

432 Class 02 Slides

thomaseLove.github.io/432

2022-01-13

Today's Agenda

- Linear Regression with Categorical Predictors
- Building Two-Way ANOVA models with interaction
- Building Analysis of Covariance Models

We'll use some BRFSS/SMART data about Ohio residents to address the following questions.

- 1 What is the association of diabetes diagnosis and smoking status on BMI?
- 2 Does adjusting for subject's age affect our assessments?
- 3 How does the subject's level of physical activity fit into this?

Chapter 2 of the Course Notes builds the BRFSS/SMART data.

Setup

```
knitr::opts_chunk$set(comment = NA)
options(width = 60)      ## for the slides

library(here)            ## project management
library(knitr)            ## mostly for kable
library(mosaic)          ## mostly for favstats
library(patchwork)       ## combine plots
library(janitor)         ## mostly for tabyl
library(naniar)          ## missing data tools
library(simputation)     ## for single imputation
library(broom)           ## for tidying model output
library(tidyverse)       ## as always (dplyr, ggplot2, etc.)

theme_set(theme_bw())    ## my personal preference
```

- I used `message = FALSE` in the code chunk setup.

Codebook of variables we'll select from smart_ohio

```
smart_ohio <- read_csv(here("data/smart_ohio.csv"))
```

```
dim(smart_ohio)
```

```
[1] 7412    99
```

We'll sample 432 observations from smart_ohio on these six variables...

Variable	Type	Description
SEQNO	ID code	we'll represent as a character
bmi	Quantity	body-mass index (in kg/m ²)
smoker	4 levels	smoking status (we'll collapse to 3 levels)
dm_status	4 levels	diabetes status (we'll collapse to 2)
activity	4 levels	physical activity level (we'll re-level)
age_imp	Quantity	age (imputed from groups, see Chapter 2)

Create our working data set (day2)

```
set.seed(43202) ## note 1

day2 <- smart_ohio %>% ## note 2
  select(SEQNO, bmi, smoker, dm_status,
         activity, age_imp) %>% ## note 3
  slice_sample(n = 432) %>% ## note 4
  type.convert(as.is = FALSE) %>% ## note 5
  mutate(SEQNO = as.character(SEQNO)) ## note 6
```

- 1 set seed for random sampling
- 2 modify smart_ohio data and place in new day2 tibble
- 3 select our six variables
- 4 take a random sample of 432 rows from the data
- 5 convert all character variables into factors
- 6 set ID code as a character variable

The day2 tibble, printed

```
day2

# A tibble: 432 x 6
  SEQNO      bmi smoker dm_status activity age_imp
  <chr>    <dbl> <fct>  <fct>    <fct>    <int>
1 2017000934   NA   Never No-Diabetes Inactive      71
2 2017000411  27.3 Never No-Diabetes <NA>         69
3 2017001084  19.2 Former No-Diabetes Inactive      NA
4 2017000065  20.5 Never No-Diabetes Highly_Acti~  22
5 2017000240  25.1 Never No-Diabetes Insufficien~  61
6 2017000300  24.8 Never Pregnancy-I~ Inactive      40
7 2017001444  31.2 Former No-Diabetes Highly_Acti~  57
8 2017000258  29.4 Never No-Diabetes Active        27
9 2017000769  26.7 Former No-Diabetes Highly_Acti~  71
10 2017000247  27.4 Former No-Diabetes Inactive      91
# ... with 422 more rows
```

Initial Data Checking (Quantities)

```
favstats(~ bmi, data = day2) %>% kable(dig = 1)
```

min	Q1	median	Q3	max	mean	sd	n	missing
16.9	23.9	27.4	31.2	56.6	28.4	6.3	400	32

```
favstats(~ age_imp, data = day2) %>% kable(dig = 1)
```

min	Q1	median	Q3	max	mean	sd	n	missing
18	43	58.5	70	96	56.9	19	424	8

- 1 Do the observed ranges of values look plausible in context?
- 2 Are there missing values we need to deal with?

Checking the Categorical Data (activity)

```
day2 %>% count(activity)
```

```
# A tibble: 5 x 2
```

	activity	n
	<fct>	<int>
1	Active	60
2	Highly_Active	114
3	Inactive	124
4	Insufficiently_Active	87
5	<NA>	47

- What does <NA> mean? What should we do with this variable?

Checking the Categorical Data (activity)

```
day2 %>% count(activity)
```

```
# A tibble: 5 x 2
```

	activity	n
	<fct>	<int>
1	Active	60
2	Highly_Active	114
3	Inactive	124
4	Insufficiently_Active	87
5	<NA>	47

- What does <NA> mean? What should we do with this variable?
- Shortly, we'll **reorder** these levels in a more sensible way (suggestions?) and then we'll have to deal with the missing values, somehow.

Checking the Categorical Data (smoker)

```
day2 %>% count(smoker)
```

```
# A tibble: 5 x 2
```

	smoker	n
	<fct>	<int>
1	Current_daily	58
2	Current_not_daily	19
3	Former	124
4	Never	216
5	<NA>	15

- OK. Some missing values to deal with. What else might we do here?

