

431 Classes 18-20

thomaselove.github.io/431

2021-10-26

Agenda for Classes 18-20

- Power / Sample Size Decisions and The March of Science
- Multiple Regression using the dm1 data
 - Using df_stats to get favstats for multiple variables at once
 - Using the naniar package to identify and summarizing missingness
 - Complete Cases and Simple imputation to deal with missingness
 - Partitioning our data into training/test samples
 - Outcome transformation: what to consider
 - Assessing the fit in the sample where we build the model
 - Using tidy to describe model coefficients
 - Using glance to study fit quality
 - Using augment to obtain predicted values and residuals
 - Residual plots to check assumptions with plot and with ggplot2
- Testing the model in new data (a holdout sample)
 - Assessing the predictions and prediction errors
 - Back-transformation and MAPE, root MSPE and max error

Schedule

We will discuss this material over Classes 18-20.

Our R Setup

```
knitr::opts_chunk$set(comment = NA)
options(dplyr.summarise.inform = FALSE)

library(simputation) # for single imputation
library(car) # for boxCox
library(GGally) # for ggpairs
library(ggrepel) # help with residual plots
library(equatiomatic) # help with equation extraction
library(knitr); library(janitor); library(magrittr)
library(patchwork); library(broom); library(naniar)
library(tidyverse)

theme_set(theme_bw())
```

On Power / Sample Size Decisions and the March of Science

The March of Science and Power Calculations

I mentioned last time that the most common scenario was to identify:

- desired significance level α
- desired power $(1 - \beta)$

and the details of the plan in terms of what comparison is to be made, and how will the data be collected to support that comparison.

This will then permit the calculation of a minimum necessary sample size to achieve these desires.

I also mentioned that $\alpha = 0.05$ and $\beta = 0.2$ were the most common selections.

- Neither 95% confidence nor 80% power is a magical choice.
- Anything below 80% power will be hard to justify in real work.

A useful metaphor?

Sometimes I like to think of science as a march towards a destination.

Actually, I suppose it's an infinitely long march towards an ever-receding destination, but let's leave the philosophy out of it for a moment.

Suppose, for example, that we're trying to make a meaningful change in the world, perhaps to treat an infection.

What we're trying to do is related to where we are in the March of Science.

Early vs. Late in the March of Science

In **early** work, we're focused more on discovery than making final decisions.

- We don't have a lot of past experience, so we bring little relevant data to the table.
- We're (often) most concerned about discovering new possibilities, and we don't have a very clear sense of where to go next.
- We're (often) less concerned about false starts than we are about missed opportunities.

In **late** work, we're focused more on making a decision about how to treat.

- We have a fair amount of relevant history to draw on, sometimes quite detailed.
- We're more concerned about testing the limits of our current knowledge than we are about missing opportunities to consider a new pathway.
- We're often concerned about doing harm if we implement the strategy that looks most promising.

How does this relate to power and significance?

If we treat our sample size and study design as fixed strategies, then there is a tradeoff between:

- reducing α , the rate of Type I error (increasing our confidence) and
- reducing β , the rate of Type II error (increasing our power)

Suppose we are testing a new treatment for some condition.

- A Type I error means we conclude this treatment is helpful, when it actually isn't.
- A Type II error means we conclude this treatment is not helpful, when it actually is.

Early Work: Power and Sample Size

In early work, we are searching for treatments of promise, and our initial study will inevitably not be the last word on the subject, but rather will be followed up by confirmatory studies. In such a setting, it is often the case that:

- We're not so concerned about getting results that cause us to continue to explore a treatment that doesn't actually do what we need it to do.
- We're really concerned about ruling out a treatment that is promising before we should.

This implies we should prioritize reducing Type II error rates (we want more power to detect small but real effects, even if this means we will occasionally identify something as promising when it isn't.)

This means setting lower confidence levels and higher power levels, potentially, than the standard 95% confidence and 80% power.

Late Work: Power and Sample Size

In late work, we have already identified promising treatments, and we are trying to confirm those results. The current study may actually be the last word on the subject, and we want to be sure we do no harm.

- We're very concerned about getting results that cause us to continue to explore a treatment that doesn't actually do what we need it to do.
- We're less concerned about ruling out a treatment that is promising but doesn't actually work.

This implies we should prioritize reducing Type I error rates (we want greater confidence, even at the expense of power, that the effect we claim based on past data holds up.)

This kind of confirmatory work is usually well suited to studies set up with higher confidence (perhaps 95% or 99% or more) and lower power (80% is the minimum I would recommend) against reasonable alternatives.

Conclusions

- ① If you're early in the March of Science (perhaps just one pilot study has been done) then I would emphasize Type II error (power) more than usual, perhaps pushing required power to 90%, at least for a reasonably substantial "minimum scientifically important difference" δ .
- ② If you're late in the March of Science (perhaps confirming the results of multiple prior studies) then I would be happier with 80% power and higher levels of confidence.
- ③ If you're in the middle, trading off Type I and Type II error is worth some thought, but I'd never recommend being under 80% power for an effect that matters.
- ④ If it's feasible to run a study large enough to have strong performance on both α and β , that's obviously ideal. Typically that doesn't happen in early work.

Multiple Regression with the dm1 data

The dm1 data: Four Variables (+ Subject)

Suppose we want to consider predicting the a1c values of 500 diabetes subjects now, based on these three predictors:

- a1c_old: subject's Hemoglobin A1c (in %) two years ago
- age: subject's age in years
- income: median income of subject's home neighborhood (3 categories)

```
dm1 <- readRDS("data/dm1.Rds")
```

```
head(dm1, 3)
```

```
# A tibble: 3 x 5
  a1c a1c_old    age income      subject
  <dbl>   <dbl>   <dbl> <fct>       <chr>
1   6.3     11.4     62 Higher_than_50K S-001
2   11      16.3     54 Between_30-50K  S-002
3   8.7     10.7     47 <NA>        S-003
```

Summarizing the dm1 tibble

```
summary(dm1)
```

	a1c	a1c_old	age
Min.	: 4.300	Min. : 4.200	Min. :31.00
1st Qu.	: 6.500	1st Qu.: 6.500	1st Qu.:49.00
Median	: 7.300	Median : 7.300	Median :56.00
Mean	: 7.898	Mean : 7.693	Mean :55.41
3rd Qu.	: 8.600	3rd Qu.: 8.300	3rd Qu.:62.00
Max.	:16.700	Max. :16.300	Max. :70.00
NA's	:4	NA's :15	
	income	subject	
Higher_than_50K	:123	Length:500	
Between_30-50K	:194	Class :character	
Below_30K	:178	Mode :character	
NA's	:	5	

What roles will these variables play?

a1c is our outcome, which we'll predict using three models . . .

- ① Model 1: Use a1c_old alone to predict a1c
- ② Model 2: Use a1c_old and age together to predict a1c
- ③ Model 3: Use a1c_old, age, and income together to predict a1c

df_stats to get favstats on multiple quantities

Suppose we want favstats results on all 3 quantitative variables.

```
dm1 %>%  
  mosaicCore::df_stats(~ a1c + a1c_old + age) %>%  
  rename(na = missing) %>% kable(dig = 2)
```

response	min	Q1	median	Q3	max	mean	sd	n	na
a1c	4.3	6.5	7.3	8.6	16.7	7.90	2.06	496	4
a1c_old	4.2	6.5	7.3	8.3	16.3	7.69	1.75	485	15
age	31.0	49.0	56.0	62.0	70.0	55.41	9.02	500	0

- The df_stats function is part of the mosaicCore package.
- Either use library(mosaic) to make df_stats() available, or specify it with mosaicCore::df_stats().

What will we do about missing data?

```
dm1 %>%
  summarize(across(everything(), ~ sum(is.na(.)))) %>%
  kable()
```

a1c	a1c_old	age	income	subject
4	15	0	5	0

- We're missing 4 values of a1c, our outcome
- and 15 values of a1c_old, a predictor (Models 1-3)
- and 5 values of income, another predictor (Model 3)

But what if we have other questions, like:

- How many observations are missing at least one of these variables?
- How many subjects (cases) are missing multiple variables?

Counting missingness by subject with naniar

The `naniar` package provides several useful functions for identifying and summarizing missingness which work within a tidy workflow.

`miss_case_table()` for instance, provides a summary table describing the number of subjects missing 0, 1, 2, ... of the variables in our tibble.

```
miss_case_table(dm1)
```

```
# A tibble: 3 x 3
  n_miss_in_case n_cases pct_cases
            <int>    <int>     <dbl>
1                 0      479     95.8
2                 1       18      3.6
3                 2        3      0.6
```

So, there are 18 subjects missing one variable, and 3 missing two.

Can we identify these cases?

miss_case_summary lists missingness for each subject

```
miss_case_summary(dm1)
```

```
# A tibble: 500 x 3
  case n_miss pct_miss
  <int>   <int>     <dbl>
1    1      2       40
2    2      2       40
3    3      2       40
4    4      1       20
5    5      1       20
6    6      1       20
7    7      1       20
8    8      1       20
9    9      1       20
10   10     1       20
# ... with 490 more rows
```

Can we summarize missingness by variable?

