	16.9	23.6	26.5	29.4	52.0	27.0	5.4	3
Former.2_VeryGood	16.1	24.8	27.4	30.1	75.5	28.0	5.5	6
Never.2_VeryGood	15.6	24.3	27.5	30.3	54.6	27.8	5.3	14
Current.3_Good	15.3	23.5	27.3	31.0	75.0	28.1	7.1	4
Former.3_Good	15.7	25.7	28.7	33.3	58.9	30.0	6.4	6
Never.3_Good	16.2	25.6	28.2	32.2	69.3	29.4	6.4	12

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`

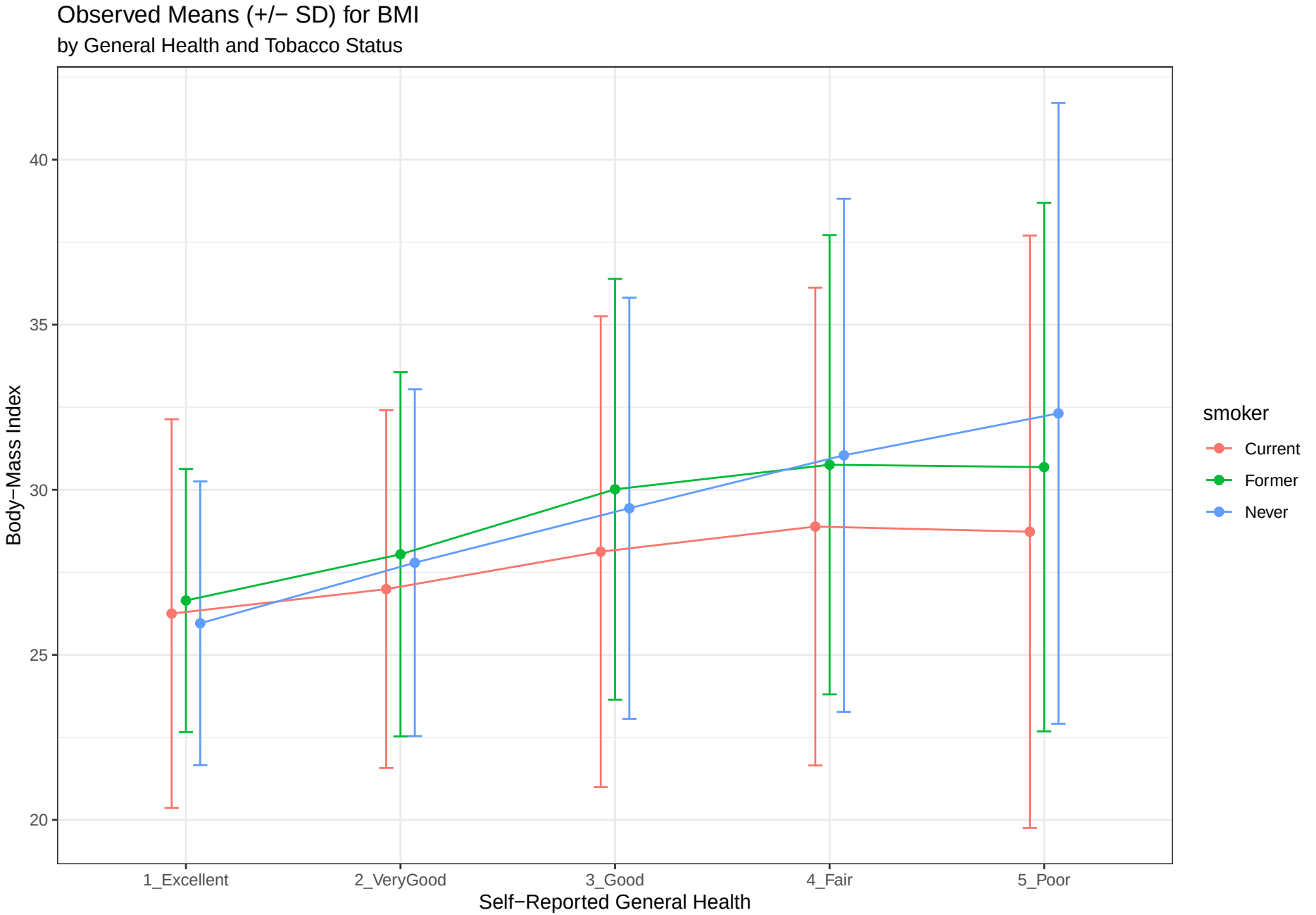
```
summaries4 <- smart1_sh %>%
  group_by(genhealth, smoker) %>%
  summarize(n = n(), mean = mean(bmi), stdev = sd(bmi))

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

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



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.val
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.
(Intercept)	26.25	0.59	25.28	2
genhealth2_VeryGood	0.74	0.68	-0.38	
genhealth3_Good	1.88	0.66	0.79	
genhealth4_Fair	2.64	0.69	1.50	
genhealth5_Poor	2.48	0.80	1.16	
smokerFormer	0.40	0.72	-0.78	
smokerNever	-0.29	0.63	-1.34	
genhealth2_VeryGood:smokerFormer	0.66	0.83	-0.71	
genhealth3_Good:smokerFormer	1.49	0.81	0.16	
genhealth4_Fair:smokerFormer	1.48	0.86	0.06	
genhealth5_Poor:smokerFormer	1.56	1.03	-0.14	
genhealth2_VeryGood:smokerNever	1.09	0.74	-0.12	

The Equations

The model with the interaction term is