Checking the Categorical Data (smoker)

```
day2 %>% count(smoker)
```

```
# A tibble: 5 x 2
```

	smoker	n
	<fct>	<int>
1	Current_daily	58
2	Current_not_daily	19
3	Former	124
4	Never	216
5	<NA>	15

- OK. Some missing values to deal with. What else might we do here?
- Shortly, we'll **collapse** this from 4 to 3 levels (how?)

Checking the Categorical Data (smoker)

```
day2 %>% count(smoker)
```

```
# A tibble: 5 x 2
```

	smoker	n
	<fct>	<int>
1	Current_daily	58
2	Current_not_daily	19
3	Former	124
4	Never	216
5	<NA>	15

- OK. Some missing values to deal with. What else might we do here?
- Shortly, we'll **collapse** this from 4 to 3 levels (how?)
- I think we'll go with Current, Former and Never

Checking the Categorical Data (dm_status)

```
day2 %>% count(dm_status)
```

```
# A tibble: 5 x 2
```

	dm_status	n
	<fct>	<int>
1	Diabetes	67
2	No-Diabetes	351
3	Pre-Diabetes	9
4	Pregnancy-Induced	4
5	<NA>	1

- Next Steps?

Checking the Categorical Data (dm_status)

```
day2 %>% count(dm_status)
```

```
# A tibble: 5 x 2
```

	dm_status	n
	<fct>	<int>
1	Diabetes	67
2	No-Diabetes	351
3	Pre-Diabetes	9
4	Pregnancy-Induced	4
5	<NA>	1

- Next Steps?
- Shortly, we'll collapse this to two levels (how might we do that?) and then we'll deal with the missing information.

Re-ordering and collapsing in day2

```
day2 <- day2 %>%
  mutate(activity =
    fct_relevel(activity, "Highly_Active",
                 "Active", "Insufficiently_Active",
                 "Inactive")) %>%
  mutate(smoker =
    fct_collapse(smoker,
                  Current = c("Current_not_daily",
                              "Current_daily"))) %>%
  mutate(dm_status =
    fct_collapse(dm_status,
                  No = c("No-Diabetes",
                        "Pre-Diabetes",
                        "Pregnancy-Induced"),
                  Yes = "Diabetes"))
```

Sanity Checks

```
day2 %>% tabyl(activity) %>% adorn_pct_formatting()
```

	activity	n	percent	valid_percent
	Highly_Active	114	26.4%	29.6%
	Active	60	13.9%	15.6%
	Insufficiently_Active	87	20.1%	22.6%
	Inactive	124	28.7%	32.2%
	<NA>	47	10.9%	-

- Still need to deal with the missing values, but now the order makes sense.

Sanity Checks

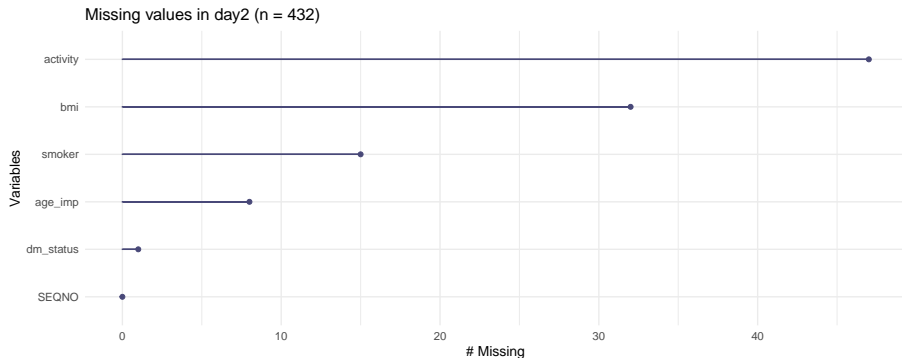
```
day2 %>% tabyl(dm_status, smoker)
```

dm_status	Current	Former	Never	NA_
Yes	10	23	33	1
No	67	100	183	14
<NA>	0	1	0	0

- OK, now we have two `dm_status` levels and three `smoker` levels, although we don't have a lot of currently smoking people with diabetes.
- Once we deal with the missing values, we should be all set.

How many missing values in each variable?

```
gg_miss_var(day2) +  
  labs(title = "Missing values in day2 (n = 432)")
```



- I used `warning = FALSE` in the code chunk setup.

Table of missing values in each variable?

We can also just get a count of missing values for each variable with `miss_var_summary()`, which is also part of `naniar`.

```
miss_var_summary(day2)
```

```
# A tibble: 6 x 3
  variable n_miss pct_miss
  <chr>    <int>    <dbl>
1 activity     47    10.9
2 bmi         32     7.41
3 smoker      15     3.47
4 age_imp      8     1.85
5 dm_status    1     0.231
6 SEQNO        0      0
```

How many missing values per row (subject)?

```
miss_case_table(day2) ## also from nanian package
```

```
# A tibble: 5 x 3  
  n_miss_in_case n_cases pct_cases  
      <int>      <int>      <dbl>  
1           0       362      83.8  
2           1        48     11.1  
3           2        14      3.24  
4           3         5      1.16  
5           4         3      0.694
```

- How many observations would we lose in a complete case analysis?
- Can we make the necessary assumption for a complete case analysis?