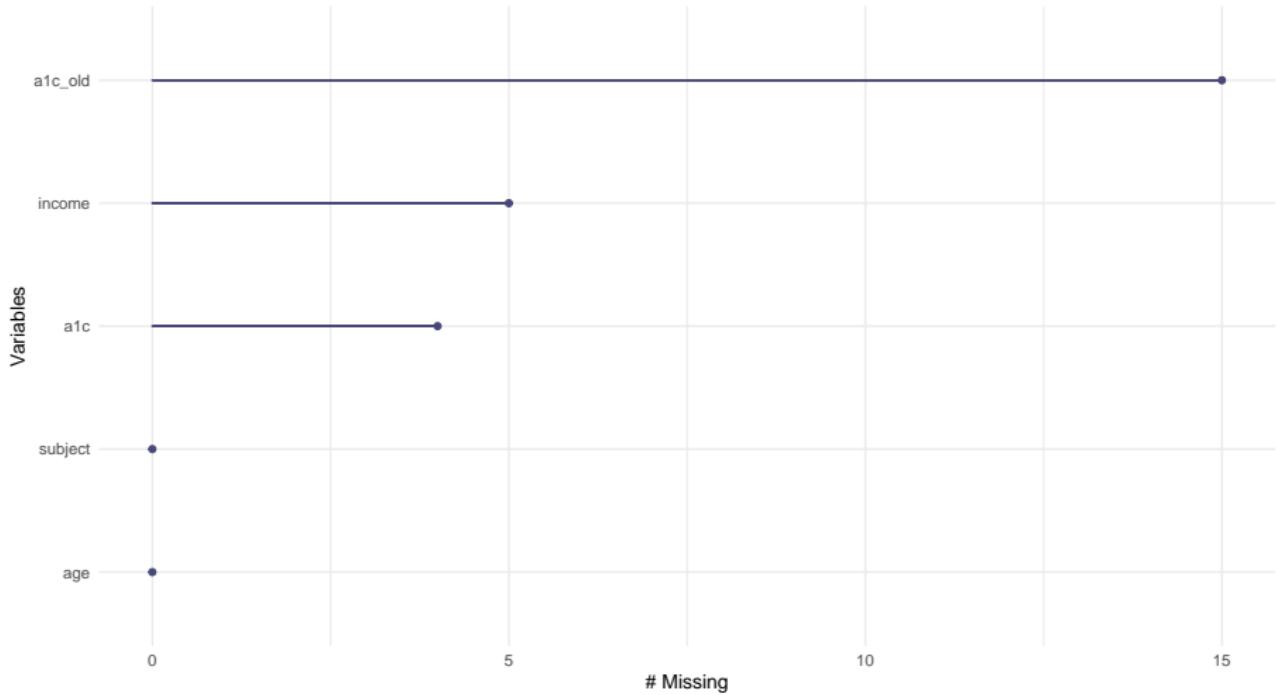
```
miss_var_summary(dm1)
```

```
# A tibble: 5 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 a1c_old      15       3
2 income        5       1
3 a1c           4      0.8
4 age            0       0
5 subject        0       0
```

There's a `miss_var_table()` function, too, if that's useful.

naniar also has helpers for plots

```
gg_miss_var(dm1)
```



Option 1: Complete Cases Only

We might assume that all of our missing values are Missing Completely At Random (MCAR) and thus that we can safely drop all observations with missing data from our data set.

```
dm1_cc <- dm1 %>% filter(complete.cases(.))
```

```
nrow(dm1)
```

```
[1] 500
```

```
nrow(dm1_cc)
```

```
[1] 479
```

- For today, I want to use the same observations in each of my 3 models.
- So we would drop 21 subjects, and fit models with the 479 subjects who have complete data on all four variables.

Simple Imputation with the simputation package

Option 2: Simple Imputation

Suppose I don't want to impute the outcome. I think people missing my outcome shouldn't be included in my models.

- We'll drop the 4 observations missing a1c from our data set.

Perhaps I'd be OK with assuming the missing values of income or a1c_old are MAR (so that we could use variables in our data to predict them.)

- This would allow us to use imputation methods to "fill in" or "impute" missing predictor values so that we can still use all of the other 496 subjects in our models.
- The simputation package provides a straightforward method to do this, while maintaining a tidy workflow.
- There are dangers in assuming everything is MCAR, so this looks helpful (MAR is a lesser assumption) but it introduces the issue of "creating" data where it didn't exist.

Simple Imputation of Missing a1c_old Values

We could use a robust linear model method to impute our quantitative a1c_old values on the basis of age, which is missing no observations in common with a1c_old (in fact, age is missing no observations.)

```
tempA <- impute_rlm(dm1, a1c_old ~ age)
```

```
tempA %>% miss_var_summary()
```

```
# A tibble: 5 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 income      5        1
2 a1c         4        0.8
3 a1c_old     0        0
4 age         0        0
5 subject     0        0
```

Simple Imputation of Missing income Values

We could use a decision tree (CART) method to impute our missing categorical income values, also on the basis of age.

```
tempB <- impute_cart(dm1, income ~ age)
```

```
tempB %>% miss_var_summary()
```

```
# A tibble: 5 x 3
```

	variable	n_miss	pct_miss
	<chr>	<int>	<dbl>
1	a1c_old	15	3
2	a1c	4	0.8
3	age	0	0
4	income	0	0
5	subject	0	0

Chaining our Simple Imputations

Or we could put all of our imputations together in a chain.

- In 431, I encourage you to try `rlm` for imputing quantitative variables, and `cart` for categorical variables.
- Were I imputing a binary categorical variable, I would present it as a factor to `impute_cart`.

```
dm1_imp <- dm1 %>%
  filter(complete.cases(a1c, subject)) %>%
  impute_rlm(a1c_old ~ age) %>%
  impute_cart(income ~ age + a1c_old)
```

- I imputed `a1c_old` using `age` and then imputed `income` using both `age` and `a1c_old`.

What is the result?

Summary of imputed tibble

dm1_imp has 496 observations (since we dropped the 4 subjects with missing a1c: our *outcome*) but no missing values left.

```
dm1_imp %>% summary()
```

a1c	a1c_old	age
Min. : 4.300	Min. : 4.200	Min. :31.00
1st Qu.: 6.500	1st Qu.: 6.500	1st Qu.:49.00
Median : 7.300	Median : 7.300	Median :56.00
Mean : 7.898	Mean : 7.691	Mean :55.35
3rd Qu.: 8.600	3rd Qu.: 8.300	3rd Qu.:62.00
Max. :16.700	Max. :16.300	Max. :70.00
income	subject	
Higher_than_50K:121	Length:496	
Between_30-50K :193	Class :character	
Below_30K :182	Mode :character	

Two approaches for dealing with missing data

- ① We could assume MCAR for all variables, and then work with the complete cases ($n = 479$) in `dm1_cc`.
- ② We could assume MAR for the predictors, and work with the simply imputed ($n = 496$) in `dm1_imp`

Neither of these, as it turns out, will be 100% satisfactory, but for now, we'll compare the impact of these two approaches on the results of our models.

OK. We'll do the complete case analysis now,
and return to the imputed data later.

How will we decide which of the models is “best”?

Our goal is accurate prediction of a1c values.

Which of these models gives us the “best” result?

- ① Model 1: Use a1c_old alone to predict a1c
- ② Model 2: Use a1c_old and age together to predict a1c
- ③ Model 3: Use a1c_old, age, and income together to predict a1c

and does our answer change depending on whether we start our work with the complete cases (`dm1_cc`: $n = 479$) or our simply imputed data (`dm1_imp`: $n = 496$)?

How shall we be guided by our data?

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience. (A. Einstein)

- often this is reduced to “make everything as simple as possible but no simpler”

Entities should not be multiplied without necessity. (Occam's razor)

- often this is reduced to “the simplest solution is most likely the right one”

George Box's aphorisms

On Parsimony: Since all models are wrong the scientist cannot obtain a “correct” one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

On Worrying Selectively: Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

- and, the most familiar version . . .

... all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.

431 approach: Which model is “most useful”?

- ① Split the data into a development (model training) sample of about 70-80% of the observations, and a holdout (model test) sample, containing the remaining observations.
- ② Develop candidate models using the development sample.
- ③ Assess the quality of fit for candidate models within the development sample.
- ④ Check adherence to regression assumptions in the development sample.
- ⑤ When you have candidates, assess them based on the accuracy of the predictions they make for the data held out (and thus not used in building the models.)
- ⑥ Select a “final” model for use based on the evidence in steps 3, 4 and especially 5.

Split the data into a model development (training) sample of about 70-80% of the observations, and a model test (holdout) sample, containing the remaining observations.

Partitioning the 479 Complete Cases

- We'll select a random sample (without replacement) of 70% of the data (70-80% is customary) for model training.
- We'll hold out the remaining 30% for model testing, using `anti_join()` to identify all `dm1_cc` subjects not in `dm1_cc_train`.

```
set.seed(20211026)
```

```
dm1_cc_train <- dm1_cc %>%
  slice_sample(prop = 0.7, replace = FALSE)
```

```
dm1_cc_test <-
  anti_join(dm1_cc, dm1_cc_train, by = "subject")
```

```
c(nrow(dm1_cc_train), nrow(dm1_cc_test), nrow(dm1_cc))
```

```
[1] 335 144 479
```

Develop candidate models using the development sample.

A look at the outcome (a1c) distribution

We'll study the outcome variable (a1c) in the development sample, to consider whether a transformation might be in order.