```
BMI = 26.25 + 0.74 (genhealth = Very Good)
            + 1.88 (genhealth = Good)
            + ...
            + 2.48 (genhealth = Poor)
            + 0.40 (smoker = Former)
            - 0.29 (smoker = Never)
            + 0.66 (genhealth = Very Good)(smoker = Former)
            + ...
            + 1.09 (genhealth = Very Good)(smoker = Never)
            + ...
            + 3.88 (genhealth = Poor)(smoker = Never)
```

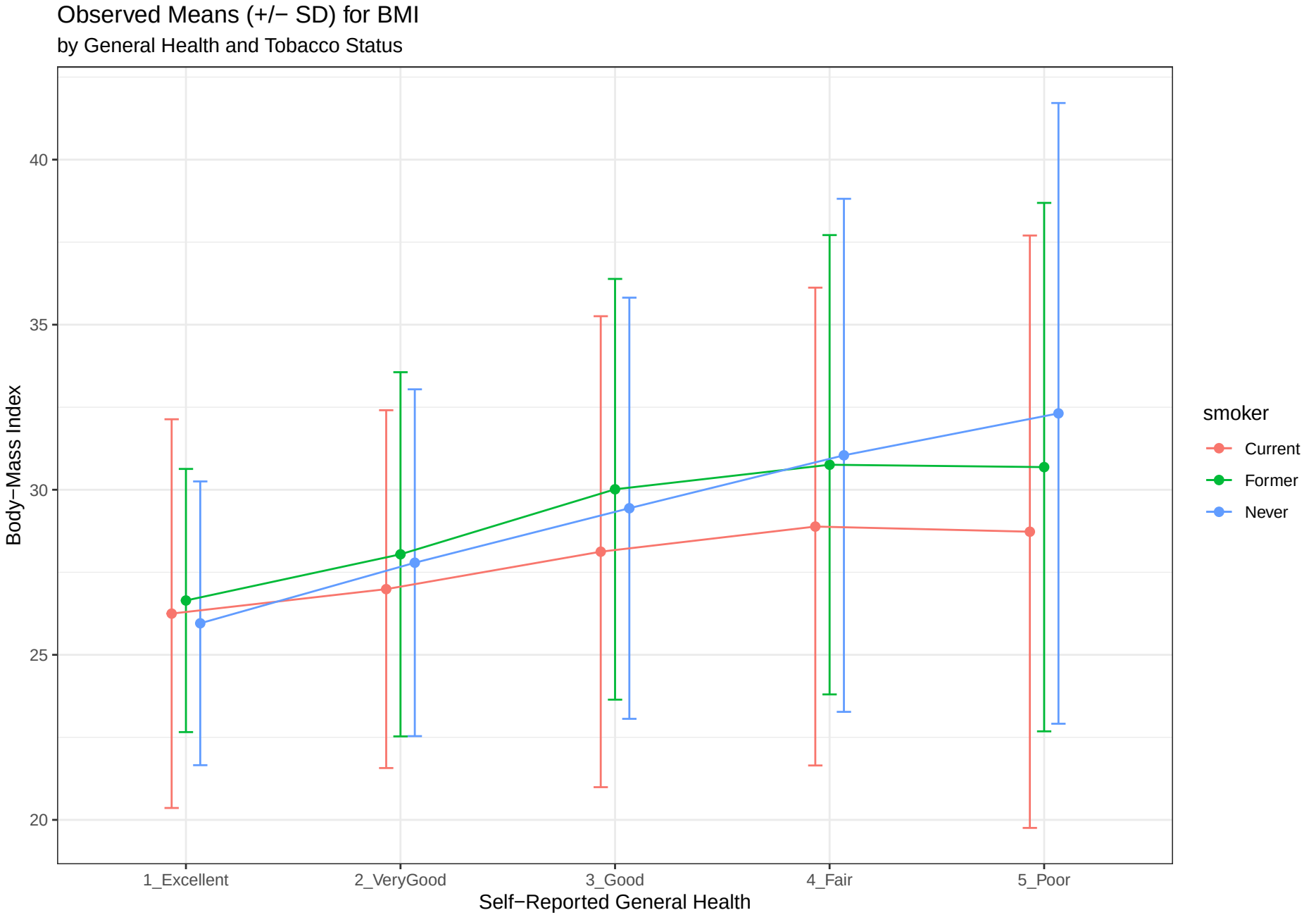
- Predict Harry, who's in Excellent health but a Current smoker
- Predict Sally, who's in Very Good Health and a Former smoker

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. . .



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
- η^2(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 + 0.658 = 27.354
- If genhealth Very Good, effect of Never vs. Current is 0.365

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)

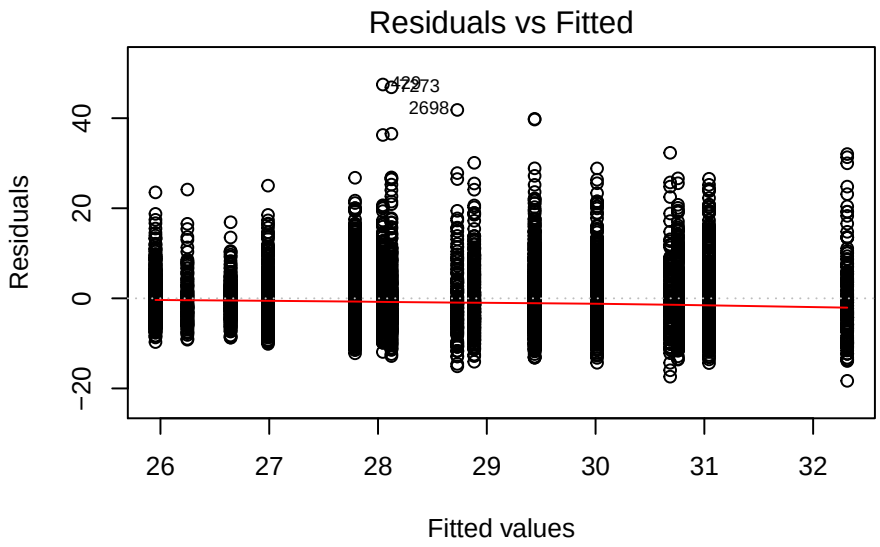
Scenario	Subject	genhealth	smoker
3	Harry	Very Good	Current
3	Sally	Poor	Current

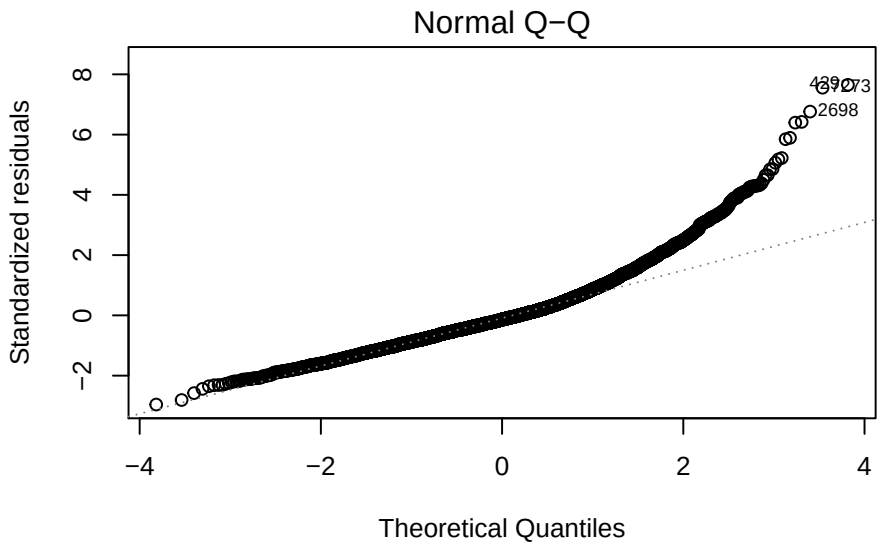
- Harry's predicted BMI is 26.248 + 0.741 = 26.989
- Sally's predicted BMI is 26.248 + 2.481 = 28.729
- If Current smoker, effect of Poor vs. Very Good is 1.740

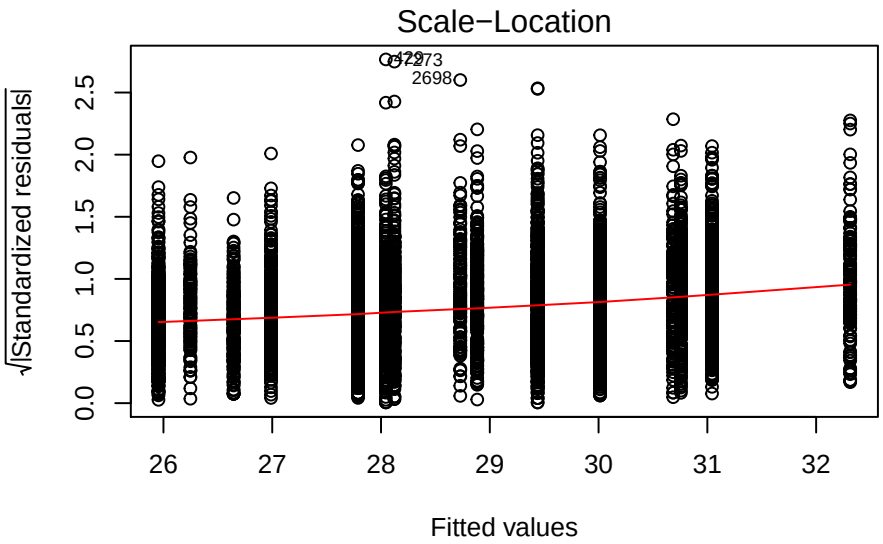
Scenario	Subject	genhealth	smoker
4	Harry	Very Good	Never
4	Sally	Poor	Never

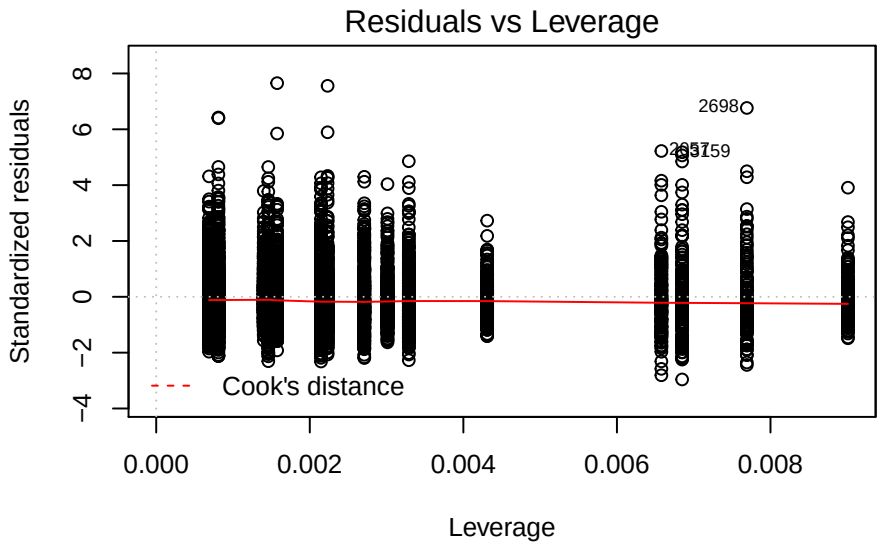
- Harry's predicted BMI is 26.248 + 0.741 - 0.293 + 1.093 = 27.789
- Sally's predicted BMI is 26.248 + 2.481 - 0.293 + 3.878 = 32.314
- If Never smoker, effect of Poor vs. Very Good is 4.525

Residual Plots for model a4







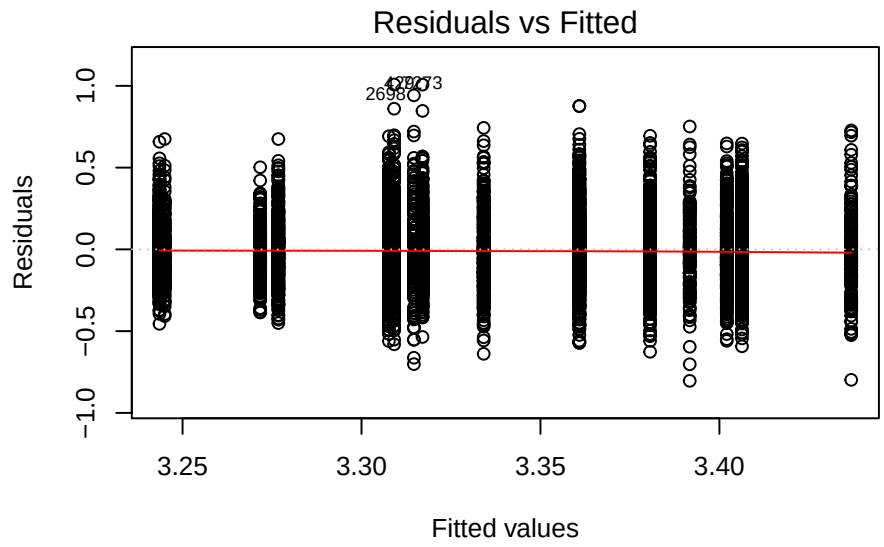


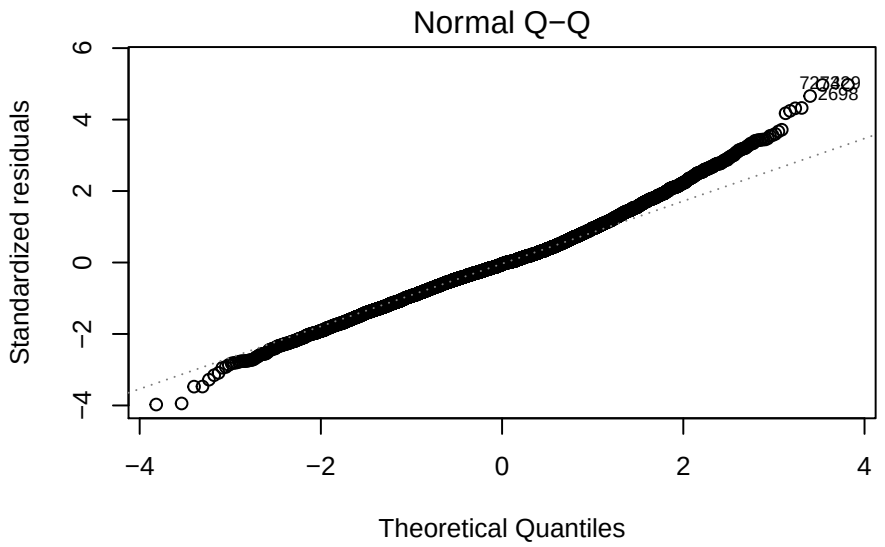
Would using `log(BMI)` make the difference?

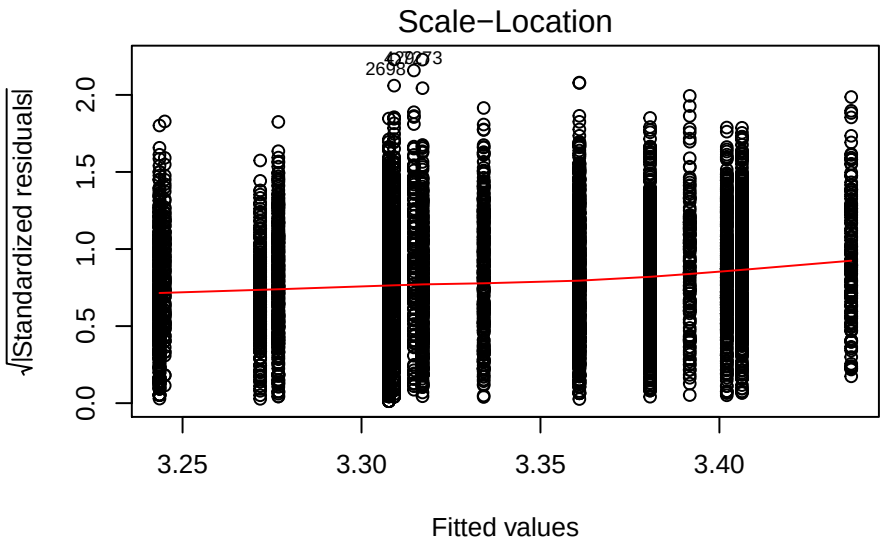
```
a4_log <- smart1_sh %$% lm(log(bmi) ~ genhealth * smoker)

