431 Class 15

thomaselove.github.io/431

2020-10-13

Today's Agenda

- Comparing Means using Regression Models
  - Comparing Two Groups
  - Comparing More Than Two Groups
  - What you'll be using in Project A
- Alternatives for Comparing Means using Two Independent Samples
  - Welch's t test (not assuming equal population variances)
  - Bootstrap methods for comparing means in 2 samples
  - Rank-based alternatives (Wilcoxon-Mann-Whitney)

Today's R Packages and Data

```
library(broom)
library(ggrepel)
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())

source("data/Love-boost.R") # new today!

dm431 <- readRDS("data/dm431_2020.Rds")
```

Comparing Means with `dm431`

Our population: ALL adults ages 31-70 seen for care this year and two years ago who live in Northeast Ohio with a diabetes diagnosis.

Our `dm431` sample: 431 of those people, drawn in a way we hope is representative (but certainly isn't random).

1. Can we estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not?
2. Can we estimate the difference between people with four types of `insurance` in terms of their population mean hemoglobin A1c? (or maybe their diastolic BP?)
3. Can we estimate the difference between females and males in terms of the population mean systolic blood pressure?

Comparing Population Means using Regression Models

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_{Statin}$ and $\bar{x}_{NoStatin}$?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_{Statin}$ and $\bar{x}_{NoStatin}$?
- How might we estimate the difference in population means, $\mu_S - \mu_N$?

dm431 Example 1 (Comparing LDL by Statin usage)

Estimate the difference in the population mean LDL cholesterol for those who have a statin prescription as compared to those who do not.

```
mosaic::favstats(ldl ~ statin, data = dm431) %>%
  kable(digits = 2)
```

statin	min	Q1	median	Q3	max	mean	sd	n	missing
0	31	76	98.0	114.5	177	97.42	29.22	72	14
1	36	70	88.5	113.0	227	96.41	35.33	322	23

- What is the outcome here?
- What are the two exposure groups we are comparing?
- What are the sample means, $\bar{x}_{Statin}$ and $\bar{x}_{NoStatin}$?
- How might we estimate the difference in population means, $\mu_S - \mu_N$?
- Is there a problem in these data we need to deal with?

How much missing data do we have?

Do we have missing values in both columns, or just one?

```
dm431 %>% summarize(across(c(statin, ldl), ~ sum(is.na(.x))))
```

```
# A tibble: 1 x 2
  statin   ldl
   <int> <int>
1      0    37
```

So what shall we do?

- Drop the 37 cases, or
- Something else?

On Missing Data

Drop the Missing = A "Complete Case" analysis

- We could drop these 37, and do a **complete case analysis** on the other 431-37 = 394 subjects.
- We'll also create a factor (`statin_f`) with the statin information.

```
dm431_cc <- dm431 %>% filter(complete.cases(ldl, statin)) %>%
  mutate(statin_f = fct_recode(factor(statin),
                               "Statin" = "1", "No" = "0"))

mosaic::favstats(ldl ~ statin_f, data = dm431_cc) %>%
  kable(dig = 2)
```

statin_f	min	Q1	median	Q3	max	mean	sd	n	missing
No	31	76	98.0	114.5	177	97.42	29.22	72	0
Statin	36	70	88.5	113.0	227	96.41	35.33	322	0

- HUGE assumption: The 37 missing `ldl` are MCAR.

Missing Completely at Random (MCAR)

Our complete case analysis requires the HUGE assumption that these 37 observations are what Donald Rubin called "missing completely at random."

Missing Completely at Random (MCAR) means that there is no relationship between whether a data point is missing and any values in the data set, missing or observed. Thus, the missing values are just a random subset of the data.

- That is the huge assumption that is both impossible to prove and that is also tacitly made in many settings, more or less by default.
- The alternative is to consider other possible mechanisms (besides MCAR) for why data might be missing.

Assuming data are Missing at Random (MAR)?

Missing at Random (MAR): the reason a data point is missing is related to some observed data, but unrelated to the actual missing values.

So we assume that we can predict the missing values effectively using other variables in the data, without causing any problems. That's a big assumption, but then we could *impute* (or fill in with predictions based on other variables) the missing data.

So to impute predicted `ldl` values for these 37 subjects, we'd need to:

- account for the fact that we're imputing in building estimates, and
- control for the variables which (together) predict why the data were missing, and
- remember that we are making a large and unverifiable assumption about why the data are missing.

If missing data aren't MCAR or MAR, then they are MNAR.