I did a little fancy work with the code (continues next slide)...

```
p1 <- ggplot(dm1_cc_train, aes(x = a1c)) +  
  geom_histogram(binwidth = 0.5,  
                 fill = "slateblue", col = "white")  
  
p2 <- ggplot(dm1_cc_train, aes(sample = a1c)) +  
  geom_qq(col = "slateblue") + geom_qq_line(col = "red")  
  
p3 <- ggplot(dm1_cc_train, aes(x = "", y = a1c)) +  
  geom_violin(fill = "slateblue", alpha = 0.3) +  
  geom_boxplot(fill = "slateblue", width = 0.3,  
                outlier.color = "red") +  
  labs(x = "") + coord_flip()
```

A look at the outcome (a1c) distribution

Putting the plots together, and titling them meaningfully...

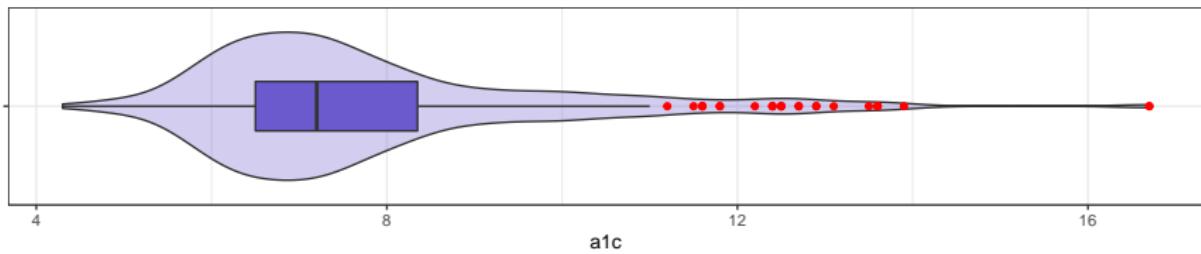
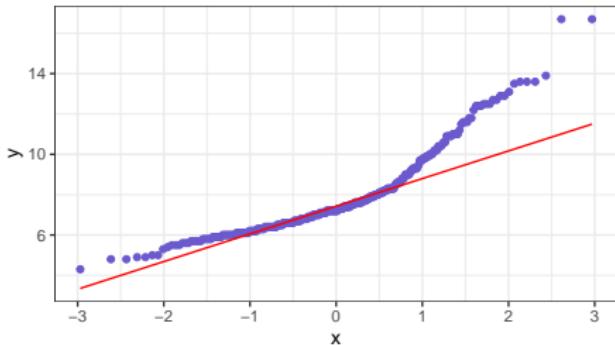
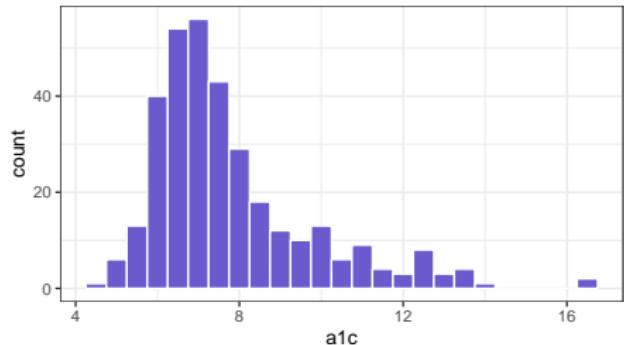
```
p1 + p2 - p3 +  
  plot_layout(ncol = 1, height = c(3, 2)) +  
  plot_annotation(title = "Hemoglobin A1c values (%)",  
                  subtitle = paste0("Model Development Sample: ",  
                                     nrow(dm1_cc_train),  
                                     " adults with diabetes"))
```

Result on the next slide...

Outcome (a1c): Model Development Sample

Hemoglobin A1c values (%)

Model Development Sample: 335 adults with diabetes



Why Transform the Outcome?

We want to try to identify a good transformation for the conditional distribution of the outcome, given the predictors, in an attempt to make the linear regression assumptions of linearity, Normality and constant variance more appropriate.

Ladder of Especially Useful (and often interpretable) transformations

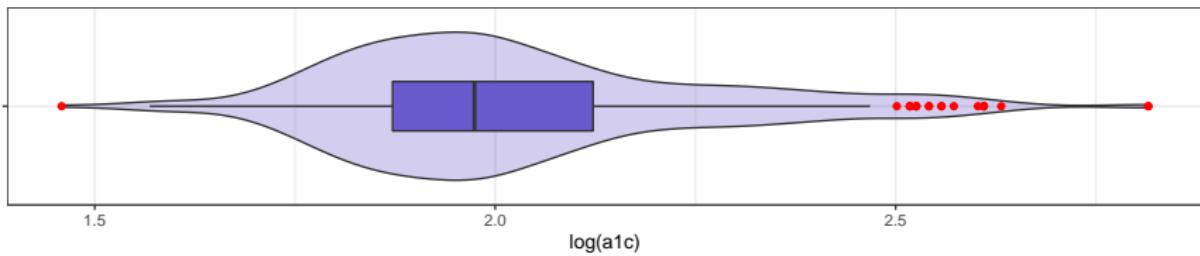
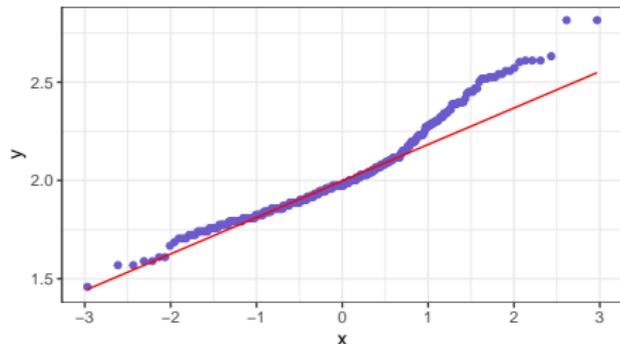
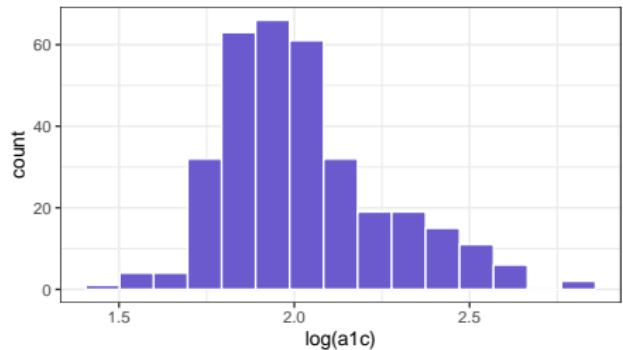
Transformation	y^2	y	\sqrt{y}	$\log(y)$	$1/y$	$1/y^2$
λ	2	1	0.5	0	-1	-2

- We see some sign of right skew in the a1c data. Let's try a log transformation.

Consider a log transformation?

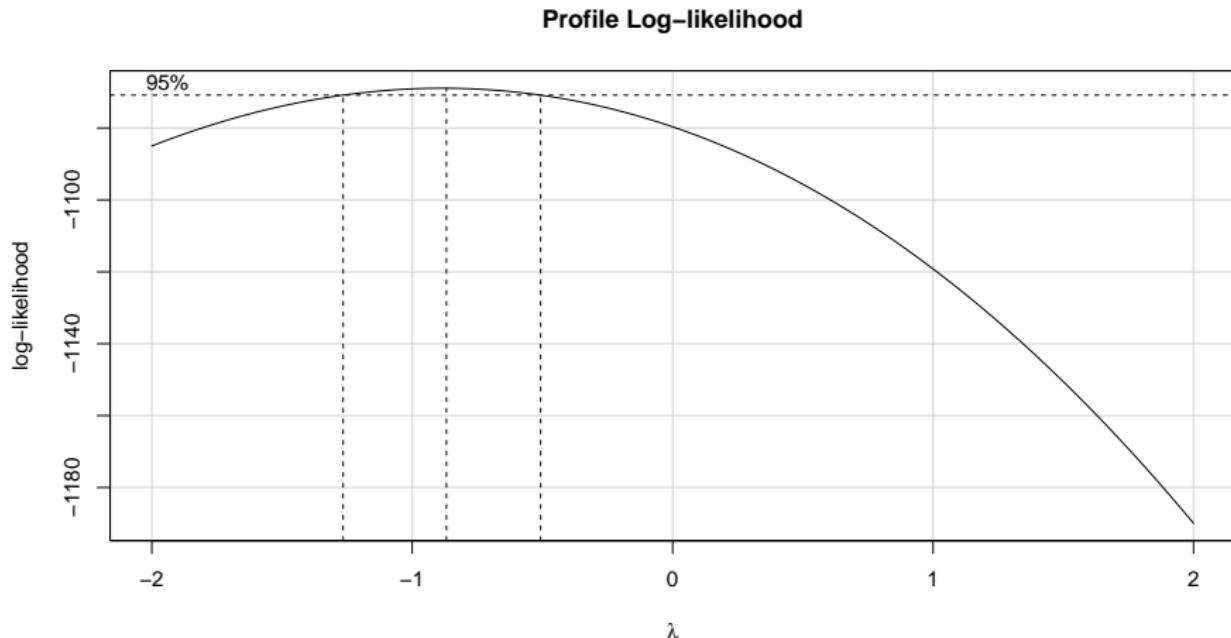
Natural Logarithm of Hemoglobin A1c

Model Development Sample: 335 adults with diabetes



Using Box-Cox to help select a transformation?

```
mod_0 <- lm(a1c ~ a1c_old + age + income,  
             data = dm1_cc_train)  
boxCox(mod_0)
```