anova(a4_log) %>% knitr::kable(digits = 3)
```

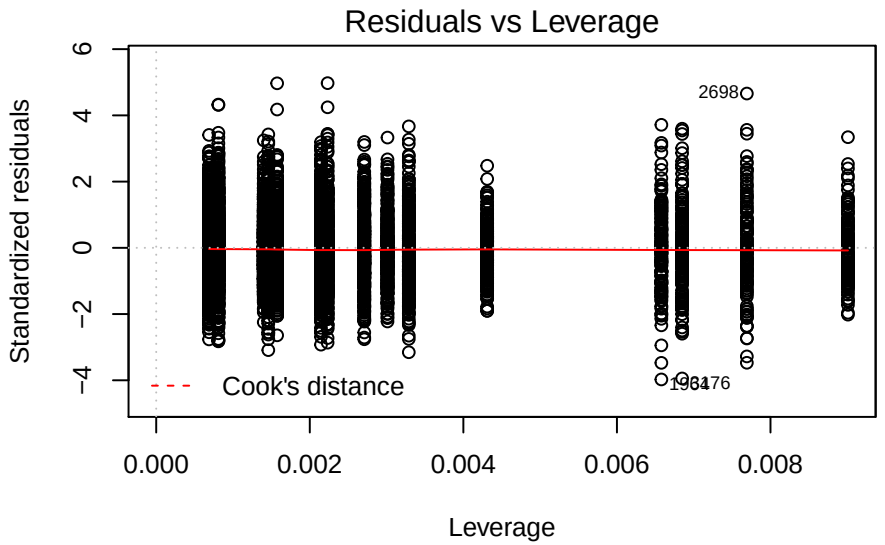
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	14.651	3.663	88.951	0.000
smoker	2	3.149	1.575	38.242	0.000
genhealth:smoker	8	0.933	0.117	2.833	0.004
Residuals	7397	304.582	0.041	NA	NA

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)
```