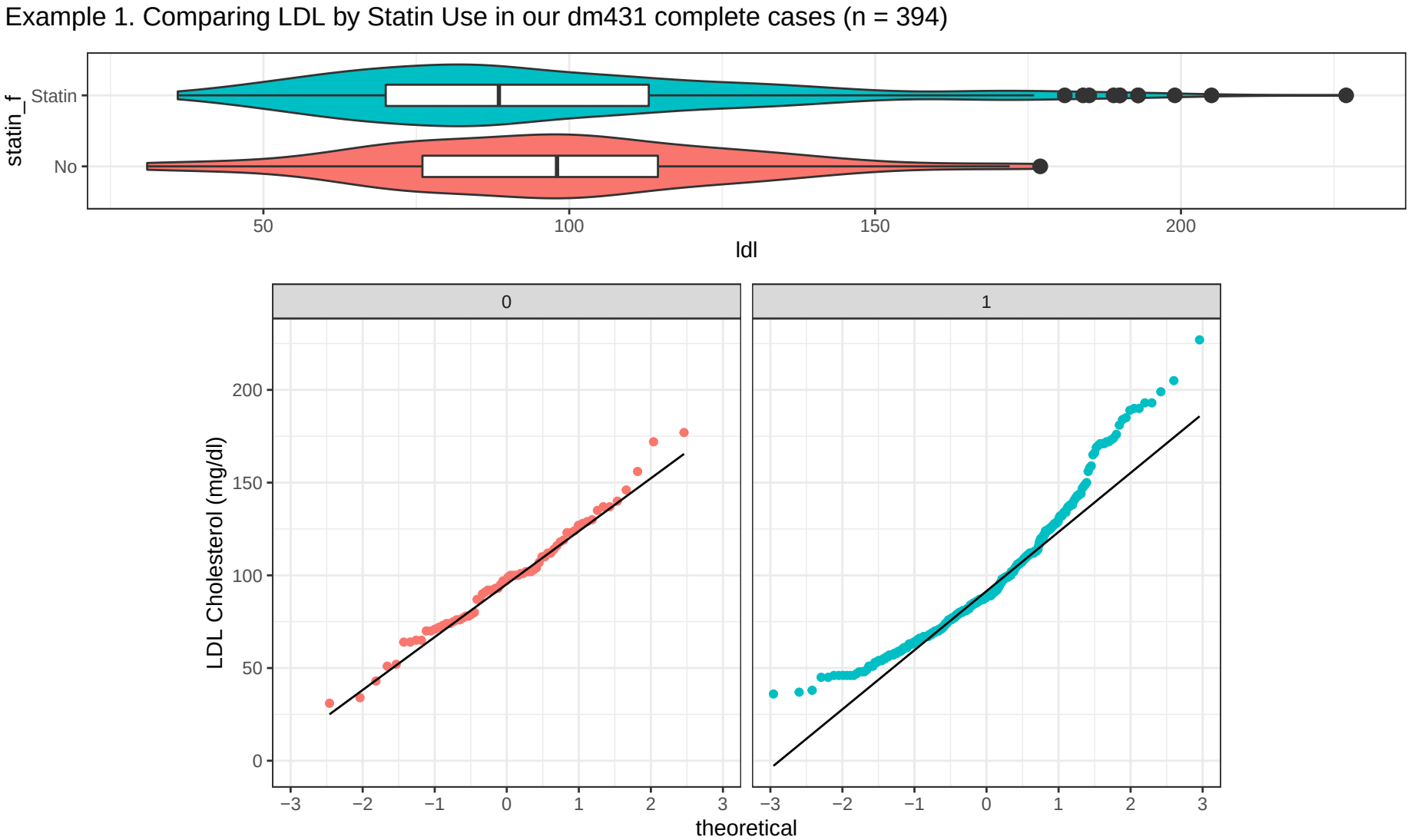
Three Types of Missingness

1. MCAR: Missing Completely At Random (ignorable nonresponse)
   - missing values are just a random subset of the data
   - unrealistically strong assumption in practice, although it's easy
   - makes a complete case analysis unbiased
2. MAR: Missing At Random
   - reason for missingness can be completely accounted for by variables where there is complete information
   - much more reasonable in many settings than MCAR, but impossible to verify statistically
   - imputing missing values here leads to a more robust conclusion
3. MNAR: Missing Not at Random (nonignorable nonresponse)
   - data are neither MCAR nor MAR
   - the reason the data is missing is related to its value, even after controlling for other variables.

These have different effects on the validity of the conclusions you build.

DTDP: Example 1. (Comparing LDL by Statin Use)

Assuming MCAR, we'll press on with a complete case analysis.



Linear Model for Example 1 (slide A)

Estimate the difference in population mean LDL cholesterol among people taking a statin as compared to those not taking a statin.

```
app1 <- lm(ldl ~ statin, data = dm431_cc)

tidy(app1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	97.42	4.04	24.09	0.00	90.75	104.08
statin	-1.01	4.47	-0.23	0.82	-8.38	6.36

- What can we learn from this output?
  - What is the sample mean `ldl` for those not on a statin?
  - What is the sample mean `ldl` for statin users?
  - The point estimate for $\mu_S - \mu_N$ is ...

Linear Model for Example 1 (slide B)

Estimate the difference in population mean LDL cholesterol among people taking a statin as compared to those not taking a statin.

```
app1 <- lm(ldl ~ statin, data = dm431_cc)

tidy(app1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	97.42	4.04	24.09	0.00	90.75	104.08
statin	-1.01	4.47	-0.23	0.82	-8.38	6.36

- What can we learn from this output?
  - The point estimate for $\mu_S - \mu_N$ is **-1.01**
  - The 90% confidence interval for $\mu_S - \mu_N$ is ...

Linear Model for Example 1 (slide C)

Estimate the difference in population mean LDL cholesterol among people taking a statin as compared to those not taking a statin.

```
app1 <- lm(ldl ~ statin, data = dm431_cc)

tidy(app1, conf.int = T, conf.level = 0.90) %>% kable(dig = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	97.42	4.04	24.09	0.00	90.75	104.08
statin	-1.01	4.47	-0.23	0.82	-8.38	6.36

- What can we learn from this output?
  - The point estimate for $\mu_S - \mu_N$ is -1.01
  - The 90% confidence interval for $\mu_S - \mu_N$ is (**-8.38**, **6.36**)

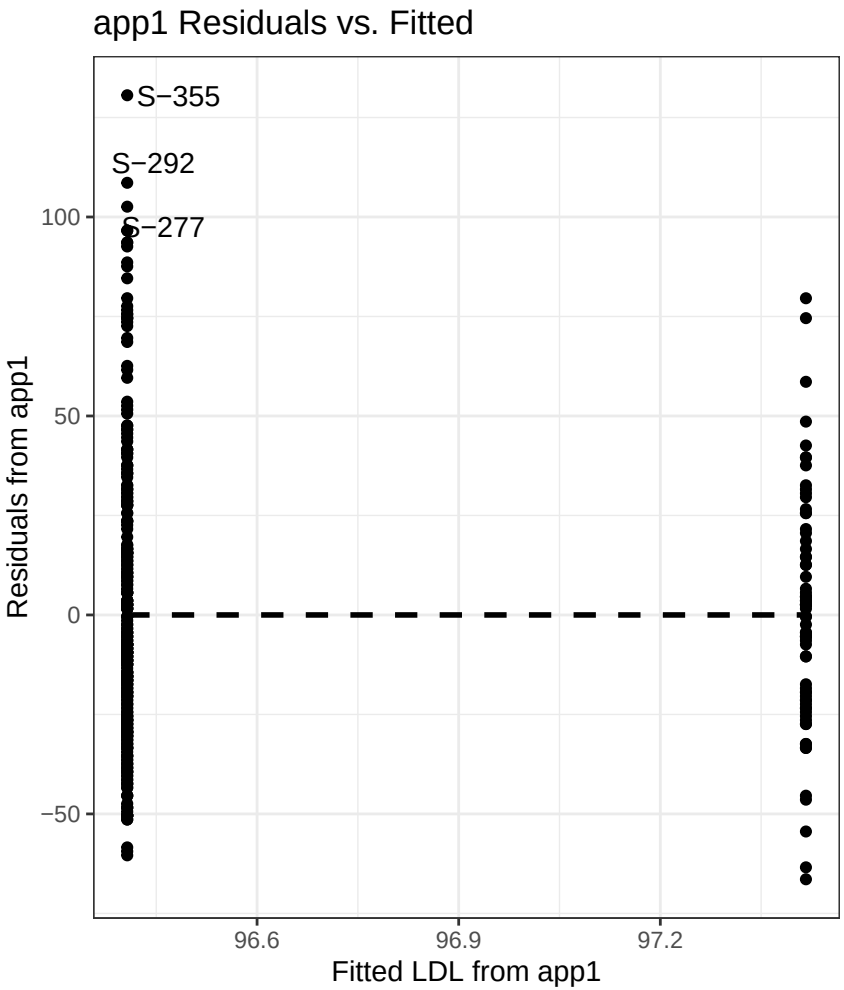
Augment our model to get fitted/residual values?

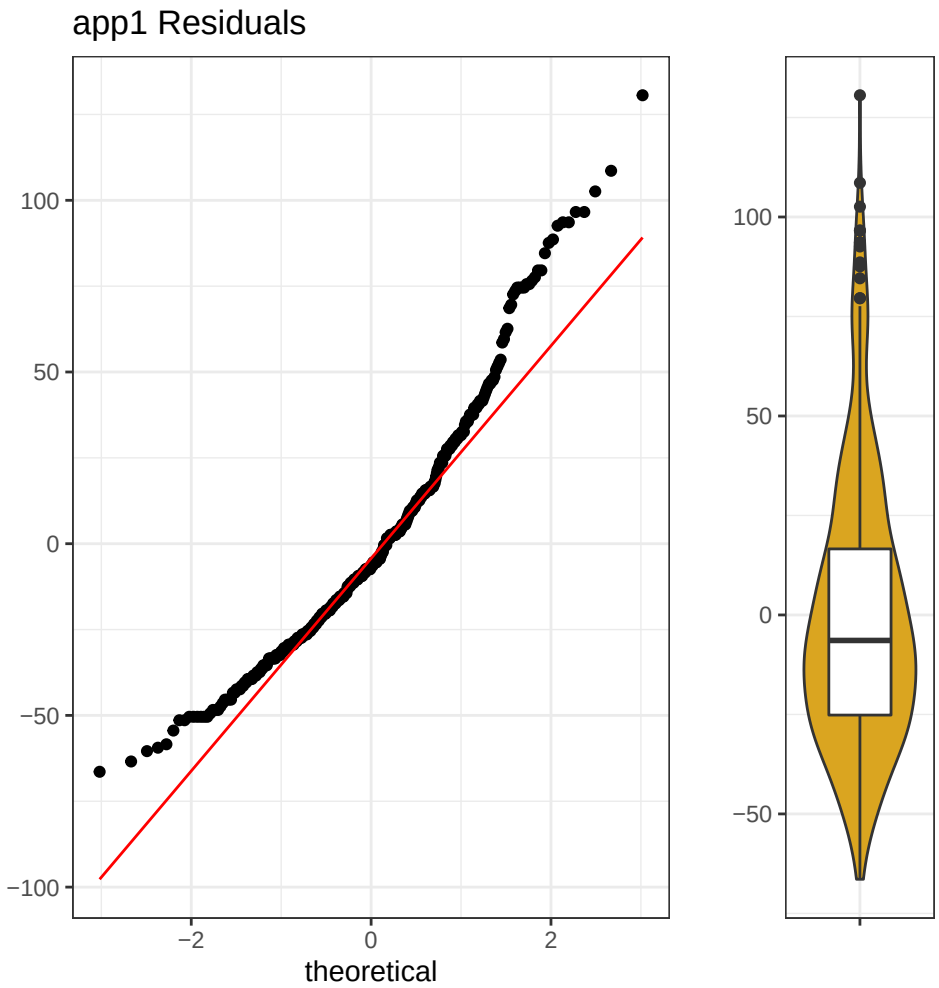
```
aug1 <- augment(app1, dm431_cc)