Using Box-Cox to help select a transformation?

```
summary(powerTransform(mod_0))
```

bcPower Transformation to Normality

	Est	Power	Rounded	Pwr	Wald	Lwr	Bnd	Wald	Upr	Bnd
Y1	-0.8838			-1			-1.2628			-0.5048

Likelihood ratio test that transformation parameter is equal to
(log transformation)

	LRT	df	pval
LR test, lambda = (0)	21.48071	1	3.5741e-06

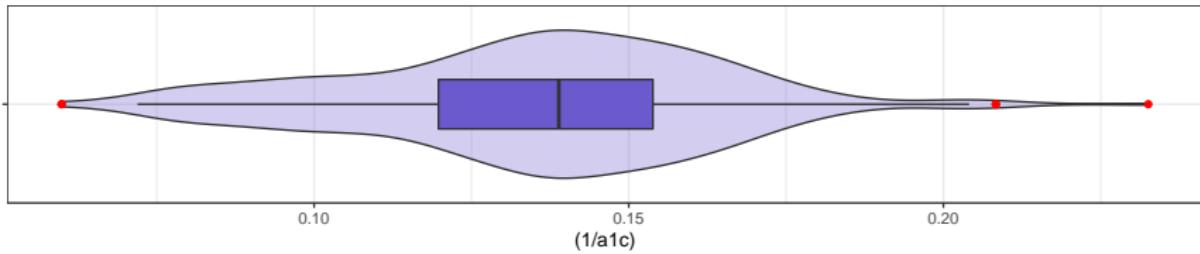
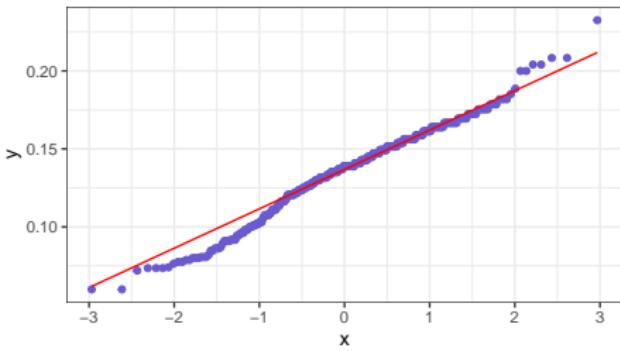
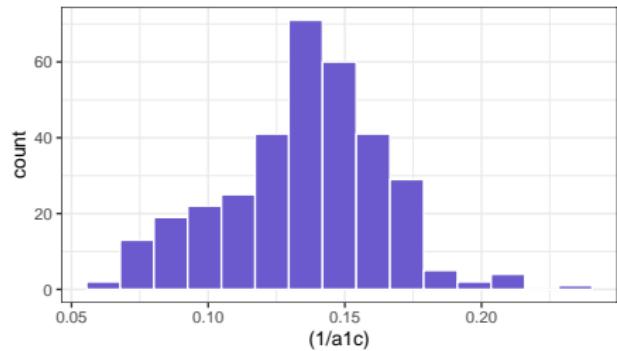
Likelihood ratio test that no transformation is needed

	LRT	df	pval
LR test, lambda = (1)	100.5543	1	< 2.22e-16

Consider the inverse?

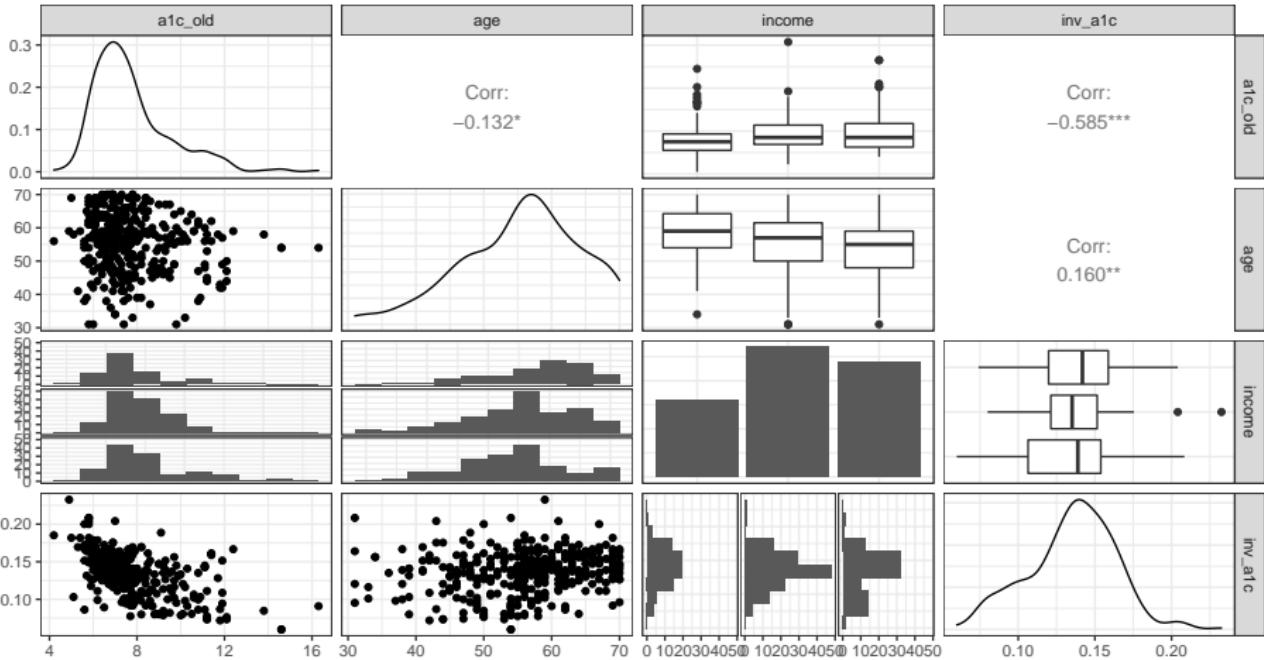
Inverse of Hemoglobin A1c

Model Development Sample: 335 adults with diabetes



Scatterplot Matrix (code on next slide)

Scatterplots: Model Development Sample



Scatterplot Matrix (Code)

```
dm1_cc_train %>%  
  mutate(inv_a1c = 1/a1c) %>%  
  select(a1c_old, age, income, inv_a1c) %>%  
  ggpairs(.,  
    title = "Scatterplots: Model Development Sample",  
    lower = list(combo = wrap("facethist", bins = 10)))
```

Note that `ggpairs` comes from the `GGally` package.

- If you have more than 4-5 predictors, it's usually necessary to split this up into two or more scatterplot matrices, each of which should include the outcome.
- I'd always put the outcome last in my selection here. That way, the bottom row will show the most important scatterplots, with the outcome on the Y axis, and each predictor, in turn on the X.

Three Regression Models We'll Fit

- Remember we're using the model development sample here.
- Let's work with the $(1/a1c)$ transformation.

```
mod_1 <- lm((1/a1c) ~ a1c_old, data = dm1_cc_train)
```

```
mod_2 <- lm((1/a1c) ~ a1c_old + age, data = dm1_cc_train)
```

```
mod_3 <- lm((1/a1c) ~ a1c_old + age + income,  
             data = dm1_cc_train)
```

**Assess the quality of fit for candidate models
within the development sample.**

Tidied coefficients (mod_1)

```
tidy_m1 <- tidy(mod_1, conf.int = TRUE, conf.level = 0.95)

tidy_m1 %>%
  select(term, estimate, std.error, p.value,
         conf.low, conf.high) %>%
knitr::kable(digits = 4)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	0.2080	0.0057	0	0.1968	0.2192
a1c_old	-0.0094	0.0007	0	-0.0108	-0.0080

The Regression Equation (mod_1)

Use the `equatiomatic` package to help here. Note the use of `results = 'asis'` in the code chunk name.

```
extract_eq(mod_1, use_coefs = TRUE, coef_digits = 4,  
           italic_vars = TRUE)
```

$$\widehat{(\text{1}/\text{a1c})} = 0.208 - 0.0094(\text{a1c_old}) \quad (1)$$

Summary of Fit Quality (mod_1)

```
glance(mod_1) %>%
  mutate(name = "mod_1") %>%
  select(name, r.squared, adj.r.squared,
         sigma, AIC, BIC) %>%
knitr::kable(digits = c(0, 3, 3, 3, 0, 0))
```

name	r.squared	adj.r.squared	sigma	AIC	BIC
mod_1	0.342	0.34	0.023	-1565	-1553

Tidied coefficients (mod_2)

```
tidy_m2 <- tidy(mod_2, conf.int = TRUE, conf.level = 0.95)

tidy_m2 %>%
  select(term, estimate, std.error, p.value,
        conf.low, conf.high) %>%
knitr::kable(digits = 4)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	0.1913	0.0105	0.0000	0.1707	0.2120
a1c_old	-0.0093	0.0007	0.0000	-0.0107	-0.0078
age	0.0003	0.0001	0.0603	0.0000	0.0006

The Regression Equation (mod_2)

Again, we'll use the `equatiomatic` package, with `results = 'asis'`.

```
extract_eq(mod_2, use_coefs = TRUE, coef_digits = 4,  
           italic_vars = TRUE)
```

$$\widehat{(\text{1}/\text{a1c})} = 0.1913 - 0.0093(\text{a1c_old}) + 3e - 04(\text{age}) \quad (2)$$