What do we lose in a complete case analysis?

```
day2_cc <- day2 %>%  
  filter(complete.cases(.))
```

```
dim(day2_cc)
```

```
[1] 362    6
```

This seems clean in some ways (and is the default approach in software), but actually it hides a very important assumption, that the data are **missing completely at random**.

```
prop_miss_case(day2); prop_miss_case(day2_cc)
```

```
[1] 0.162037
```

```
[1] 0
```

Missing Data Mechanisms (Notes, Chapter 3)

- Missing Completely at Random (MCAR)
 - The probability of missing data is the same for every subject, so that throwing out cases with missing data does not bias inferences.
- Missing at Random (MAR)
 - Here, the probability that a variable is missing depends only on available information in your data (the other variables we have). If this is so, then **imputation** is the most appropriate option.
- Missing Not at Random (MNAR)
 - Whether data are missing is dependent on either unobserved predictors, or on the actual true (but unavailable) value of the observation itself or both. Even imputation cannot solve the problem.

What should we assume in our day2 scenario?

Formulating a single imputation plan

```
miss_var_summary(day2) %>% kable()
```

variable	n_miss	pct_miss
activity	47	10.8796296
bmi	32	7.4074074
smoker	15	3.4722222
age_imp	8	1.8518519
dm_status	1	0.2314815
SEQNO	0	0.0000000

Today, we'll use a *naïve* approach to generating a single imputation.

- Impute `dm_status` with a random draw from its distribution.
- Use CART to impute `smoker` and `activity` from `dm_status`.
- Impute quantities via robust linear models using the factors.

Single imputation in day2 to yield day2_im

```
set.seed(432021)
day2_im <- day2 %>%
  data.frame() %>%
  impute_rhd(., dm_status ~ 1) %>%
  impute_cart(., smoker + activity ~ dm_status) %>%
  impute_rlm(., age_imp + bmi ~
              dm_status + smoker + activity) %>%
  tibble()
```

- `impute_rhd` (random hot deck) for `dm_status`
- `impute_cart` (classification and regression trees) for other factors
- `impute_rlm` (robust linear model) for `age_imp` and `bmi`

```
prop_miss_case(day2_im)
```

```
[1] 0
```

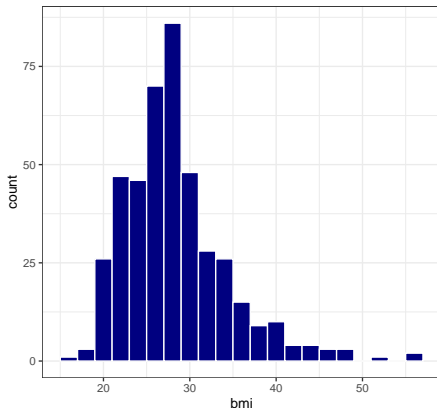

Draw the outcome?

- We're interested in diabetes and smoking's association with BMI
 - What do the BMI data look like? (plots shown on next slide)

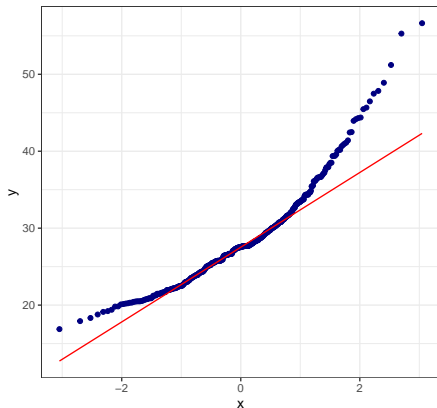
```
p1 <- ggplot(day2_im, aes(x = bmi)) +  
  geom_histogram(fill = "navy", col = "white",  
                 binwidth = 2) +  
  labs(title = "Histogram of BMI")  
  
p2 <- ggplot(day2_im, aes(sample = bmi)) +  
  geom_qq(col = "navy") + geom_qq_line(col = "red") +  
  labs(title = "Normal Q-Q plot of BMI")  
  
p1 + p2
```

BMI data in day2_im

Histogram of BMI

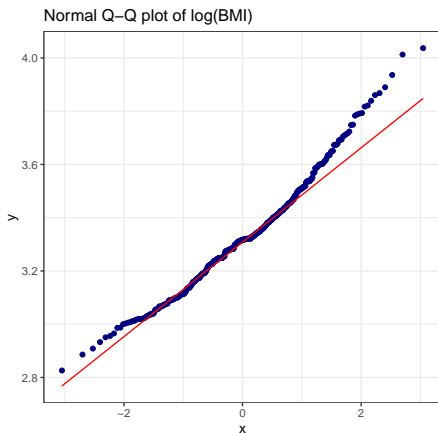
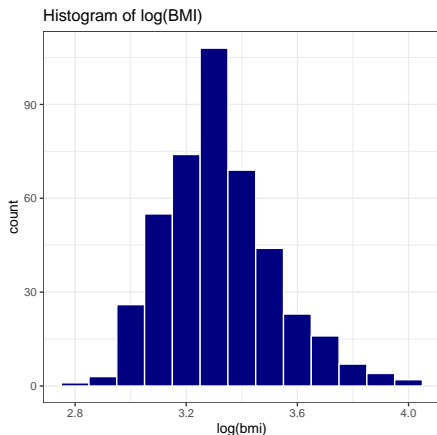