aug1 %>% select(subject, statin, ldl, .fitted, .resid) %>%
  slice(1, 6, 206, 394)
```

```
# A tibble: 4 x 5
  subject statin   ldl .fitted .resid
  <chr>     <int> <int>   <dbl>  <dbl>
1 S-001         1   126    96.4  29.6
2 S-006         1    65    96.4 -31.4
3 S-224         0   100    97.4   2.58
4 S-431         0    77    97.4 -20.4
```

Here, I'm just using `slice` to pick out four values from the distribution (two with statin, two without and two with a positive and two with a negative residual.)

Residual Plots for Example 1 app1?





Conclusions So Far: Example 1

- The point estimate for $\mu_S - \mu_N$ is -1.01
- The 90% confidence interval for $\mu_S - \mu_N$ is (**-8.38**, **6.36**)
- There is some evidence of non-Normality in the residuals after this regression model.
  - Perhaps the assumption that the difference $\mu_S - \mu_N$ is Normally distributed is in question. This will eventually lead to alternatives to the t test, discussed later in these slides.

But for now, let's look at an example where we compare means across more than just two groups.

Example 2 (Comparing Hemoglobin A1c by Insurance type)

Comparing A1c by Insurance Type in dm431

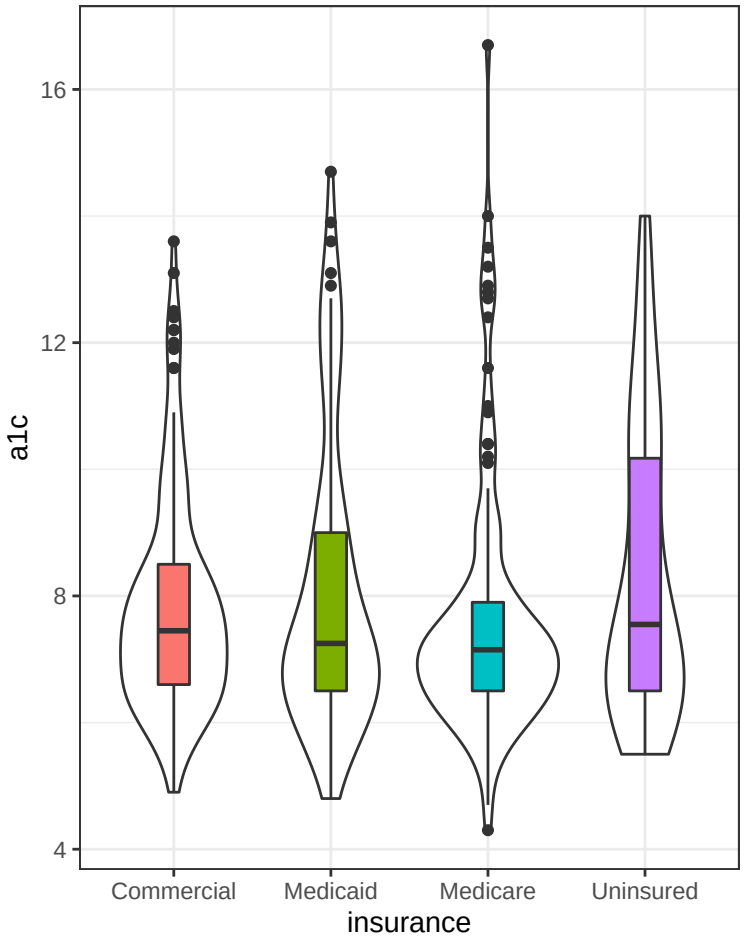
```
dm431 %>% select(insurance, a1c) %>% glimpse()
```

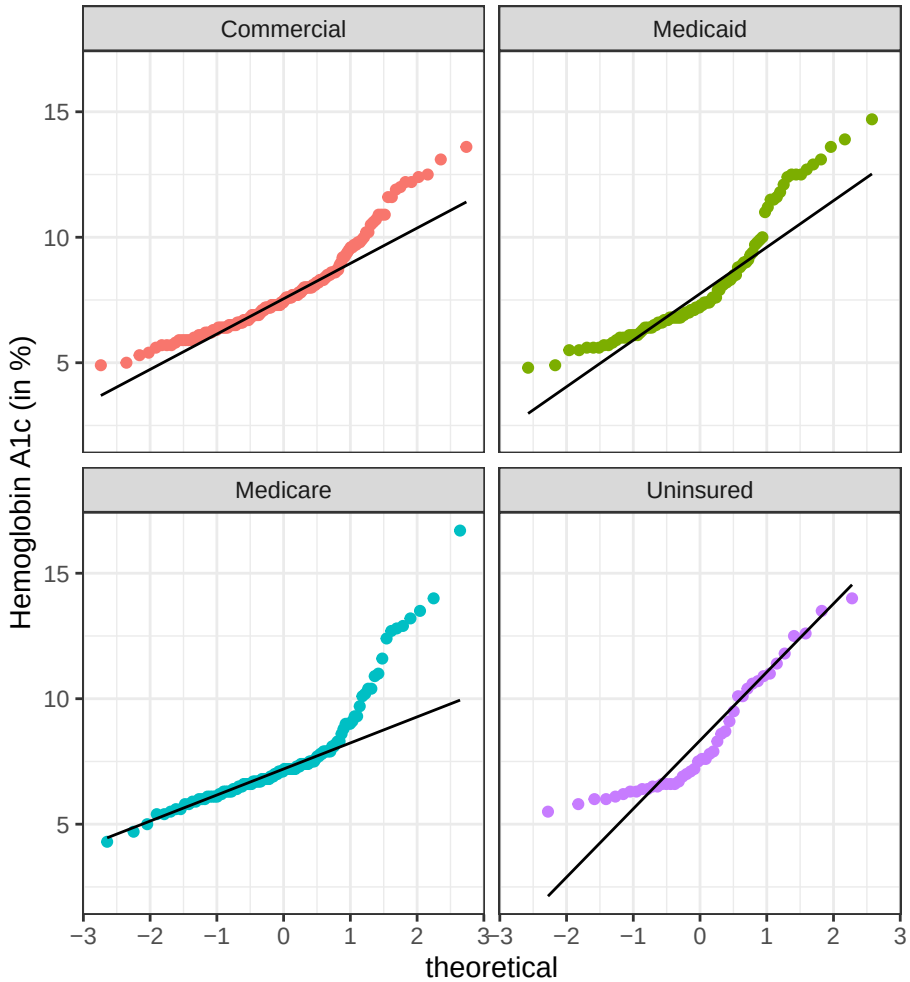
```
Rows: 431
Columns: 2
$ insurance <fct> Commercial, Uninsured, Uninsured...
$ a1c       <dbl> 6.3, 11.0, 8.7, 6.5, 6.7, 5.8, 9...
```