Summary of Fit Quality (mod_2)

```
glance(mod_2) %>%
  mutate(name = "mod_2") %>%
  select(name, r.squared, adj.r.squared,
         sigma, AIC, BIC) %>%
knitr::kable(digits = c(0, 3, 3, 3, 0, 0))
```

name	r.squared	adj.r.squared	sigma	AIC	BIC
mod_2	0.349	0.345	0.023	-1566	-1551

Tidied coefficients (mod_3)

```
tidy_m3 <- tidy(mod_3, conf.int = TRUE, conf.level = 0.95)

tidy_m3 %>%
  select(term, estimate, se = std.error,
        low = conf.low, high = conf.high, p = p.value) %>%
  knitr::kable(digits = c(4,4,4,4,3))
```

term	estimate	se	low	high	p
(Intercept)	0.1922	0.0110	0.1707	0.214	0.0000
a1c_old	-0.0092	0.0007	-0.0107	-0.008	0.0000
age	0.0003	0.0001	0.0000	0.001	0.0771
incomeBetween_30-50K	0.0002	0.0033	-0.0063	0.007	0.9456
incomeBelow_30K	-0.0016	0.0034	-0.0084	0.005	0.6384

The Regression Equation (mod_3)

Again, we'll use the `equatiomatic` package.

```
extract_eq(mod_3, use_coefs = TRUE, coef_digits = 4,  
          italic_vars = TRUE, wrap = TRUE, terms_per_line = 2)
```

$$\widehat{(\text{1}/\text{a1c})} = 0.1922 - 0.0092(\text{a1c_old}) + \\ 3e - 04(\text{age}) + 2e - 04(\text{income}_{\text{Between_}30-\text{50K}}) - \quad (3) \\ 0.0016(\text{income}_{\text{Below_}30K})$$

Summary of Fit Quality (mod_3)

```
glance(mod_3) %>%
  mutate(name = "mod_3") %>%
  select(name, r.squared, adj.r.squared,
         sigma, AIC, BIC) %>%
knitr::kable(digits = c(0, 3, 3, 3, 0, 0))
```

name	r.squared	adj.r.squared	sigma	AIC	BIC
mod_3	0.35	0.342	0.023	-1563	-1540

Could we have fit other predictor sets?

Perhaps an automated procedure, like stepwise regression, would suggest a better alternative?

- Three predictor candidates, so we could have used any of these predictor sets:
- a1c_old alone (our mod_1)
- age alone
- income alone
- a1c_old and age (our mod_2)
- a1c_old and income
- age and income
- a1c_old, age and income (our mod_3)

step(mod_3)

Stepwise Regression Results (part 1 of 2)

We'll try backwards elimination, where we let R's `step` function start with the full model (`mod_3`) including all three predictors, and then remove the predictor whose removal causes the largest drop in AIC, until we reach a point where eliminating another predictor will not improve the AIC.

- Remember the smaller (more negative, here) the AIC, the better.

```
step(mod_3)
```

Start: AIC=-2515.5
(1/a1c) ~ a1c_old + age + income

	Df	Sum of Sq	RSS	AIC
- income	2	0.000236	0.17847	-2519.1
<none>			0.17823	-2515.5
- age	1	0.001698	0.17993	-2514.3
- a1c_old	1	0.086785	0.26502	-2384.6

Stepwise Regression Results (part 2 of 2)

Step: AIC=-2519.05
 $(1/\text{a1c}) \sim \text{a1c_old} + \text{age}$

	Df	Sum of Sq	RSS	AIC
<none>		0.17847	-2519.1	
- age	1	0.00191	0.18038	-2517.5
- a1c_old	1	0.08858	0.26705	-2386.0

Call:

```
lm(formula = (1/a1c) ~ a1c_old + age, data = dm1_cc_train)
```

Coefficients:

(Intercept)	a1c_old	age
0.1913429	-0.0092565	0.0002749

and we wind up here with just our mod_2.

An Important Point

- There is an **enormous** amount of evidence that variable selection causes severe problems in estimation and inference.
- Stepwise regression, in particular, is an egregiously bad choice.
- Disappointingly, there really isn't a good choice. The task itself just isn't one we can do well in a uniform way across all of the different types of regression models we'll build.

More on this in 432.

Comparing Summary Measures of Fit

in the development (model training) sample...

```
bind_rows(glance(mod_1), glance(mod_2), glance(mod_3)) %>%
  mutate(model_vars = c("1_a1c_old", "2_age", "3_income"))
  select(model_vars, r2 = r.squared, adj_r2 = adj.r.squared,
         sigma, AIC, BIC, df, df_res = df.residual) %>%
  kable(digits = c(0, 4, 4, 5, 1, 0, 0, 0))
```

model_vars	r2	adj_r2	sigma	AIC	BIC	df	df_res
1_a1c_old	0.3418	0.3398	0.02327	-1564.8	-1553	1	333
2_age	0.3487	0.3448	0.02319	-1566.4	-1551	2	332
3_income	0.3496	0.3417	0.02324	-1562.8	-1540	4	330

In the data we used to build the model, these are our results.

Which Model Looks Best In-Sample?

For each of these summaries, which model looks best in the training sample?

model	vars	r2	adj_r2	sigma	AIC	BIC
mod_1	a1c_old	0.3418	0.3398	0.02327	-1564.8	-1553
mod_2	+ age	0.3487	0.3448	0.02319	-1566.4	-1551
mod_3	+ income	0.3496	0.3417	0.02324	-1562.8	-1540

- By r^2 , the largest model will always look best (raw r^2 is greedy)
- Adjusted r^2 penalizes for lack of parsimony. Model 2 looks best there.
- For σ , AIC and BIC, we want small (more negative) values.
 - Model 2 looks best by σ and AIC, as well.
 - Model 1 looks a little better than Model 2 by BIC.
- Overall, what should we conclude about in-sample fit quality?

**Check adherence to regression assumptions in
the development sample.**

Using augment to add fits, residuals, etc.

```
aug1 <- augment(mod_1, data = dm1_cc_train) %>%  
  mutate(inv_a1c = 1/a1c) # add in our model's outcome
```

aug1 includes all variables in dm_cc_train to which we've added:

- inv_a1c = $1/a1c$, the transformed outcome that mod_1 predicts
- .fitted = fitted (predicted) values of $1/a1c$
- .resid = residual (observed outcome - fitted outcome) values, so that larger values (positive or negative) mean poorer fit points
- .std.resid = standardized residuals (residuals scaled to SD = 1, remember that the residual mean is already 0)
- .hat statistic = measures leverage (larger values of .hat indicate unusual combinations of predictor values)
- .cooksrd = Cook's distance (or Cook's d), a measure of the subject's influence on the model (larger Cook's d values indicate that removing the point will materially change the model's coefficients)
- plus .sigma = estimated σ if this point is dropped from the model

augment results for the first 2 subjects

```
aug1 %>% select(subject, a1c:income, inv_a1c) %>%  
tail(2) %>% kable(dig = 3)
```

subject	a1c	a1c_old	age	income	inv_a1c
S-060	7.0	7.3	45	Higher_than_50K	0.143
S-116	6.4	6.5	63	Between_30-50K	0.156

```
aug1 %>% select(subject, .fitted:.cooksdi) %>%  
tail(2) %>% kable(dig = 3)
```

subject	.fitted	.resid	.hat	.sigma	.cooksdi
S-060	0.139	0.004	0.003	0.023	0
S-116	0.147	0.010	0.004	0.023	0

augment for models mod_2 and mod_3

We need the augment results for our other two models: mod_2 and mod_3.

```
aug2 <- augment(mod_2, data = dm1_cc_train) %>%  
  mutate(inv_a1c = 1/a1c) # add in our model's outcome
```

```
aug3 <- augment(mod_3, data = dm1_cc_train) %>%  
  mutate(inv_a1c = 1/a1c) # add in our model's outcome
```

Checking Regression Assumptions

Four key assumptions we need to think about:

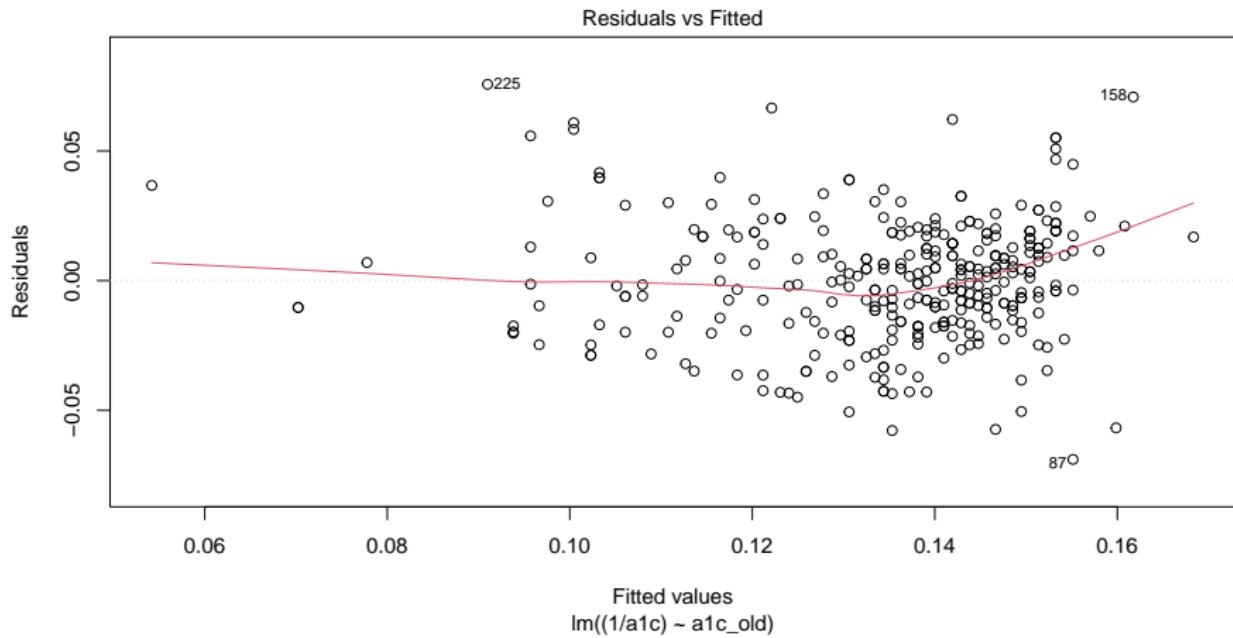
- ① Linearity
- ② Constant Variance (Homoscedasticity)
- ③ Normality
- ④ Independence

How do we assess 1, 2, and 3? Residual plots.

There are five automated ones that we could obtain using `plot(mod_1)...`

Residuals vs. Fitted Values Plot (Model mod_1)

```
plot(mod_1, which = 1)
```



Which points are highlighted in that plot?

Note that the points labeled 87, 158 and 225 are the 87th, 158th and 225th rows in our dm1_cc_train data file, or, equivalently, in our aug1 file.

```
aug1 %>% slice(c(87, 158, 225)) %>% select(a1c:.resid)
```

```
# A tibble: 3 x 7
  a1c a1c_old age income subject .fitted .resid
  <dbl>    <dbl> <dbl> <fct>    <chr>    <dbl>    <dbl>
1 11.6      5.6   54 Below_30K S-168     0.155 -0.0689
2  4.3      4.9   59 Between_~ S-386     0.162  0.0708
3   6       12.4   59 Below_30K S-105     0.0910 0.0757
```

These are subjects S-168, S-386, and S-105, respectively.

Another way to confirm who the plot is identifying

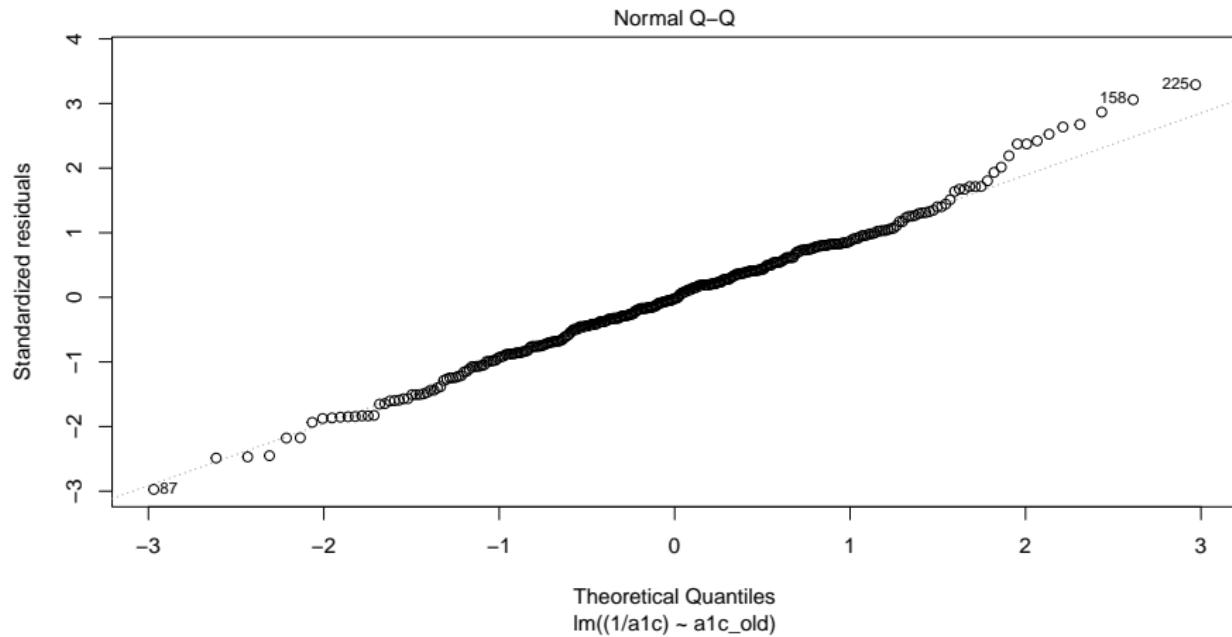
As mentioned, we think the identifiers (87, 158 and 225) of the points with the largest residual (in absolute value) describe subjects S-168, S-386, and S-105, respectively. Does this make sense?

```
aug1 %>% select(subject, .resid) %>%
  arrange(desc(abs(.resid))) %>% head()
```

```
# A tibble: 6 x 2
  subject   .resid
  <chr>     <dbl>
1 S-105    0.0757
2 S-386    0.0708
3 S-168   -0.0689
4 S-071    0.0666
5 S-052    0.0621
6 S-341    0.0609
```

Normal Q-Q of Standardized Residuals (mod_1)

```
plot(mod_1, which = 2)
```



Are the outliers we see there completely out of line?

```
nrow(aug1)
```

```
[1] 335
```

```
aug1 %>% select(subject, .std.resid) %>%
  arrange(desc(abs(.std.resid)))
```

```
# A tibble: 335 x 2
  subject .std.resid
  <chr>     <dbl>
1 S-105      3.29
2 S-386      3.06
3 S-168     -2.97
4 S-071      2.87
5 S-052      2.67
6 S-341      2.64
7 S-001      2.53
8 S-009     -2.49
```

Is a Z score of 3.34 for the biggest outlier scary here?

```
outlierTest(mod_1)
```

No Studentized residuals with Bonferroni p < 0.05

Largest |rstudent|:

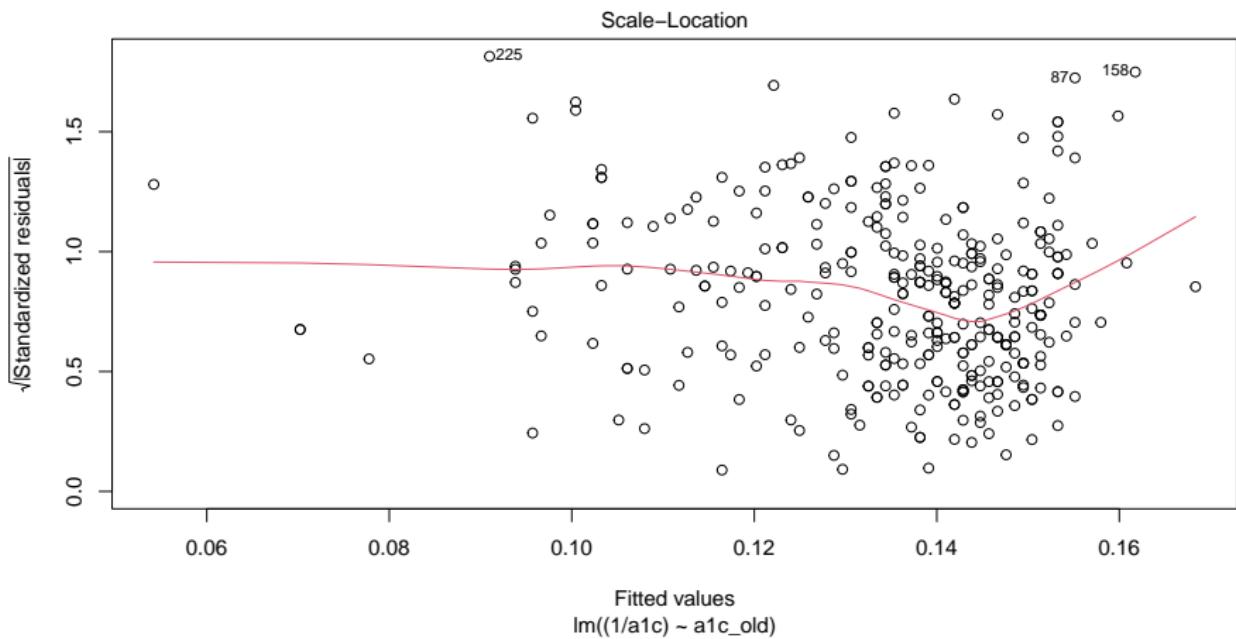
	rstudent	unadjusted p-value	Bonferroni p
225	3.340755	0.00093066	0.31177

For now, a studentized residual is just another way to standardize the residuals that has some useful properties in this setting.

- There's no indication that having a maximum absolute value of 3.34 in a sample of 335 studentized residuals is a major concern about the assumption of Normality, given the Bonferroni p value of 0.31.