Normal Q-Q plot of BMI

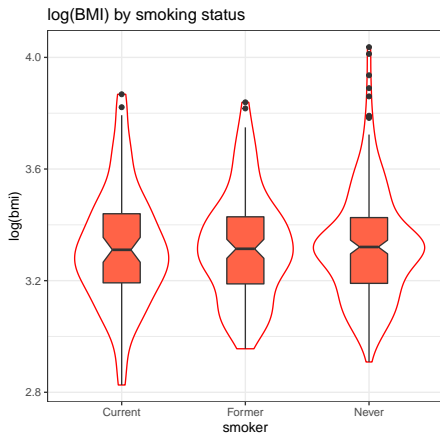
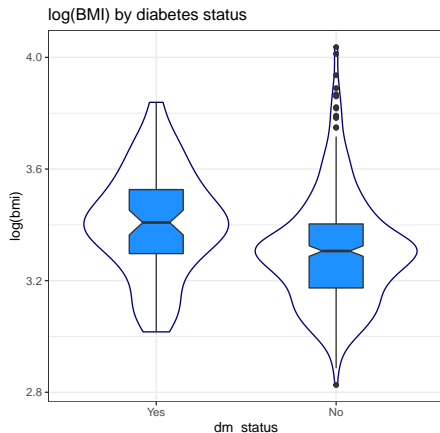


- These data are a little right-skewed. Transform?

Consider a logarithmic transformation of BMI?



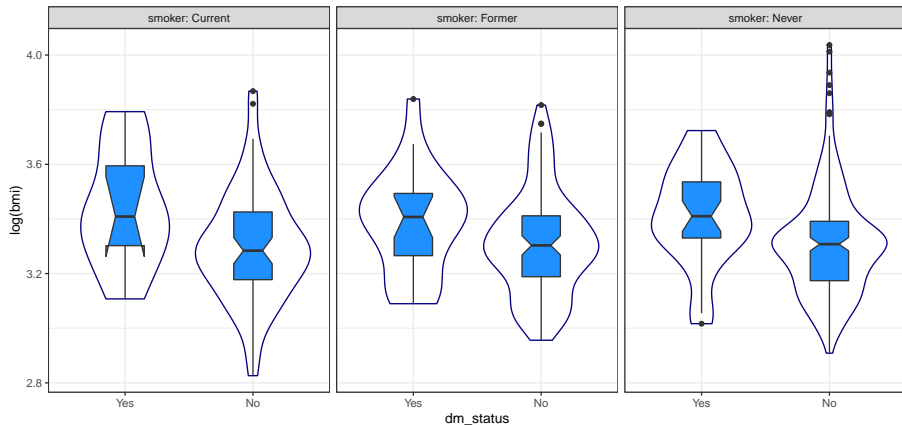
Compare $\log(\text{BMI})$ by diabetes and by smoking



log(BMI) by diabetes and smoking together

notch went outside hinges. Try setting notch=FALSE.

log(BMI) by diabetes and smoking status



Finding the Means of Each Group

We'll plot the mean of `log(bmi)` in six combinations:

- two levels of `dm_status` combined with
- three levels of `smoker`

```
summaries_1 <- day2_im %>%  
  group_by(dm_status, smoker) %>%  
  summarise(n = n(), mean = mean(log(bmi)),  
            stdev = sd(log(bmi)))
```

``summarise()`` has grouped output by `'dm_status'`. You can overr

We can suppress this message with `message = FALSE` in the code chunk label.

Here are the means of log(BMI) in each group

```
summaries_1 %>% kable(digits = 2)
```

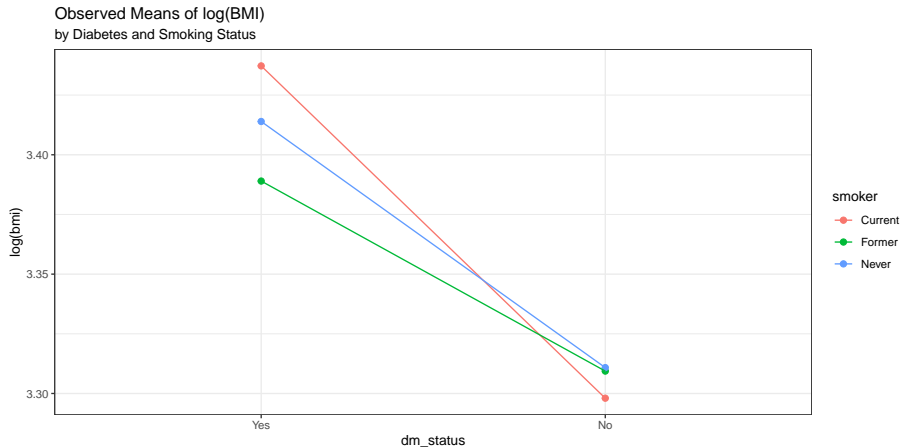
dm_status	smoker	n	mean	stdev
Yes	Current	10	3.44	0.22
Yes	Former	23	3.39	0.19
Yes	Never	34	3.41	0.18
No	Current	67	3.30	0.20
No	Former	101	3.31	0.19
No	Never	197	3.31	0.20

- Can we plot this information?