```
dm431 %$% mosaic::favstats(a1c ~ insurance) %>%
  rename(na = missing) %>% kable(dig = 2)
```

insurance	min	Q1	median	Q3	max	mean	sd	n	na
Commercial	4.9	6.6	7.45	8.50	13.6	7.83	1.72	162	2
Medicaid	4.8	6.5	7.25	9.00	14.7	8.07	2.30	100	0
Medicare	4.3	6.5	7.15	7.90	16.7	7.64	2.04	122	1
Uninsured	5.5	6.5	7.55	10.17	14.0	8.35	2.33	44	0

Distribution of A1c in insurance groups





Code for previous slide

```
dm_comp <- dm431 %>%
  filter(complete.cases(a1c, insurance))

p1 <- ggplot(dm_comp, aes(x = insurance, y = a1c)) +
  geom_violin() +
  geom_boxplot(aes(fill = insurance), width = 0.2) +
  guides(fill = FALSE)

p2 <- ggplot(dm_comp, aes(sample = a1c, col = insurance)) +
  geom_qq() + geom_qq_line(col = "black") +
  guides(col = FALSE) +
  theme(aspect.ratio = 1) +
  labs(y = "Hemoglobin A1c (in %)") +
  facet_wrap(~ insurance)

p1 + p2 + plot_layout(widths = c(2,3))
```

We'll assume MCAR and run a model

```
dm_comp <- dm431 %>%
  filter(complete.cases(a1c, insurance))

modA <- lm(a1c ~ insurance, data = dm_comp)
modA
```

```
Call:
lm(formula = a1c ~ insurance, data = dm_comp)

Coefficients:
       (Intercept)   insuranceMedicaid
            7.8272              0.2468
 insuranceMedicare  insuranceUninsured
           -0.1919              0.5251
```

- It was very helpful that `insurance` was a factor already.

Model A Fit Summary

```
glance(modA) %>%
  select(r.squared, statistic, df, df.residual,
         p.value, sigma, nobs) %>%
  kable(dig = c(3, 2, 0, 0, 4, 2, 0))
```

r.squared	statistic	df	df.residual	p.value	sigma	nobs
0.012	1.73	3	424	0.1598	2.03	428

What can we conclude about whether `insurance` is an effective predictor of `a1c` in these data?

Model A Coefficients

```
tidy(modA, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  kable(dig = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	7.83	0.16	7.56	8.09
insuranceMedicaid	0.25	0.26	-0.18	0.67
insuranceMedicare	-0.19	0.24	-0.59	0.21
insuranceUninsured	0.53	0.34	-0.04	1.09

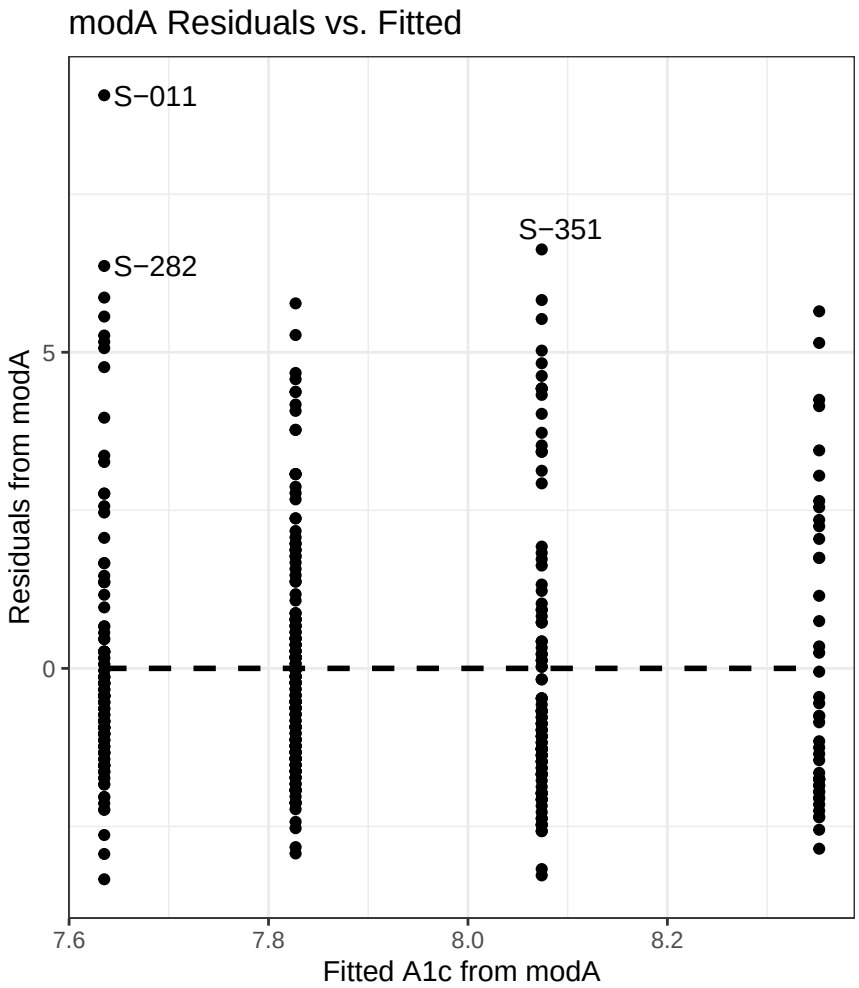
- Which insurance type is associated with the highest (worst) A1c?
- Which has the lowest predicted A1c? Are these results surprising?