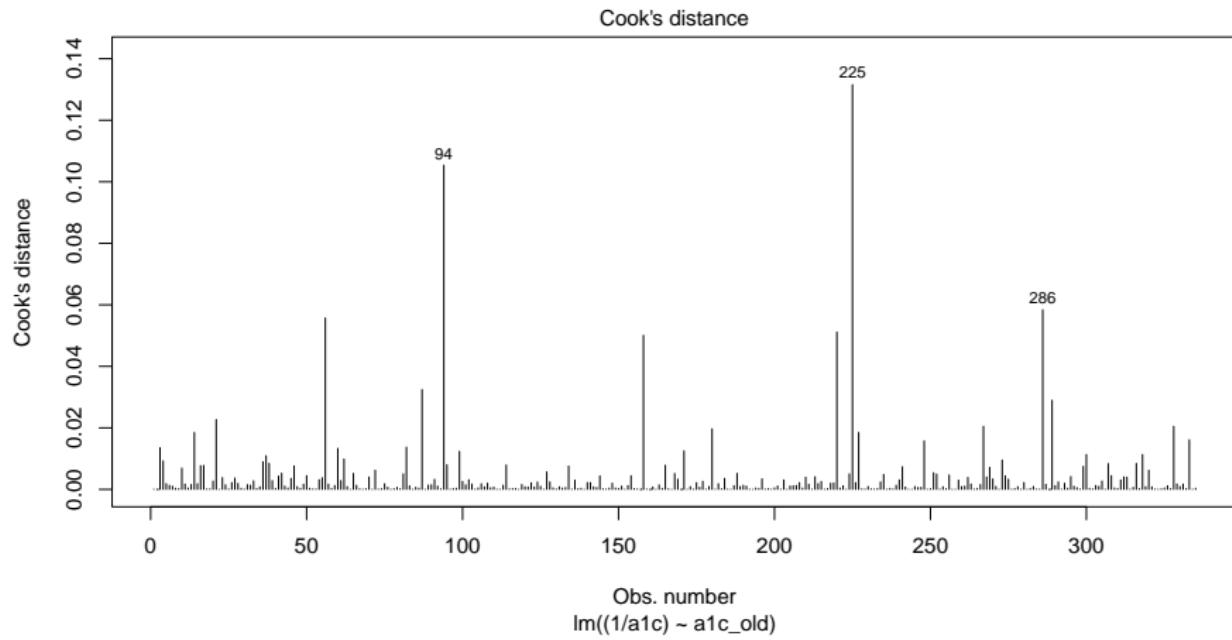
Scale-Location: Check for heteroscedasticity (mod_1)

```
plot(mod_1, which = 3)
```



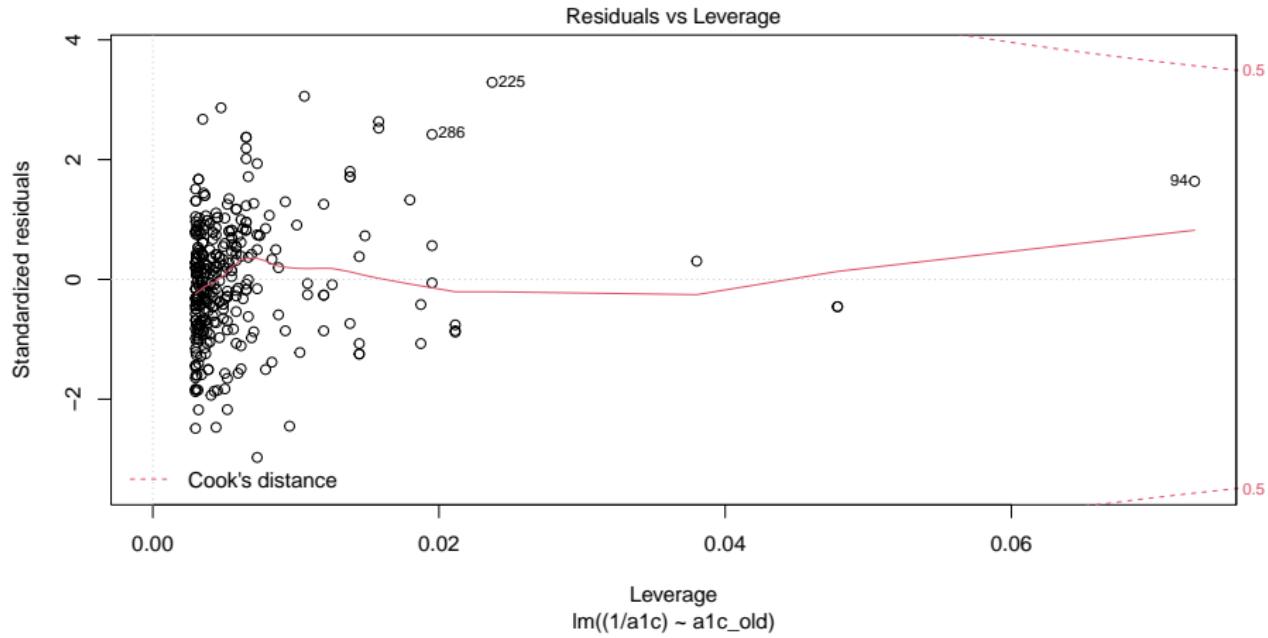
Index plot of Cook's distance for influence (mod_1)

```
plot(mod_1, which = 4)
```



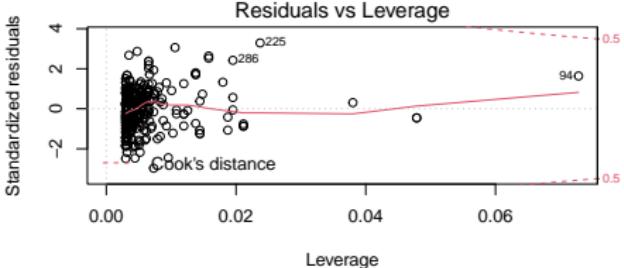
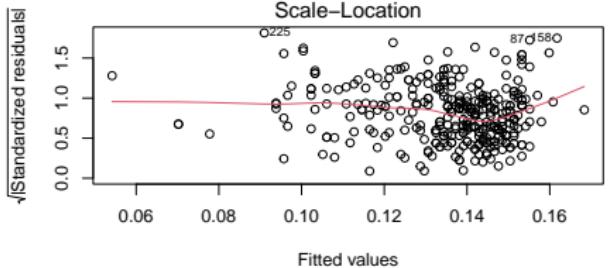
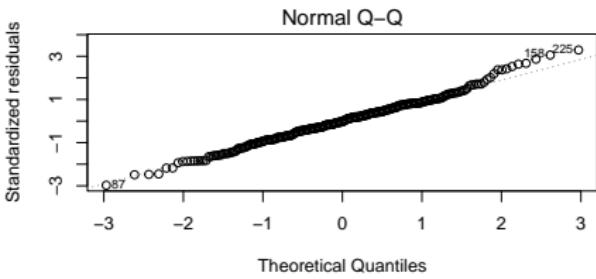
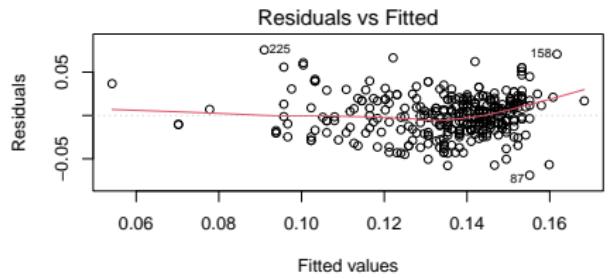
Residuals, Leverage and Influence plot (mod_1)

```
plot(mod_1, which = 5)
```

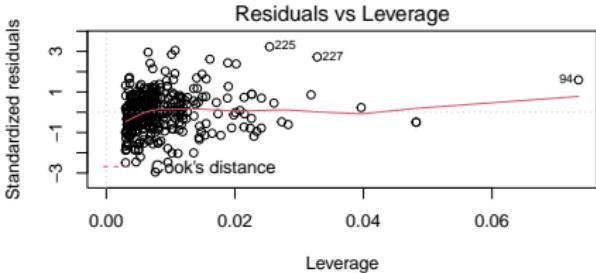
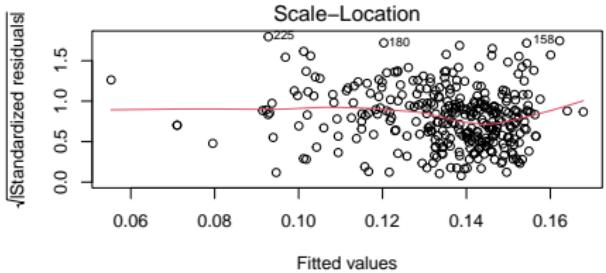
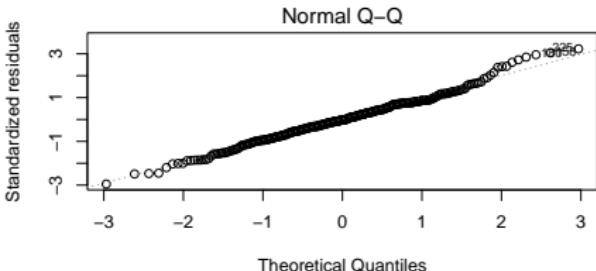
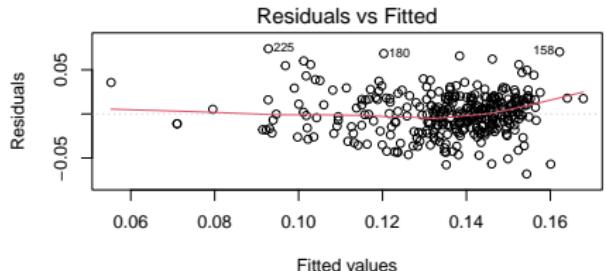