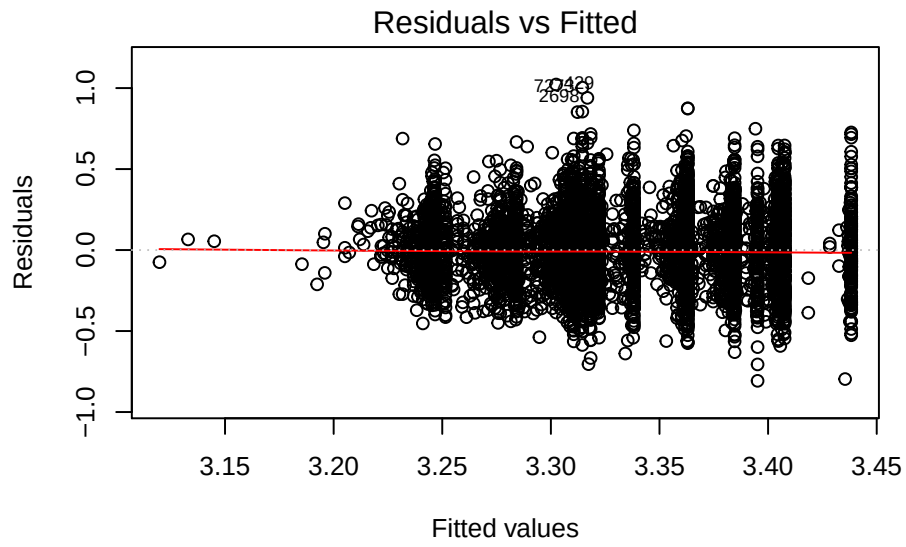
```
Analysis of Variance Table

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
```

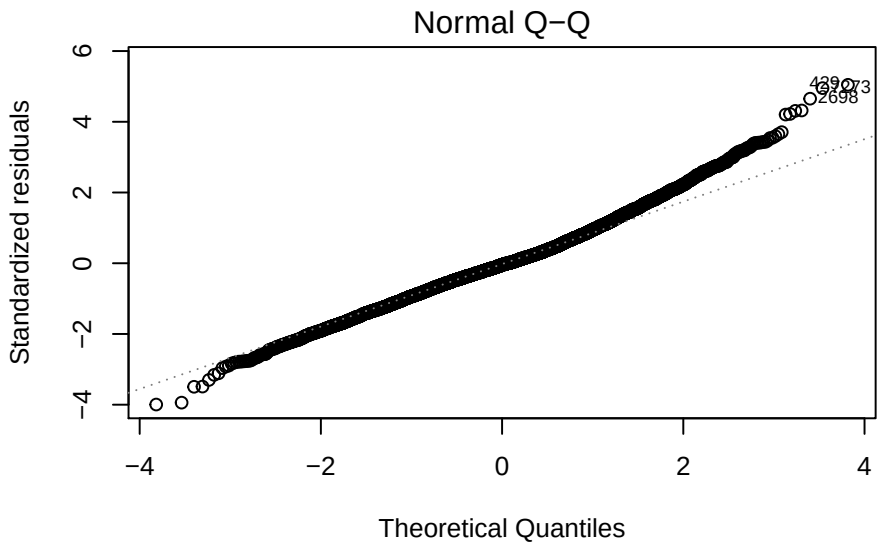
Comparing Models

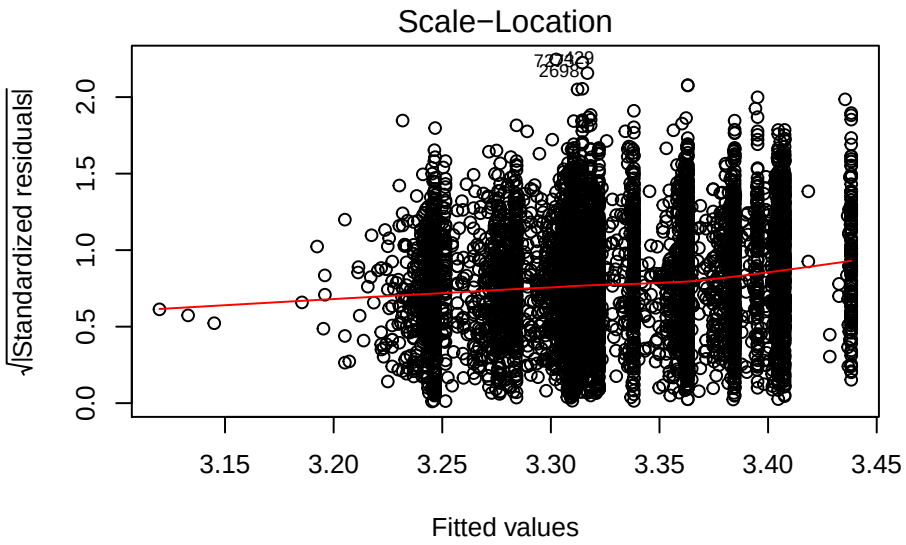
```
bind_rows(glance(a4_log), glance(a5_log)) %>%
  mutate(model = c("ANOVA", "+ drinks_c")) %>%
  select(model, r2 = r.squared, sigma, AIC, BIC,
         adjr2 = adj.r.squared) %>%
  knitr::kable(digits = c(0, 3, 3, 0, 0, 3))
```