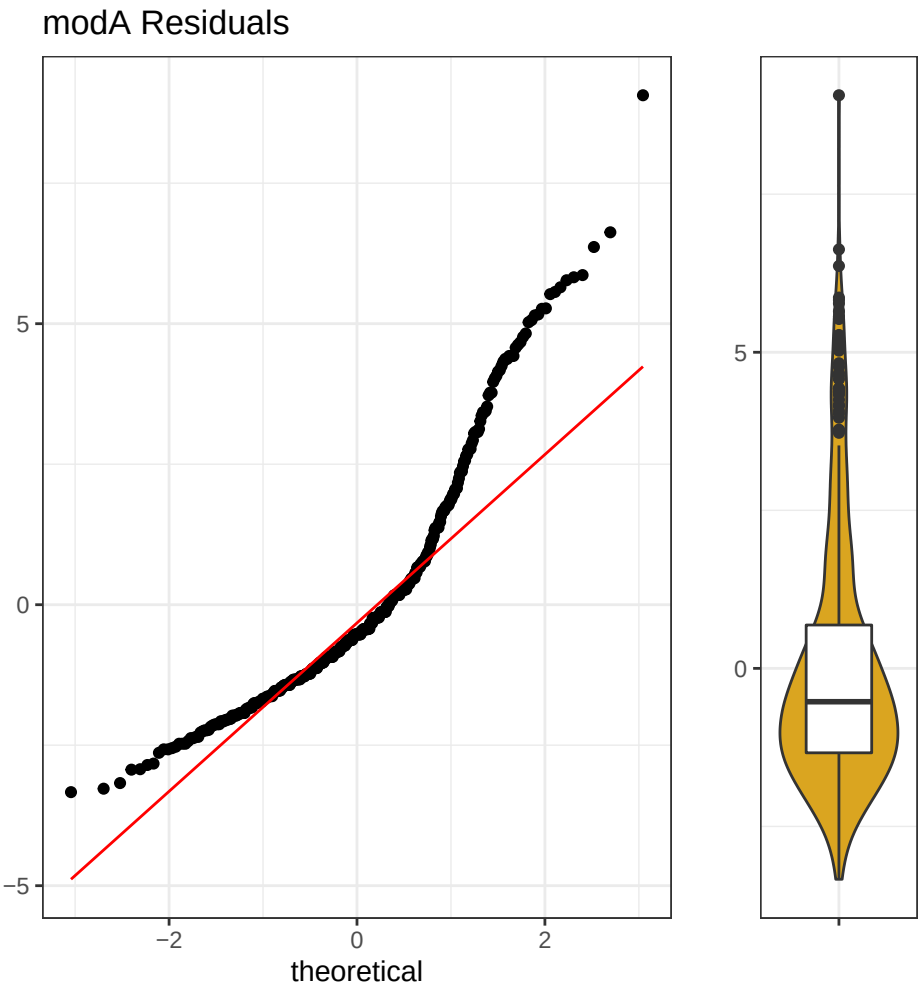
Making Predictions with augment

```
augA <- augment(modA, dm_comp)

augA %>% select(subject, insurance, a1c, .fitted, .resid) %>%
  head() %>% kable(dig = 2)
```

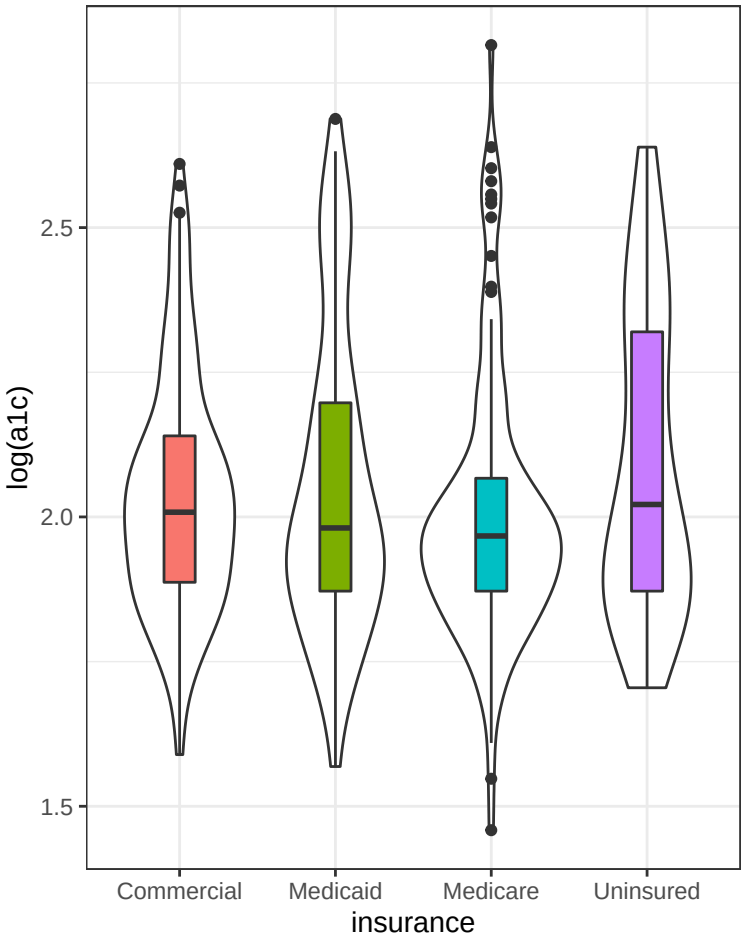
subject	insurance	a1c	.fitted	.resid
S-001	Commercial	6.3	7.83	-1.53
S-002	Uninsured	11.0	8.35	2.65
S-003	Uninsured	8.7	8.35	0.35
S-004	Commercial	6.5	7.83	-1.33
S-005	Commercial	6.7	7.83	-1.13
S-006	Medicare	5.8	7.64	-1.84

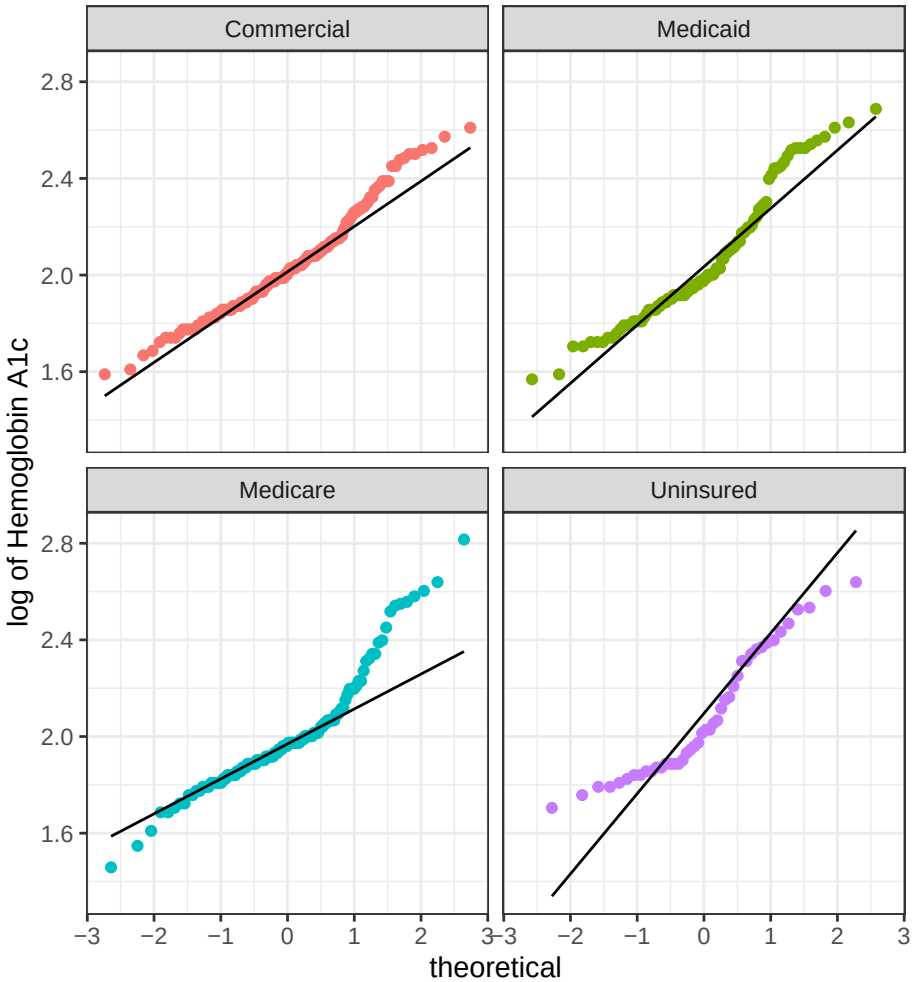
Residual Plots for `modA`





Try `log(a1c)` as our outcome instead?