Plotting the Means (code)

```
ggplot(summaries_1, aes(x = dm_status, y = mean,  
                        col = smoker)) +  
  geom_point(size = 2) +  
  geom_line(aes(group = smoker)) +  
  labs(y = "log(bmi)",  
       title = "Observed Means of log(BMI)",  
       subtitle = "by Diabetes and Smoking Status")
```

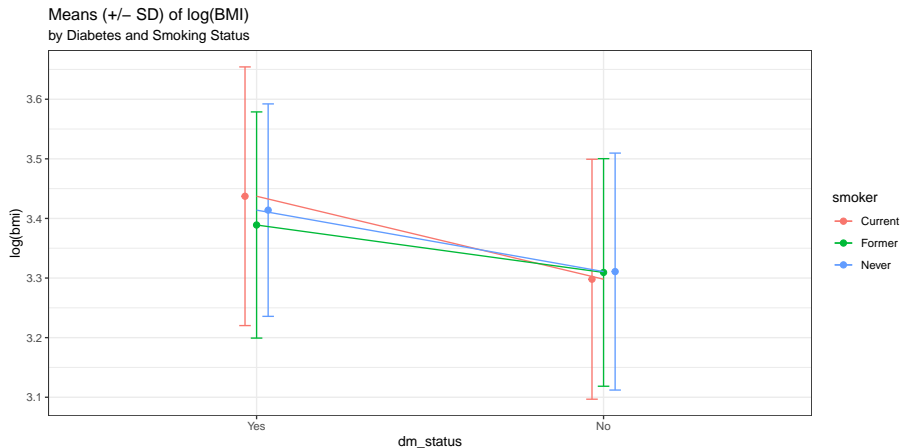

Plotting the Means (results)



Adding in standard deviations (code)

```
pd <- position_dodge(0.1)
ggplot(summaries_1, aes(x = dm_status, y = mean,
                        col = smoker)) +
  geom_errorbar(aes(ymin = mean - stdev,
                    ymax = mean + stdev),
                width = 0.1, position = pd) +
  geom_point(size = 2, position = pd) +
  geom_line(aes(group = smoker)) +
  labs(y = "log(bmi)",
       title = "Means (+/- SD) of log(BMI)",
       subtitle = "by Diabetes and Smoking Status")
```

Adding in standard deviations (code)



Review: One-Factor Analysis of Variance

```
m1 <- lm(log(bmi) ~ dm_status, data = day2_im)

anova(m1)
```

Analysis of Variance Table

Response: log(bmi)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dm_status	1	0.5748	0.57482	15.113	0.0001172 ***
Residuals	430	16.3546	0.03803		

Signif. codes:

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

Tidied m1 output

```
tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate,  
         low90 = conf.low, high90 = conf.high,  
         se = std.error, t = statistic, p = p.value) %>%  
  kable(digits = c(0,2,2,2,2,2,5))
```

term	estimate	low90	high90	se	t	p
(Intercept)	3.41	3.37	3.45	0.02	143.07	0.00000
dm_statusNo	-0.10	-0.14	-0.06	0.03	-3.89	0.00012

Revising the order?

```
day2_im <- day2_im %>%  
  mutate(dm_status = fct_relevel(dm_status, "No", "Yes"))  
  
m1 <- lm(log(bmi) ~ dm_status, data = day2_im)  
  
tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate,  
         low90 = conf.low, high90 = conf.high,  
         se = std.error, t = statistic, p = p.value) %>%  
  kable(digits = c(0,2,2,2,2,2,5))
```

term	estimate	low90	high90	se	t	p
(Intercept)	3.31	3.29	3.32	0.01	324.07	0.00000
dm_statusYes	0.10	0.06	0.14	0.03	3.89	0.00012

Glancing at m1

```
glance(m1) %>%  
  select(r.squared, adj.r.squared, sigma, AIC, BIC) %>%  
  kable(digits = c(3, 3, 2, 1, 1))
```

r.squared	adj.r.squared	sigma	AIC	BIC
0.034	0.032	0.2	-182.4	-170.2

Developing a Two-Factor Model

We want to describe the mean of $\log(\text{BMI})$ as a function of **both**

- the two-level factor `dm_status`, and
- the three-level factor `smoker`

One decision is whether we'll consider an **interaction** term between these two factors.

- A model with the interaction will fit the data a bit better, by some measures.
- A model with the interaction is most appropriate if we believe the `dm_status` relationship with $\log(\text{BMI})$ changes depending on the level of `smoker`.
 - or at least if we are unwilling to assume the `smoker` effect is the same regardless of `dm_status`

What is an interaction term (a product term)?