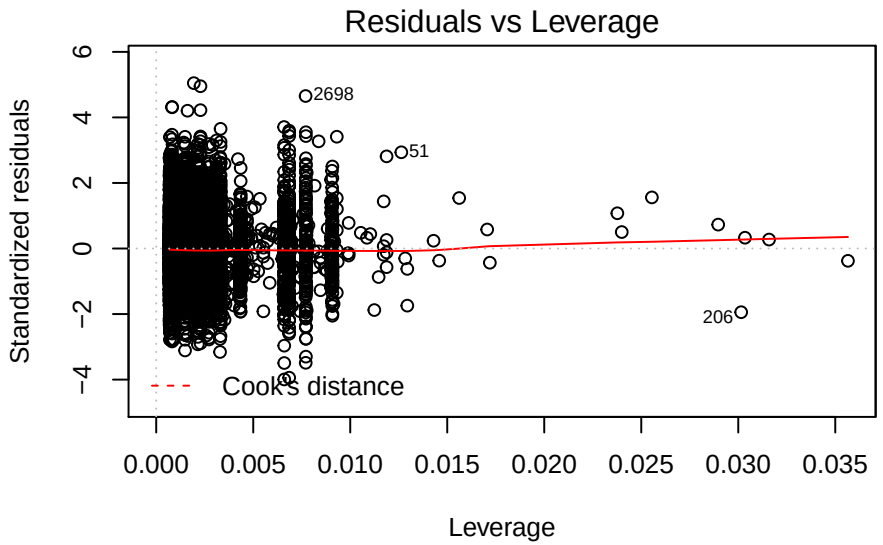
model	r2	sigma	AIC	BIC	adjr2
ANOVA	0.058	0.203	-2592	-2482	0.056
+ drinks_c	0.060	0.203	-2604	-2487	0.058

Residual Plots for model a5_log









What if we have a binary outcome?

Let's predict the probability that BMI < 30

```
smart1_sh <- smart1_sh %>%
  mutate(bmilt30 = as.numeric(bmi < 30),
         dm_status = fct_relevel(dm_status, "No"))

smart1_sh %>% tabyl(bmilt30) %>% adorn_pct_formatting()
```

```
 bmilt30    n percent
       0 2343   31.6%
       1 5069   68.4%
```

Linear Probability Model

- Create a binary (1/0) outcome.
- Run a linear regression model to predict the 1/0 value.
- Interpret the result as Prob(outcome = 1).

Any clear problems with this?

Two-Factor Linear Probability model for bmilt30

```
a6 <- smart1_sh %$%
  lm(bmilt30 ~ dm_status * genhealth)

anova(a6) %>% knitr::kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dm_status	1	56.215	56.215	276.320	0.000
genhealth	4	38.003	9.501	46.700	0.000
dm_status:genhealth	4	2.273	0.568	2.793	0.025
Residuals	7402	1505.867	0.203	NA	NA

Equation for model a6

```
tidy(a6) %>%
  select(term, estimate) %>% knitr::kable(digits = 3)
```

term	estimate
(Intercept)	0.847
dm_statusYes	-0.120
genhealth2_VeryGood	-0.090
genhealth3_Good	-0.193
genhealth4_Fair	-0.213
genhealth5_Poor	-0.189
dm_statusYes:genhealth2_VeryGood	-0.101
dm_statusYes:genhealth3_Good	-0.041
dm_statusYes:genhealth4_Fair	-0.047
dm_statusYes:genhealth5_Poor	-0.198

- Prediction for a subject without diabetes who is in Excellent Health?

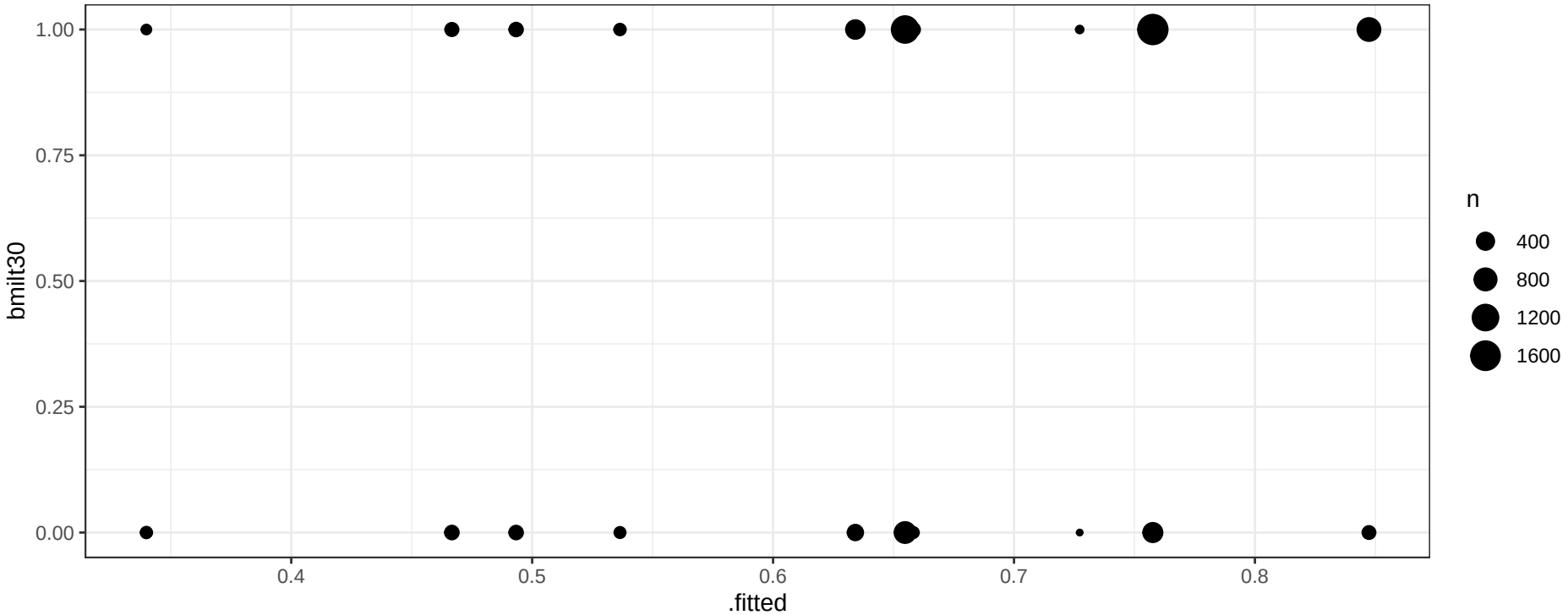
Get predictions for all subjects in our data

```
a6_aug <- augment(a6)

a6_aug %>% count(.fitted, dm_status, genhealth)
```

```
# A tibble: 10 x 4
   .fitted dm_status genhealth       n
     <dbl> <fct>     <fct>        <int>
 1   0.340 Yes       5_Poor         153
 2   0.467 Yes       4_Fair         360
 3   0.493 Yes       3_Good         375
 4   0.536 Yes       2_VeryGood     192
 5   0.634 No        4_Fair         779
 6   0.655 No        3_Good        1993
 7   0.658 No        5_Poor         275
 8   0.727 Yes       1_Excellent     22
 9   0.758 No        2_VeryGood    2228
10   0.847 No        1_Excellent   1035
```

Plot observed vs. predicted values

```
ggplot(a6_aug, aes(x = .fitted, y = bmilt30)) +
  geom_count()
```



Logistic Regression Model

- Create a binary (1/0) outcome.
- Run a logistic regression model to predict the logarithm of the odds that the outcome is 1.
- Exponentiate to describe the result in terms of odds(outcome = 1).

Remember that $odds(Y = 1) = \frac{Pr(Y=1)}{1+Pr(Y=1)}$.

Why is this helpful?

- log(odds(Y = 1)) or logit(Y = 1) covers all real numbers.
- Prob(Y = 1) is restricted to [0, 1].

How do we fit a simple logistic regression model?