log(A1c) by Insurance Type in dm431

```
dm431 %$% mosaic::favstats(log(a1c) ~ insurance) %>%
  rename(na = missing) %>% kable(dig = 3)
```

insurance	min	Q1	median	Q3	max	mean	sd	n
Commercial	1.589	1.887	2.008	2.140	2.610	2.036	0.205	162
Medicaid	1.569	1.872	1.981	2.197	2.688	2.053	0.261	100
Medicare	1.459	1.872	1.967	2.067	2.815	2.004	0.232	122
Uninsured	1.705	1.872	2.022	2.320	2.639	2.087	0.263	44

We'll assume MCAR and run the logged A1c model

```
dm_comp <- dm431 %>%
  filter(complete.cases(a1c, insurance))

modB <- lm(log(a1c) ~ insurance, data = dm_comp)
modB
```

```

Call:
lm(formula = log(a1c) ~ insurance, data = dm_comp)

Coefficients:
       (Intercept)   insuranceMedicaid
           2.03576             0.01718
 insuranceMedicare  insuranceUninsured
          -0.03202             0.05171
```

Model B Fit Summary

```
glance(modB) %>%
  select(r.squared, statistic, df, df.residual,
         p.value, sigma, nobs) %>%
  kable(dig = c(3, 2, 0, 0, 4, 2, 0))
```

r.squared	statistic	df	df.residual	p.value	sigma	nobs
0.012	1.67	3	424	0.1727	0.23	428

What can we conclude about whether `insurance` is an effective predictor of `log(a1c)` in these data?

Model B Coefficients

```
tidy(modB, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  kable(dig = 3)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	2.036	0.018	2.006	2.066
insuranceMedicaid	0.017	0.030	-0.032	0.066
insuranceMedicare	-0.032	0.028	-0.078	0.014
insuranceUninsured	0.052	0.040	-0.014	0.117

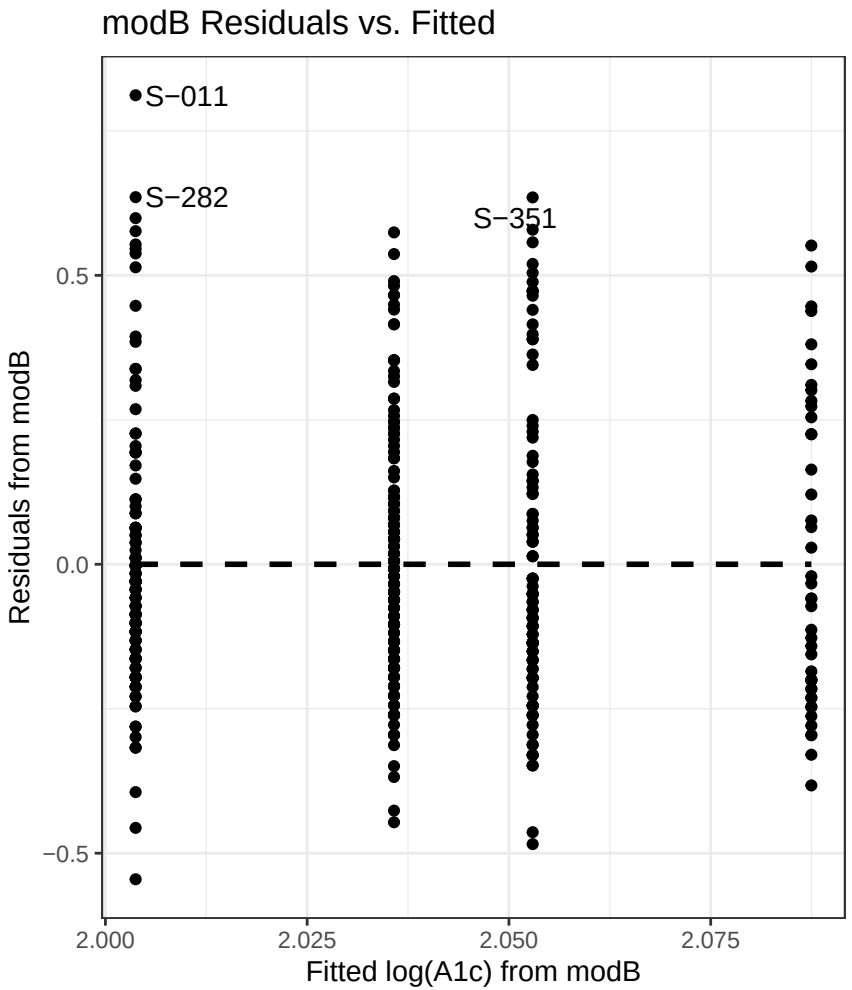
- Which insurance type is associated with the highest (worst) A1c?
- Which has the lowest predicted A1c? Are these results surprising?

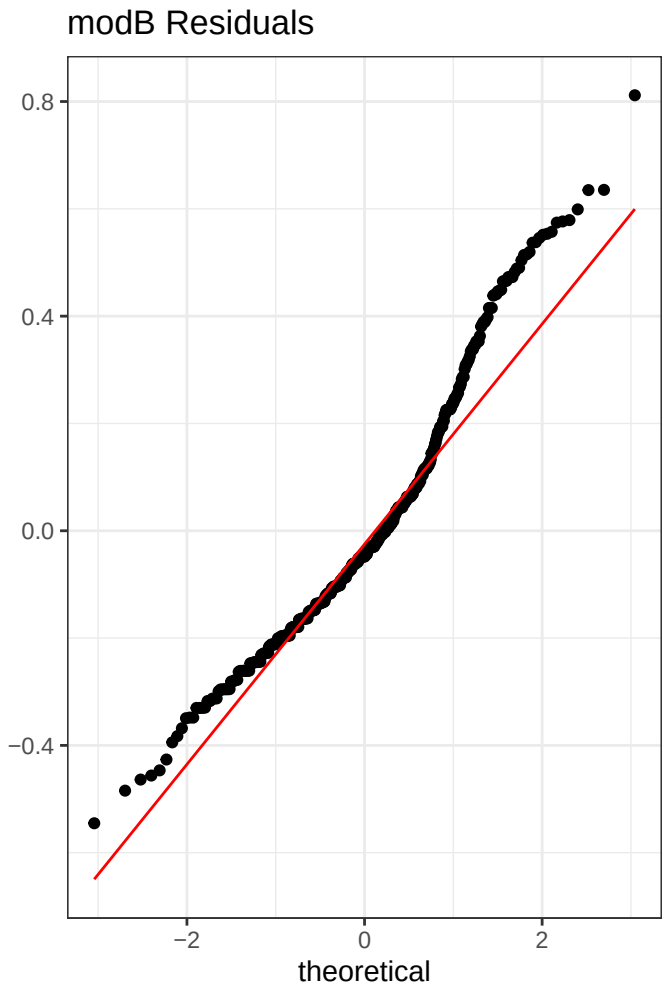
Making Predictions with augment

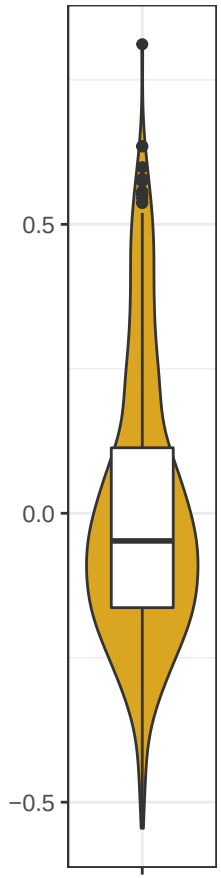
```
augB <- augment(modB, dm_comp) %>%
  mutate(log_a1c = log(a1c))