When we build our two-way model with interaction, we'll include a product term

```
m2 <- lm(log(bmi) ~ dm_status*smoker, data = day2_im)
```

as compared to a model without interaction, which we'd fit with:

```
m2_no <- lm(log(bmi) ~ dm_status + smoker, data = day2_im)
```

Our main tool in thinking about these will be a means plot.

Two-Way ANOVA with Interaction

```
m2 <- lm(log(bmi) ~ dm_status*smoker, data = day2_im)
```

```
anova(m2)
```

Analysis of Variance Table

Response: log(bmi)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
dm_status	1	0.5748	0.57482	14.9970	0.0001246	***
smoker	2	0.0051	0.00253	0.0659	0.9362544	
dm_status:smoker	2	0.0214	0.01070	0.2791	0.7566000	
Residuals	426	16.3282	0.03833			

Signif. codes:

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

Tidied m2 coefficients

```
tidy(m2, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate,  
         low90 = conf.low, high90 = conf.high,  
         se = std.error, p = p.value) %>%  
  kable(digits = c(0,3,2,2,2,3))
```

term	estimate	low90	high90	se	p
(Intercept)	3.298	3.26	3.34	0.02	0.000
dm_statusYes	0.139	0.03	0.25	0.07	0.037
smokerFormer	0.011	-0.04	0.06	0.03	0.713
smokerNever	0.013	-0.03	0.06	0.03	0.644
dm_statusYes:smokerFormer	-0.060	-0.19	0.07	0.08	0.459
dm_statusYes:smokerNever	-0.036	-0.16	0.09	0.08	0.634

Interpreting m2 (the interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.298 + 0.139 (\text{dm_status} = \text{Yes}) \\ & + 0.011 (\text{smoker} = \text{Former}) + 0.013 (\text{smoker} = \text{Never}) \\ & - 0.060 (\text{dm} = \text{Yes})(\text{smoker} = \text{Former}) \\ & - 0.036 (\text{dm} = \text{Yes})(\text{smoker} = \text{Never})\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?

Interpreting m2 (the interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.298 + 0.139 (\text{dm_status} = \text{Yes}) \\ & + 0.011 (\text{smoker} = \text{Former}) + 0.013 (\text{smoker} = \text{Never}) \\ & - 0.060 (\text{dm} = \text{Yes})(\text{smoker} = \text{Former}) \\ & - 0.036 (\text{dm} = \text{Yes})(\text{smoker} = \text{Never})\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298$, so estimated BMI = $\exp(3.298) = 27.06$

Interpreting m2 (the interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.298 + 0.139 (\text{dm_status} = \text{Yes}) \\ & + 0.011 (\text{smoker} = \text{Former}) + 0.013 (\text{smoker} = \text{Never}) \\ & - 0.060 (\text{dm} = \text{Yes})(\text{smoker} = \text{Former}) \\ & - 0.036 (\text{dm} = \text{Yes})(\text{smoker} = \text{Never})\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298$, so estimated BMI = $\exp(3.298) = 27.06$
- Estimated mean for a never smoker with no diabetes diagnosis?

Interpreting m2 (the interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.298 + 0.139 (\text{dm_status} = \text{Yes}) \\ & + 0.011 (\text{smoker} = \text{Former}) + 0.013 (\text{smoker} = \text{Never}) \\ & - 0.060 (\text{dm} = \text{Yes})(\text{smoker} = \text{Former}) \\ & - 0.036 (\text{dm} = \text{Yes})(\text{smoker} = \text{Never})\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298$, so estimated BMI = $\exp(3.298) = 27.06$
- Estimated mean for a never smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298 + 0.013 = 3.311$, so BMI = $\exp(3.311) = 27.41$

Interpreting m2 (the interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.298 + 0.139 (\text{dm_status} = \text{Yes}) \\ & + 0.011 (\text{smoker} = \text{Former}) + 0.013 (\text{smoker} = \text{Never}) \\ & - 0.060 (\text{dm} = \text{Yes})(\text{smoker} = \text{Former}) \\ & - 0.036 (\text{dm} = \text{Yes})(\text{smoker} = \text{Never})\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298$, so estimated BMI = $\exp(3.298) = 27.06$
- Estimated mean for a never smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298 + 0.013 = 3.311$, so BMI = $\exp(3.311) = 27.41$
- Estimated mean for a never smoker with a diabetes diagnosis?