```
a7 <- smart1_sh %$%
  glm(bmilt30 ~ dm_status, family = binomial(link = logit))
```

How do we interpret the coefficients?

```
tidy(a7) %>% select(term, estimate) %>%
  knitr::kable(digits = 3)
```

term	estimate
(Intercept)	0.946
dm_statusYes	-1.044

Equation: `logit(BMI < 30) = 0.946 - 1.044 (dm_status = Yes)`

How can we interpret this result?

Interpreting our Logistic Regression Equation

`logit(BMI < 30) = 0.946 - 1.044 (dm_status = Yes)`

- Harry has diabetes.
  - His predicted `logit(BMI < 30)` is 0.946 - 1.044 (1) = -0.098
- Sally does not have diabetes.
  - Her predicted `logit(BMI < 30)` is 0.946 - 1.044 (0) = 0.946

Now, `logit(BMI < 30)` = `log(odds(BMI < 30))`, so exponentiate to get the odds. . .

- Harry has predicted `odds(BMI < 30)` = exp(-0.098) = 0.9066
- Sally has predicted `odds(BMI < 30)` = exp(0.946) = 2.575

Can we convert these `odds` into something more intuitive?

Converting Odds to Probabilities

- Harry has predicted `odds(BMI < 30)` = exp(-0.098) = 0.9066
- Sally has predicted `odds(BMI < 30)` = exp(0.946) = 2.575

$$odds(BMI < 30) = \frac{Pr(BMI < 30)}{1 - Pr(BMI < 30)}$$

so, a little algebra tells us that:

$$Pr(BMI < 30) = \frac{odds(BMI < 30)}{odds(BMI < 30) + 1}$$

- So Harry's predicted `Pr(BMI < 30)` = 0.9066 / 1.9066 = 0.48
- Sally's predicted `Pr(BMI < 30)` = 2.575 / 3.575 = 0.72
- odds range from 0 to ∞, and log(odds) range from $-\infty$ to ∞.
- odds > 1 if probability > 0.5. If odds $= 1$, then probability $= 0.5$.

What about the odds ratio?

`logit(BMI < 30) = 0.946 - 1.044 (dm_status = Yes)`

- Harry, with diabetes, has odds(BMI < 30) $= 0.9066$
- Sally, without diabetes, has odds(BMI < 30) $= 2.575$

Odds Ratio for BMI < 30 associated with having diabetes (vs. not) $=$

$$\frac{0.9066}{2.575} = 0.352$$

- Our model estimates that a subject with diabetes has 35.2% of the odds of a subject without diabetes of having BMI < 30.

Can we calculate the odds ratio from the equation's coefficients?

- Yes, `exp(-1.044)` $= 0.352$.

Tidy with exponentiation

```
tidy(a7, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.9) %>%
  select(term, estimate, conf.low, conf.high) %>%
  knitr::kable(digits = 3)
```

term	estimate	conf.low	conf.high
(Intercept)	2.575	2.459	2.697
dm_statusYes	0.352	0.316	0.393

- The odds ratio for BMI < 30 among subjects with diabetes as compared to those without diabetes is 0.352
- The odds of BMI < 30 are 35.2% as large for subjects with diabetes as they are for subjects without diabetes, according to this model.
- A 90% uncertainty interval for the odds ratio estimate includes (0.316, 0.393).

Interpreting these summaries

Connecting the Odds Ratio and Log Odds Ratio to probability statements. . .

- If the probabilities were the same (for diabetes and non-diabetes subjects) of having BMI < 30, then the odds would also be the same, and so the odds ratio would be 1.
- If the probabilities of BMI < 30 were the same and thus the odds were the same, then the log odds ratio would be `log(1)` $= 0$.

`logit(BMI < 30) = 0.946 - 1.044 (dm_status = Yes)`

1. If the log odds of a coefficient (like `diabetes = Yes`) are negative, then what does that imply?
2. What if we flipped the order of the levels for diabetes so our model was about `diabetes = No`?

New model: `logit(BMI < 30) = -0.098 + 1.044 (dm_status = No)`

A Two-Factor Logistic Regression

First, let's try a model without interaction.

```
a8_without <- smart1_sh %$%
      glm(bmilt30 ~ dm_status + genhealth,
          family = binomial()) # logit is default link

tidy(a8_without) %>% select(term, estimate) %>%
  knitr::kable(digits = 3)
```

term	estimate
(Intercept)	1.716
dm_statusYes	-0.813
genhealth2_VeryGood	-0.595
genhealth3_Good	-1.051
genhealth4_Fair	-1.124
genhealth5_Poor	-1.244

Our model a8_without

```
logit(BMI < 30) = log(odds(BMI < 30))
   = 1.72 - 0.81 (dm_status = Yes)
          - 0.60 (genhealth = Very Good)
          - 1.05 (genhealth = Good)
          - 1.12 (genhealth = Fair)
          - 1.24 (genhealth = Poor)
```

1. How do we interpret the meaning of the -0.81 coefficient for `dm_status = Yes` in this model?
2. How do we interpret the meaning of the -1.05 coefficient for `genhealth = Good`?

Our model a8_without

```
logit(BMI < 30) =
   = 1.72 - 0.81 (dm = Yes) - 0.60 (Very Good) - 1.05 (Good)
      - 1.12 (Fair) - 1.24 (Poor)
```

1. How do we interpret the meaning of the -0.81 coefficient for `dm_status = Yes` in this model?

If Harry and Sally have the **same `genhealth` status**, but Harry has diabetes and Sally does not, the model predicts that Harry's `log(odds(BMI < 30))` will be 0.81 lower than Sally's.

- Harry: `logit(BMI < 30) = (1.72 - 0.81) - 0.60 (Very Good) - 1.05 (Good) - 1.12 (Fair) - 1.24 (Poor)`
- Sally: `logit(BMI < 30) = 1.72 - 0.60 (VG) - 1.05 (G) - 1.12 (F) - 1.24 (P)`

Suppose that, for example, Harry and Sally each had Excellent `genhealth`...

Question 1 (continued)

```
logit(BMI < 30) =
   = 1.72 - 0.81 (dm = Yes) - 0.60 (Very Good) - 1.05 (Good)
      - 1.12 (Fair) - 1.24 (Poor)
```

1. How do we interpret the meaning of the -0.81 coefficient for `dm_status = Yes` in this model?

Subject	Harry	Sally
genhealth	Excellent	Excellent
dm_status	Yes	No
log(odds(BMI < 30))	1.72 - 0.81 = 0.91	1.72
odds(BMI < 30)	exp(0.91) = 2.484	exp(1.72) = 5.585
Pr(BMI < 30)	2.484/3.484 = 0.71	5.585/6.585 = 0.85

Our model `a8_without`