augB %>% select(subject, insurance, a1c,
                log_a1c, .fitted, .resid) %>%
  head() %>% kable(dig = 3)
```

subject	insurance	a1c	log_a1c	.fitted	.resid
S-001	Commercial	6.3	1.841	2.036	-0.195
S-002	Uninsured	11.0	2.398	2.087	0.310
S-003	Uninsured	8.7	2.163	2.087	0.076
S-004	Commercial	6.5	1.872	2.036	-0.164
S-005	Commercial	6.7	1.902	2.036	-0.134
S-006	Medicare	5.8	1.758	2.004	-0.246

Residual Plots for `modB`







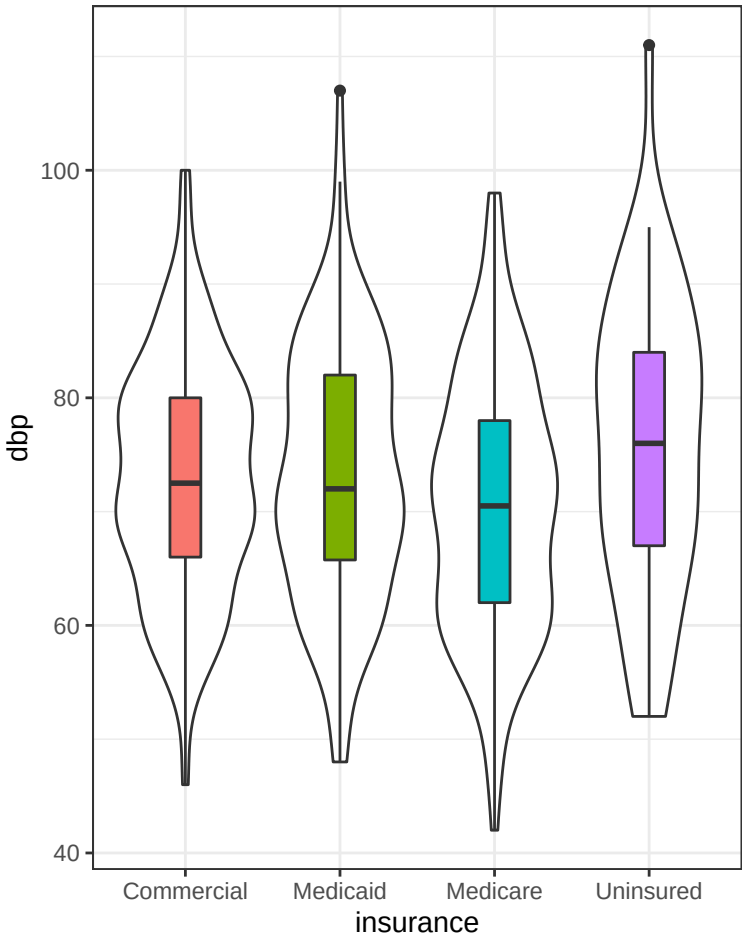
Try `dbp` as our outcome instead?

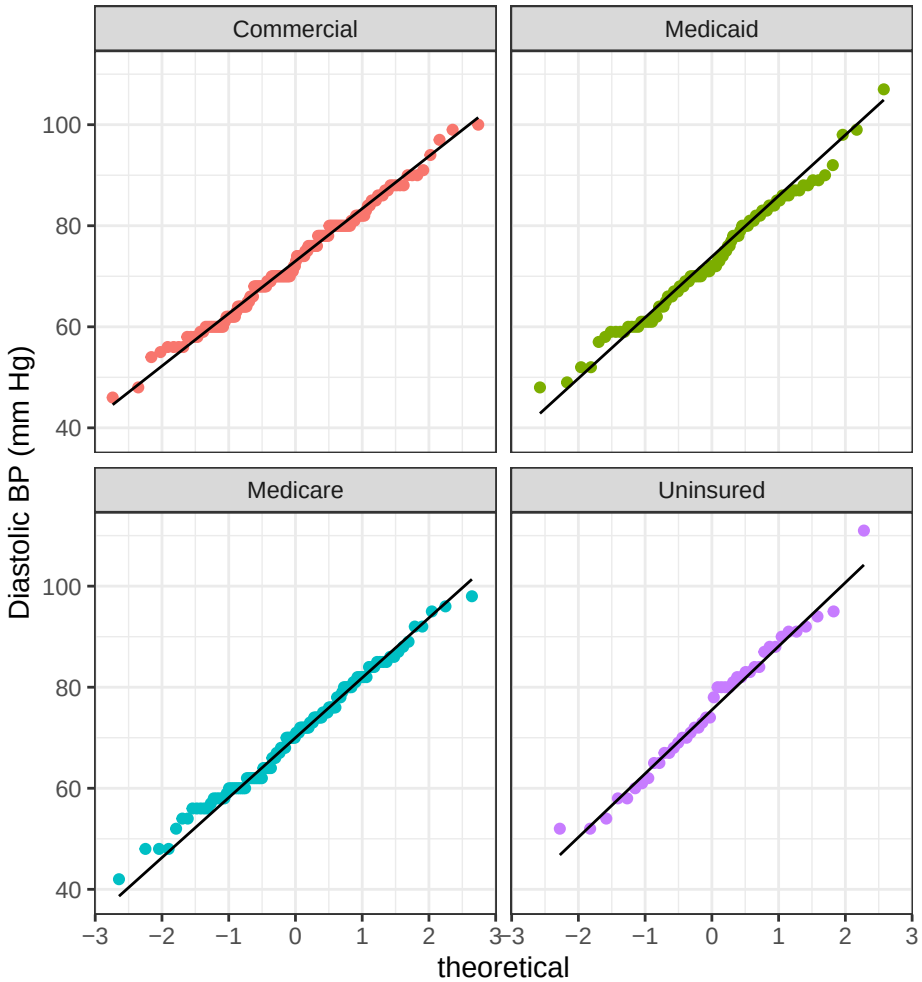
Diastolic BP by Insurance Type in dm431

```
dm431 %$% mosaic::favstats(dbp ~ insurance) %>%
  rename(na = missing) %>% kable(dig = 1)
```

insurance	min	Q1	median	Q3	max	mean	sd	n	na
Commercial	46	66.0	73.5	80	100	73.1	10.2	164	0
Medicaid	48	65.8	72.0	82	107	73.2	11.3	100	0
Medicare	42	62.0	70.0	78	98	70.3	11.1	123	0
Uninsured	52	67.0	76.0	84	111	75.9	13.0	44	0

Compare dbp across insurance types?





We'll assume MCAR and try to predict dbp