Interpreting m2 (the interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.298 + 0.139 (\text{dm_status} = \text{Yes}) \\ & + 0.011 (\text{smoker} = \text{Former}) + 0.013 (\text{smoker} = \text{Never}) \\ & - 0.060 (\text{dm} = \text{Yes})(\text{smoker} = \text{Former}) \\ & - 0.036 (\text{dm} = \text{Yes})(\text{smoker} = \text{Never})\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298$, so estimated BMI = $\exp(3.298) = 27.06$
- Estimated mean for a never smoker with no diabetes diagnosis?
- $\log(\text{BMI}) = 3.298 + 0.013 = 3.311$, so BMI = $\exp(3.311) = 27.41$
- Estimated mean for a never smoker with a diabetes diagnosis?
- $\log(\text{BMI}) = 3.298 + 0.139 + 0.013 - 0.036 = 3.414$; BMI = 30.39

What if we assume there's no interaction?

```
m2_no <- lm(log(bmi) ~ dm_status + smoker, data = day2_im)

anova(m2_no)
```

Analysis of Variance Table

Response: log(bmi)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dm_status	1	0.5748	0.57482	15.0477	0.0001213 ***
smoker	2	0.0051	0.00253	0.0661	0.9360460
Residuals	428	16.3496	0.03820		

Signif. codes:

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

Tidied m2_no coefficients

```
tidy(m2_no, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate,  
         low90 = conf.low, high90 = conf.high,  
         se = std.error, p = p.value) %>%  
  kable(digits = c(0,3,2,2,2,3))
```

term	estimate	low90	high90	se	p
(Intercept)	3.303	3.27	3.34	0.02	0.000
dm_statusYes	0.101	0.06	0.14	0.03	0.000
smokerFormer	0.002	-0.04	0.05	0.03	0.932
smokerNever	0.008	-0.03	0.05	0.03	0.751

Interpreting m2_no (no interaction model)

m2 estimates derived from the indicator (1/0) variables

$$\begin{aligned}\log(\text{BMI}) = & 3.303 + 0.101 \text{ (dm_status = Yes)} \\ & + 0.002 \text{ (smoker = Former)} \\ & + 0.008 \text{ (smoker = Never)}\end{aligned}$$

- Estimated mean for a current smoker with no diabetes diagnosis?
 - $\log(\text{BMI}) = 3.303$, so estimated BMI = $\exp(3.303) = 27.19$
- Estimated mean for a never smoker with no diabetes diagnosis?
 - $\log(\text{BMI}) = 3.303 + 0.008 = 3.311$, so BMI = $\exp(3.311) = 27.41$
- Estimated mean for a never smoker with a diabetes diagnosis?
 - $\log(\text{BMI}) = 3.303 + 0.008 + 0.101 = 3.412$ so BMI = 30.33

What did we see in the data?

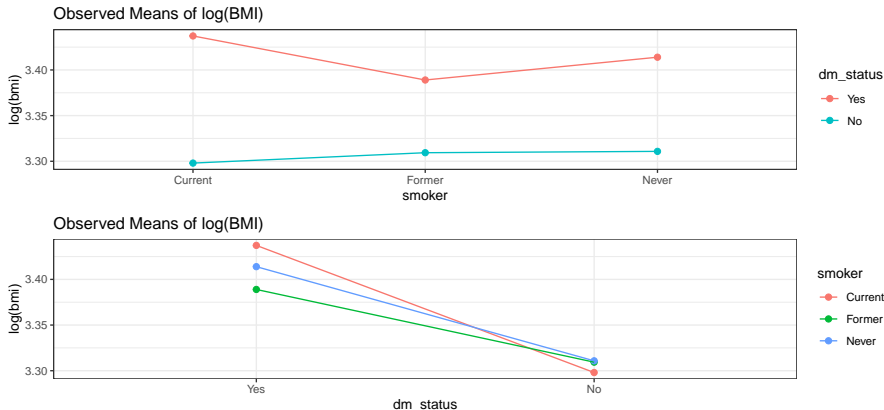
Estimates of $\text{mean}(\log(\text{BMI}))$ from the two models, vs. the actual data.

dm_status	smoker	n	actual	m2 est.	m2_no est.
No	Current	67	3.298	3.298	3.303
No	Never	197	3.311	3.311	3.311
Yes	Never	34	3.414	3.414	3.412
No	Former	101	3.309	3.309	3.305
Yes	Current	10	3.437	3.437	3.404
Yes	Former	23	3.389	3.389	3.406

- The two-way ANOVA model with interaction simply reproduces the observed means.
- Not clear we want to assume the interaction effect is actually zero.

Interaction Plot shows non-parallel lines?

Two versions of the Means Plot



How about % of variation explained measures?

```
tidy(anova(m2)) %>% select(term, df, sumsq) %>%  
  kable(dig = c(0, 0, 4))
```

term	df	sumsq
dm_status	1	0.5748
smoker	2	0.0051
dm_status:smoker	2	0.0214
Residuals	426	16.3282

- R^2 associated with the interaction term?
 - $SS(\text{interaction})$ is 0.0214
 - $SS(m2) = 0.5748 + 0.0051 + 0.0214 = 0.6013$
 - Interaction accounts for 3.6% of the variation explained by $m2$
 - Interaction accounts for $0.0214 / (0.6013 + 16.3282) = 0.0013$, or about 0.13% of the variation in $\log(\text{BMI})$

Comparison of Fit across the models?

```
comp_table <- bind_rows( glance(m2), glance(m2_no) ) %>%  
  mutate(mod = c("m2", "m2_no"))  
  
comp_table %>%  
  select(mod, r.squared, adj.r.squared, sigma, AIC, BIC) %>%  
  kable(dig = c(0, 3, 3, 3, 1, 1))
```

mod	r.squared	adj.r.squared	sigma	AIC	BIC
m2	0.036	0.024	0.196	-175.1	-146.6
m2_no	0.034	0.027	0.195	-178.5	-158.2

- Is there much to choose from in comparing the in-sample performance?