```
logit(BMI < 30) =
   = 1.72 - 0.81 (dm = Yes) - 0.60 (Very Good) - 1.05 (Good)
       - 1.12 (Fair) - 1.24 (Poor)
```

2. How do we interpret the meaning of the -1.05 coefficient for `genhealth = Good`?

If Harry and Sally have the **same `dm_status`**, but Harry has Good `genhealth` and Sally has Excellent `genhealth`, the model predicts that Harry's `log(odds(BMI < 30))` will be 1.05 lower than Sally's.

- Harry: `logit(BMI < 30) = 1.72 - 0.81 (dm = Yes) - 1.05`
- Sally: `logit(BMI < 30) = 1.72 - 0.81 (dm = Yes)`

Why are we comparing Harry at Good to Sally at Excellent here?

Question 2 (continued)

```
logit(BMI < 30) =
   = 1.72 - 0.81 (dm = Yes) - 0.60 (Very Good) - 1.05 (Good)
      - 1.12 (Fair) - 1.24 (Poor)
```

2. How do we interpret the meaning of the -1.05 coefficient for `genhealth = Good`?

Subject	Harry	Sally
`genhealth`	Good	Excellent
`dm_status`	No	No
`log(odds(BMI < 30))`	1.72 - 1.05 = 0.67	1.72
`odds(BMI < 30)`	exp(0.67) = 1.954	exp(1.72) = 5.585
`Pr(BMI < 30)`	1.954/2.954 = 0.66	5.585/6.585 = 0.85

- What is the odds ratio for BMI < 30 comparing Harry to Sally? $1.954/5.585 = 0.350$
- Now, what if Harry and Sally each had diabetes?

Question 2 (continued)

```
logit(BMI < 30) =
   = 1.72 - 0.81 (dm = Yes) - 0.60 (Very Good) - 1.05 (Good)
      - 1.12 (Fair) - 1.24 (Poor)
```

2. How do we interpret the meaning of the -1.05 coefficient for `genhealth = Good`?

Subject	Harry	Sally
`genhealth`	Good	Excellent
`dm_status`	Yes	Yes
`log(odds(BMI < 30))`	1.72 - 1.05 - 0.81 = -0.14	1.72 - 0.81 = 0.91
`odds(BMI < 30)`	exp(-0.14) = 0.869	exp(0.91) = 2.484
`Pr(BMI < 30)`	0.869/1.869 = 0.46	2.484/3.484 = 0.71

Now what is the odds ratio for BMI $<$ 30 comparing Harry to Sally?
0.869/2.484 = 0.350

Tidying our a8_without model

```
tidy(a8_without, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high) %>%
  knitr::kable(digits = 3)
```

term	estimate	conf.low	conf.high
(Intercept)	5.565	4.848	6.416
dm_statusYes	0.444	0.396	0.498
genhealth2_VeryGood	0.551	0.469	0.646
genhealth3_Good	0.350	0.298	0.409
genhealth4_Fair	0.325	0.272	0.387
genhealth5_Poor	0.288	0.232	0.358

How do we interpret the odds ratios here?

What about including an interaction term?

```
a8_with <- smart1_sh %$%
  glm(bmilt30 ~ dm_status * genhealth, family = binomial())

tidy(a8_with) %>%
  select(term, estimate, std.error, p.value) %>%
  knitr::kable(digits = 3)
```

Results on next slide. . .

Coefficients of model a8_with

```
a8_with <- smart1_sh %$%
  glm(bmilt30 ~ dm_status * genhealth, family = binomial())
```

term	estimate	std.error	p.value
(Intercept)	1.714	0.086	0.000
dm_statusYes	-0.733	0.486	0.132
genhealth2_VeryGood	-0.574	0.100	0.000
genhealth3_Good	-1.074	0.098	0.000
genhealth4_Fair	-1.164	0.114	0.000
genhealth5_Poor	-1.059	0.154	0.000
dm_statusYes:genhealth2_VeryGood	-0.261	0.510	0.609
dm_statusYes:genhealth3_Good	0.066	0.500	0.894
dm_statusYes:genhealth4_Fair	0.050	0.503	0.922
dm_statusYes:genhealth5_Poor	-0.586	0.531	0.270

Interpreting `a8_with` Coefficients

Equation for log(odds(BMI $<$ 30)) =

```
1.71 - 0.73 (dm = Yes)
- 0.57 (Very Good) - 1.07 (Good) - 1.16 (Fair) - 1.06 (Poor)
- 0.26 (dm = Yes)(Very Good) + 0.07 (dm = Yes)(Good)
+ 0.05 (dm = Yes)(Fair) - 0.59 (dm = Yes)(Poor)
```

How do we understand the -0.59 coefficient here?

Suppose Cersei has Excellent and Jaime has Poor `genhealth`. What are their model equations for `log(odds(BMI < 30))`?

- Cersei: 1.71 - 0.73 `dm_status`
- Jaime: (1.71 - 1.06) + ((-0.73) + (-0.59)) `dm_status`,
- so Jaime: 0.65 - 1.32 `dm_status`.

Making Predictions with `a8_with` (1)

Equation for log(odds(BMI $<$ 30)) =

```
1.71 - 0.73 (dm = Yes)
- 0.57 (Very Good) - 1.07 (Good) - 1.16 (Fair) - 1.06 (Poor)
- 0.26 (dm = Yes)(Very Good) + 0.07 (dm = Yes)(Good)
+ 0.05 (dm = Yes)(Fair) - 0.59 (dm = Yes)(Poor)
```

Subject	dm_status	genhealth	log(odds(BMI $<$ 30))
Harry	No	Excellent	1.71
Sally	No	Poor	1.71 - 1.06 = 0.65
Cersei	Yes	Excellent	1.71 - 0.73 = 0.98
Jaime	Yes	Poor	1.71 - 0.73 - 1.06 - 0.59 = -0.67

Getting R to make the predictions

(Reducing rounding errors)

```
new4 <- tibble(
  subject = c("Harry", "Sally", "Cersei", "Jaime"),
  dm_status = c("No", "No", "Yes", "Yes"),
  genhealth = c("1_Excellent", "5_Poor",
                "1_Excellent", "5_Poor"))

predict(a8_with, newdata = new4, type = "link")
```

```
         1          2          3          4
 1.7139120  0.6552022  0.9808293 -0.6638768
```

Making Predictions with `a8_with` (2)

```
1.71 - 0.73 (dm = Yes)
- 0.57 (Very Good) - 1.07 (Good) - 1.16 (Fair) - 1.06 (Poor)
- 0.26 (dm = Yes)(Very Good) + 0.07 (dm = Yes)(Good)
+ 0.05 (dm = Yes)(Fair) - 0.59 (dm = Yes)(Poor)
```

Subject	dm	genhealth	odds(BMI $<$ 30)
Harry	No	Excellent	exp(1.71) = 5.53
Sally	No	Poor	exp(0.65) = 1.92
Cersei	Yes	Excellent	exp(0.98) = 2.66
Jaime	Yes	Poor	exp(-0.67) = 0.51

Getting R to make the predictions

(Reducing rounding errors)