```
modD <- lm(dbp ~ insurance, data = dm431)
modD
```

```
Call:
lm(formula = dbp ~ insurance, data = dm431)

Coefficients:
       (Intercept)   insuranceMedicaid
           73.0854              0.1546
 insuranceMedicare  insuranceUninsured
           -2.7358              2.7783
```

Model D Fit Summary

```
glance(modD) %>%
  select(r.squared, statistic, df, df.residual,
         p.value, sigma, nobs) %>%
  kable(dig = c(3, 2, 0, 0, 4, 2, 0))
```

r.squared	statistic	df	df.residual	p.value	sigma	nobs
0.022	3.2	3	427	0.0234	11.05	431

What can we conclude about whether `insurance` is an effective predictor of dbp in these data?

Model D Coefficients

```
tidy(modD, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  kable(dig = 1)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	73.1	0.9	71.7	74.5
insuranceMedicaid	0.2	1.4	-2.2	2.5
insuranceMedicare	-2.7	1.3	-4.9	-0.6
insuranceUninsured	2.8	1.9	-0.3	5.9

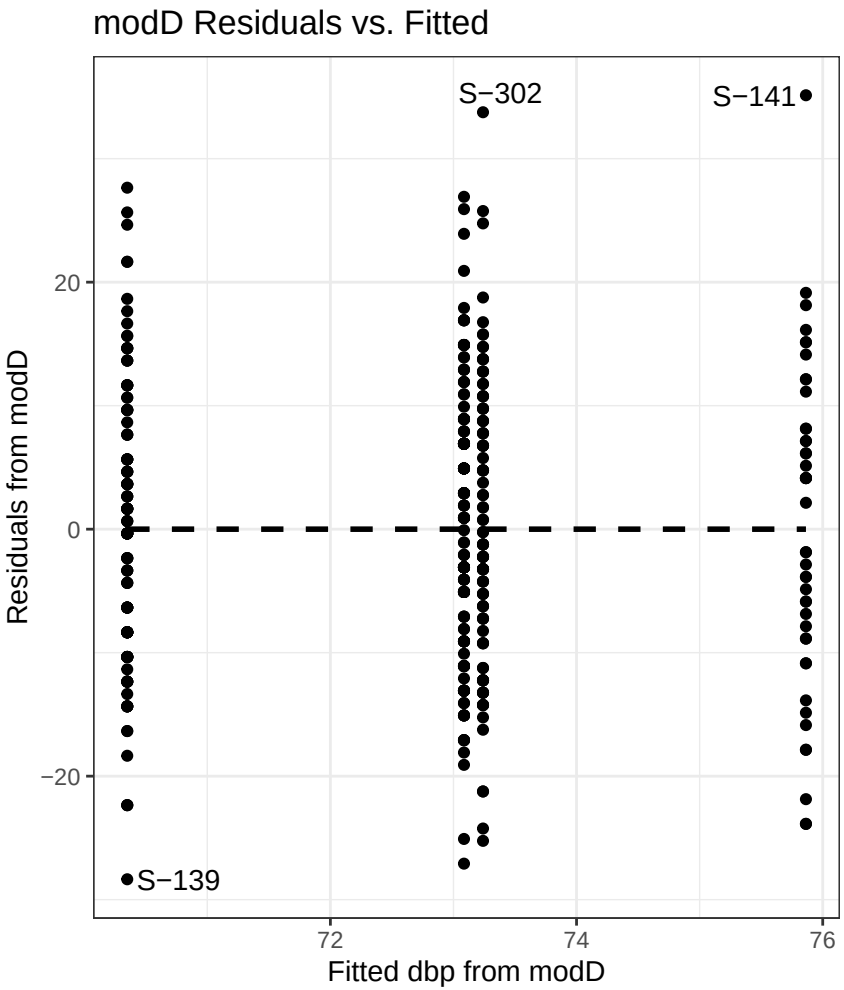
- Which insurance type is associated with the highest (worst) dbp?
- Which has the lowest predicted dbp? Are these results surprising?

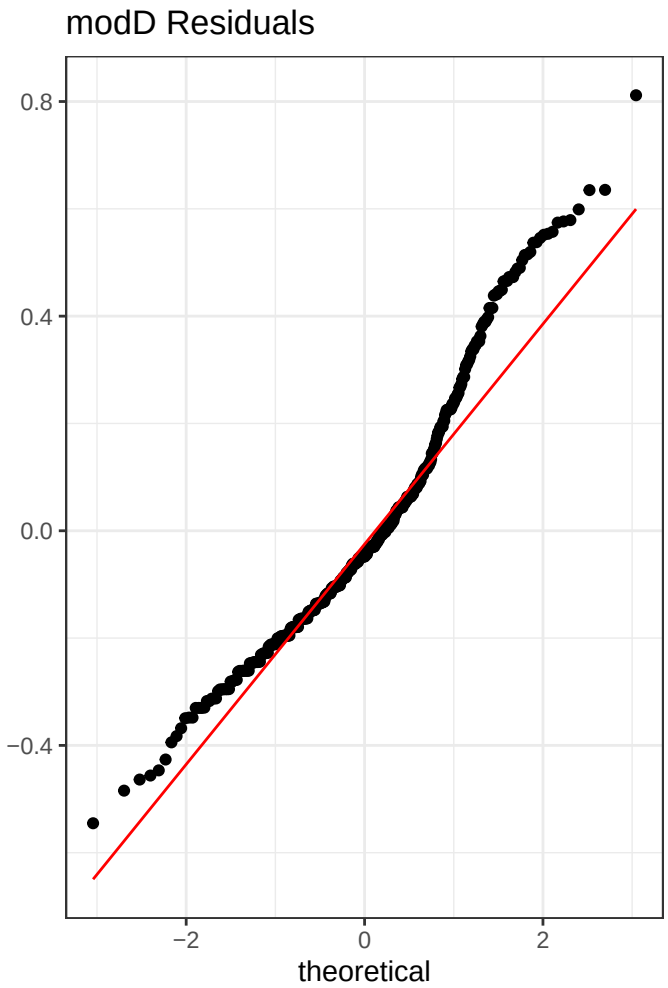
Making Predictions with augment

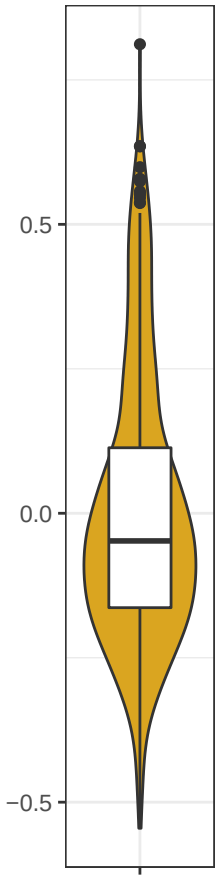
```
augD <- augment(modD, dm431)

augD %>% select(subject, insurance, dbp, .fitted, .resid) %>%
  head() %>% kable(dig = 2)
```

subject	insurance	dbp	.fitted	.resid
S-001	Commercial	64	73.09	-9.09
S-002	Uninsured	84	75.86	8.14
S-003	Uninsured	95	75.86	19.14
S-004	Commercial	87	73.09	13.91
S-005	Commercial	58	73.09	-15.09
S-006	Medicare	60	70.35	-10.35

Residual Plots for `modD`







That's the end of the material I expect you to use in Project A. All remaining slides were moved to Class 16.