How else can we assess the fit of these models?

We're not keen on making model decisions based on significance tests. Model selection doesn't actually work well, in practice.

```
anova(m2, m2_no)
```

Analysis of Variance Table

Model 1: $\log(\text{bmi}) \sim \text{dm_status} * \text{smoker}$

Model 2: $\log(\text{bmi}) \sim \text{dm_status} + \text{smoker}$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	426	16.328				
2	428	16.350	-2	-0.021396	0.2791	0.7566

- We'd rather think about how the two models reflect the data we have *and* predict on new data.

OK, what if we add age as a covariate?

ANCOVA model `m3` takes our model `m2` (with interaction) and adds in `age_imp` (centered by subtracting off its mean) as a predictor.

```
day2_imp <- day2_imp %>%  
  mutate(age_c = age_imp - mean(age_imp))  
  
day2_imp %>% select(age_imp, age_c) %>% summary()
```

age_imp	age_c
Min. :18.00	Min. : -38.942
1st Qu.:44.00	1st Qu.: -12.942
Median :59.00	Median : 2.058
Mean :56.94	Mean : 0.000
3rd Qu.:70.00	3rd Qu.: 13.058
Max. :96.00	Max. : 39.058

OK, what if we add age as a covariate?

Here's an analysis of **covariance** model `m3` with a factor-factor interaction, plus a centered quantitative covariate.

```
m3 <- lm(log(bmi) ~ dm_status * smoker + age_c,  
         data = day2_im)
```

Does this change the nature of the relationship we see between `dm_status`, `smoker` and `bmi`?

Model m3 coefficients

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%  
  select(term, estimate, low90 = conf.low,  
         high90 = conf.high) %>% kable(dig = 3)
```

term	estimate	low90	high90
(Intercept)	3.288	3.249	3.328
dm_statusYes	0.148	0.039	0.257
smokerFormer	0.026	-0.025	0.077
smokerNever	0.019	-0.026	0.064
age_c	-0.001	-0.002	-0.001
dm_statusYes:smokerFormer	-0.056	-0.188	0.075
dm_statusYes:smokerNever	-0.025	-0.149	0.099

ANOVA results for m2 and m3

```
tidy(anova(m3)) %>% select(term, df, sumsq) %>% kable(dig = 3)
```

term	df	sumsq
dm_status	1	0.575
smoker	2	0.005
age_c	1	0.307
dm_status:smoker	2	0.021
Residuals	425	16.021

```
tidy(anova(m2)) %>% select(term, df, sumsq) %>% kable(dig = 3)
```

term	df	sumsq
dm_status	1	0.575
smoker	2	0.005
dm_status:smoker	2	0.021
Residuals	426	16.328

Does age_imp improve quality of fit?

```
comp_table <- bind_rows( glance(m2), glance(m3) ) %>%  
  mutate(mod = c("m2", "m3"))  
  
comp_table %>%  
  select(mod, r.squared, adj.r.squared, sigma, AIC, BIC) %>%  
  kable(dig = c(0, 3, 3, 3, 1, 1))
```

mod	r.squared	adj.r.squared	sigma	AIC	BIC
m2	0.036	0.024	0.196	-175.1	-146.6
m3	0.054	0.040	0.194	-181.3	-148.7

How about if we add a third factor (activity)?

```
m4 <- lm(log(bmi) ~ dm_status * smoker + age_c + activity,  
          data = day2_im)  
tidy(anova(m4)) %>% select(term, df, sumsq) %>%  
  kable(dig = 3)
```

term	df	sumsq
dm_status	1	0.575
smoker	2	0.005
age_c	1	0.307
activity	3	0.269
dm_status:smoker	2	0.036
Residuals	422	15.737

Does activity improve quality of fit?

```
comp_table <- bind_rows( glance(m1), glance(m2), glance(m3),  
                          glance(m4)) %>%  
  mutate(mod = c("m1", "m2", "m3", "m4"))  
  
comp_table %>%  
  select(mod, r.squared, adj.r.squared, sigma, AIC, BIC) %>%  
  kable(dig = c(0, 3, 3, 3, 1, 1))
```

mod	r.squared	adj.r.squared	sigma	AIC	BIC
m1	0.034	0.032	0.195	-182.4	-170.2
m2	0.036	0.024	0.196	-175.1	-146.6
m3	0.054	0.040	0.194	-181.3	-148.7
m4	0.070	0.051	0.193	-183.0	-138.3

What's next?

- Is it feasible to look at the assumptions of these models?
- Could we consider additional interaction terms?
 - factor-factor interactions?
 - factor-covariate interactions?
- Interaction is just one type of non-linearity. Can we include other types?
- Should we think more about back-transformation in this setting?
- Could we split the sample and consider how well we'd predict in new data?
- Is `lm` the best way to fit regression models to a quantitative outcome like `log(bmi)`?

Can we build up our framework for developing regression models?