```
new4 <- tibble(
  subject = c("Harry", "Sally", "Cersei", "Jaime"),
  dm_status = c("No", "No", "Yes", "Yes"),
  genhealth = c("1_Excellent", "5_Poor",
                "1_Excellent", "5_Poor"))

predict(a8_with, newdata = new4, type = "link") # logit
```

```
        1         2         3         4
1.7139120 0.6552022 0.9808293 -0.6638768
```

```
exp(predict(a8_with, newdata = new4, type = "link")) # odds
```

```
        1         2         3         4
5.5506329 1.9255319 2.6666667 0.5148515
```

Making Predictions with `a8_with` (3)

```
1.71 - 0.73 (dm = Yes)
- 0.57 (Very Good) - 1.07 (Good) - 1.16 (Fair) - 1.06 (Poor)
- 0.26 (dm = Yes)(Very Good) + 0.07 (dm = Yes)(Good)
+ 0.05 (dm = Yes)(Fair) - 0.59 (dm = Yes)(Poor)
```

How do we understand the -0.59 coefficient here?

Subject	dm	genhealth	Pr(BMI $<$ 30)
Harry	No	Excellent	$5.53/6.53 = 0.85$
Sally	No	Poor	$1.92/2.92 = 0.66$
Cersei	Yes	Excellent	$2.66/3.66 = 0.73$
Jaime	Yes	Poor	$0.51/1.51 = 0.34$

Getting R to make the predictions

(Reducing rounding errors)

```
new4 <- tibble(
  subject = c("Harry", "Sally", "Cersei", "Jaime"),
  dm_status = c("No", "No", "Yes", "Yes"),
  genhealth = c("1_Excellent", "5_Poor",
                "1_Excellent", "5_Poor"))

predict(a8_with, newdata = new4, type = "response") # probs
```

```
        1         2         3         4
0.8473430 0.6581818 0.7272727 0.3398693
```

Model a8_with Results (from R's predict)

Subject	dm	genhealth	logit	odds	Pr(BMI $<$ 30)
Harry	No	Excellent	1.714	5.551	0.847
Sally	No	Poor	0.655	1.926	0.658
Cersei	Yes	Excellent	0.981	2.667	0.727
Jaime	Yes	Poor	-0.664	0.515	0.340

Calculating Odds Ratios

- Comparing DM to No DM (if GenHealth = Excellent) = 2.667/5.551 = 0.480
- Comparing Poor to Excellent (if no DM) = 1.926 / 5.551 = 0.347
- Comparing DM to No DM (if GenHealth = Poor) = 0.515/1.926 = 0.267
- Comparing Poor to Excellent (if DM) = 0.515 / 2.667 = 0.193

Exponentiating the a8_with Coefficients

```
tidy(a8_with, exponentiate = TRUE, conf.int = TRUE,
     conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high) %>%
  knitr::kable(digits = 3)
```

Results on the next slide. . .

Exponentiating the `a8_with` Coefficients

term	estimate	conf.low	conf.high
(Intercept)	5.551	4.826	6.414
dm_statusYes	0.480	0.224	1.132
genhealth2_VeryGood	0.563	0.477	0.662
genhealth3_Good	0.342	0.290	0.401
genhealth4_Fair	0.312	0.259	0.376
genhealth5_Poor	0.347	0.270	0.447
dm_statusYes:genhealth2_VeryGood	0.771	0.316	1.726
dm_statusYes:genhealth3_Good	1.068	0.445	2.349
dm_statusYes:genhealth4_Fair	1.051	0.435	2.326
dm_statusYes:genhealth5_Poor	0.557	0.221	1.292

1. Interpret the `dm_statusYes` coefficient (0.480).
2. Interpret the `genhealth5_Poor` coefficient (0.347).

Model `a8_with` Predictions, Again

1. Interpret the `dm_statusYes` coefficient (0.480).
2. Interpret the `genhealth5_Poor` coefficient (0.347).

Subject	dm	genhealth	odds(BMI $<$ 30)
Harry	No	Excellent	5.551
Sally	No	Poor	1.926
Cersei	Yes	Excellent	2.667
Jaime	Yes	Poor	0.515

Odds Ratios we calculated earlier. . .

1. Comparing DM to No DM (if GenHealth = Excellent) = 2.667/5.551 = 0.480
2. Comparing Poor to Excellent (if no DM) = 1.926 / 5.551 = 0.347

Exponentiating the `a8_with` Coefficients

term	estimate	conf.low	conf.high
(Intercept)	5.551	4.826	6.414
dm_statusYes	0.480	0.224	1.132
genhealth2_VeryGood	0.563	0.477	0.662
genhealth3_Good	0.342	0.290	0.401
genhealth4_Fair	0.312	0.259	0.376
genhealth5_Poor	0.347	0.270	0.447
dm_statusYes:genhealth2_VeryGood	0.771	0.316	1.726
dm_statusYes:genhealth3_Good	1.068	0.445	2.349
dm_statusYes:genhealth4_Fair	1.051	0.435	2.326
dm_statusYes:genhealth5_Poor	0.557	0.221	1.292

3. How do we interpret the interaction coefficients, like 0.557 for (DM = Yes)(GenHealth = Poor)?

Interpreting `a8_with` Interaction Odds Ratios

3. How do we interpret the interaction coefficients, like 0.557 for (DM = Yes)(GenHealth = Poor)?

Odds Ratios we calculated earlier. . .

- Comparing DM to No DM (if GenHealth = Poor) $\approx$ 0.267
- Comparing DM to No DM (if GenHealth = Excellent) $\approx$ 0.480
- Comparing Poor to Excellent (if DM) $\approx$ 0.193
- Comparing Poor to Excellent (if no DM) $\approx$ 0.347

Within rounding error,

$$\frac{0.267}{0.480} \approx \frac{0.193}{0.347} \approx 0.557$$

Using glance on these models

```
bind_rows(glance(a8_with), glance(a8_without)) %>%
  mutate(model = c("With Interaction", "No Interaction"),
         deviance_diff = null.deviance - deviance,
         df_diff = df.null - df.residual) %>%
  select(model, AIC, BIC, deviance_diff, df_diff) %>%
  knitr::kable(digits = 1)
```

model	AIC	BIC	deviance_diff	df_diff
With Interaction	8821.6	8890.7	447.1	9
No Interaction	8823.5	8864.9	437.2	5

Logistic Regression Comparisons via `anova`

Based on Likelihood Ratio Test

```
anova(a8_without, a8_with, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: bmilt30 ~ dm_status + genhealth
Model 2: bmilt30 ~ dm_status * genhealth
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      7406     8811.5
2      7402     8801.6  4   9.8769  0.04255 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Other options include Rao's efficient score test (`test = "Rao"`) and Pearson's chi-square test (`test = "Chisq"`)

Logistic Regression Comparisons via `anova`

Another potentially attractive option compares the models based on Mallows' C_p statistic, which is closely related to the AIC, in general, and identical to what `glance` provides for AIC, at least in this implementation of logistic regression analysis of deviance.

```
anova(a8_without, a8_with, test = "Cp")
```

```
Analysis of Deviance Table

Model 1: bmilt30 ~ dm_status + genhealth
Model 2: bmilt30 ~ dm_status * genhealth
  Resid. Df Resid. Dev Df Deviance     Cp
1      7406     8811.5                8823.5
2      7402     8801.6  4   9.8769 8821.6
```

Next Step. . .

- Adding in continuous predictors / covariates in the logistic regression setting