Residual Plots for Model mod_1

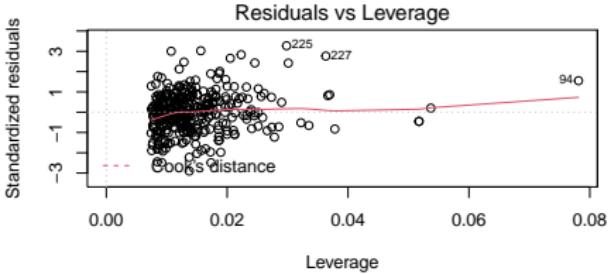
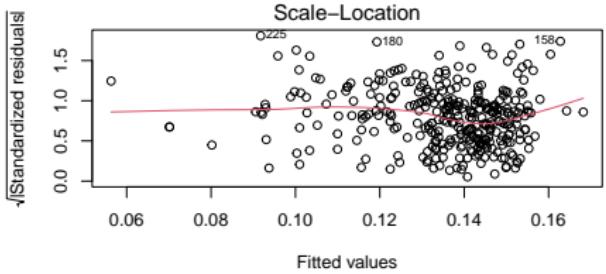
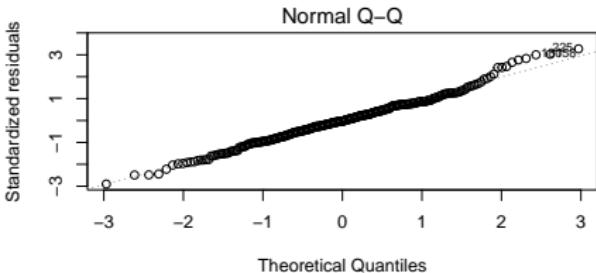
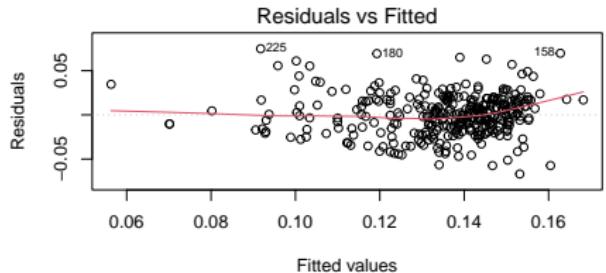
```
par(mfrow = c(2,2)); plot(mod_1); par(mfrow = c(1,1))
```



Residual Plots for Model mod_2



Residual Plots for Model mod_3



Is collinearity a serious issue here?

```
car::vif(mod_3)
```

	GVIF	Df	GVIF ^{(1/(2*Df))}
a1c_old	1.031609	1	1.015682
age	1.050249	1	1.024816
income	1.052156	2	1.012791

- Collinearity = correlated predictors
- (generalized) Variance Inflation Factor tells us something about how the standard errors of our coefficients are inflated as a result of correlation between predictors.
 - We tend to worry most about VIFs in this output that exceed 5.
 - Remember that the scatterplot matrix didn't suggest any strong correlations between our predictors.

What would we do if we had strong collinearity? Drop a predictor?

Conclusions so far?

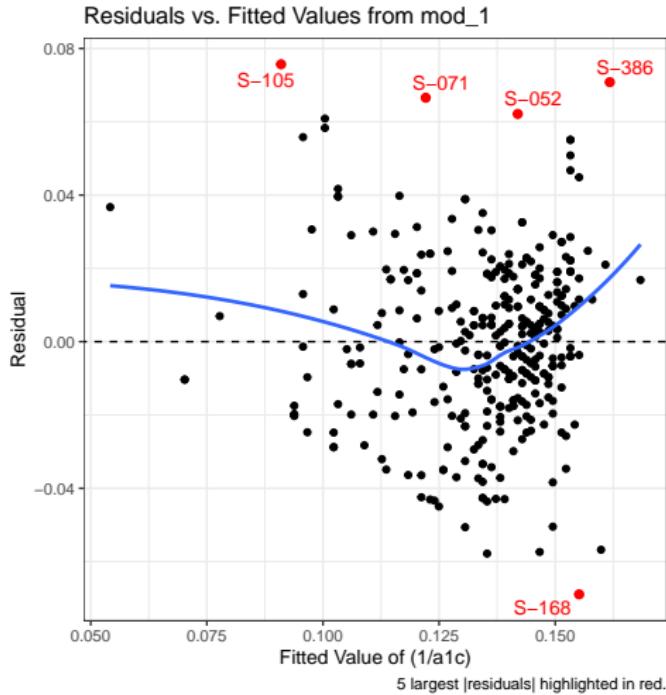
- ① In-sample model predictions are about equally accurate for each of the three models. Model 2 looks better in terms of adjusted R^2 and AIC, but model 1 looks better on BIC. There's really not much to choose from there.
- ② Residual plots look similarly reasonable for linearity, Normality and constant variance in all three models.

Using ggplot2 to create these residual plots?

- ① Residuals vs. Fitted Values plots are straightforward, with the use of the `augment` function from the `broom` package.
- We can also plot residuals against individual predictors, if we like.
- ② Similarly, plots to assess the Normality of the residuals, like a Normal Q-Q plot, are straightforward, and can use either raw residuals or standardized residuals.
- ③ The scale-location plot of the square root of the standardized residuals vs. the fitted values is also pretty straightforward.
- ④ The `augment` function can be used to obtain Cook's distance, standardized residuals and leverage values, so we can mimic both the index plot (of Cook's distance) as well as the residuals vs. leverage plot with Cook's distance contours, if we like.

Demonstrations on the next few slides, followed by the code.

Residuals vs. Fitted Values Plot via ggplot2



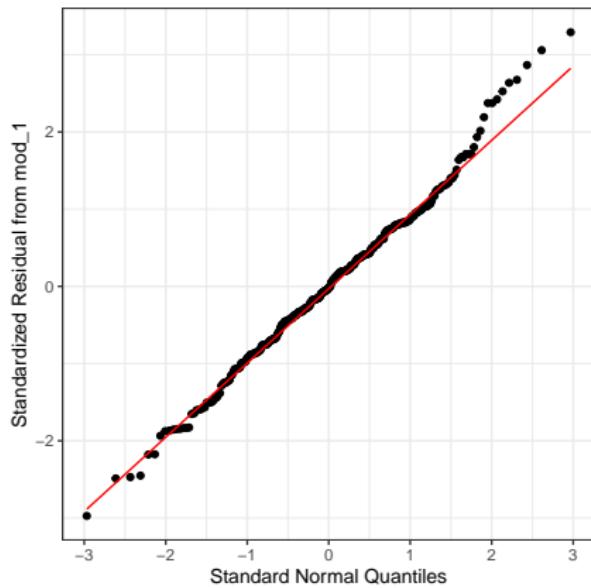
Code for Residuals vs. Fitted Values

```
ggplot(aug1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_point(data = aug1 %>%
    slice_max(abs(.resid), n = 5),
    col = "red", size = 2) +
  geom_text_repel(data = aug1 %>%
    slice_max(abs(.resid), n = 5),
    aes(label = subject), col = "red") +
  geom_abline(intercept = 0, slope = 0, lty = "dashed") +
  geom_smooth(method = "loess", formula = y ~ x, se = F) +
  labs(title = "Residuals vs. Fitted Values from mod_1",
       caption = "5 largest |residuals| highlighted in red.",
       x = "Fitted Value of (1/a1c)", y = "Residual") +
  theme(aspect.ratio = 1)
```

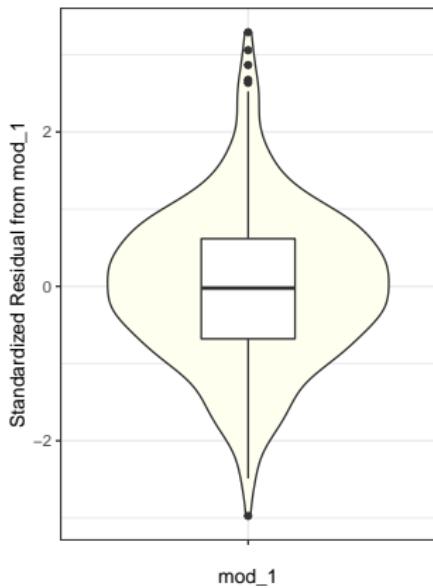
Normality of Standardized Residuals via ggplot2

Normality of Standardized Residuals from mod_1

Normal Q-Q plot



Box and Violin Plots



n = 335 residual values are plotted here.

Code for Normality Checks (1 of 2)

```
p1 <- ggplot(aug1, aes(sample = .std.resid)) +  
  geom_qq() +  
  geom_qq_line(col = "red") +  
  labs(title = "Normal Q-Q plot",  
       y = "Standardized Residual from mod_1",  
       x = "Standard Normal Quantiles") +  
  theme(aspect.ratio = 1)  
  
p2 <- ggplot(aug1, aes(y = .std.resid, x = "")) +  
  geom_violin(fill = "ivory") +  
  geom_boxplot(width = 0.3) +  
  labs(title = "Box and Violin Plots",  
       y = "Standardized Residual from mod_1",  
       x = "mod_1")
```

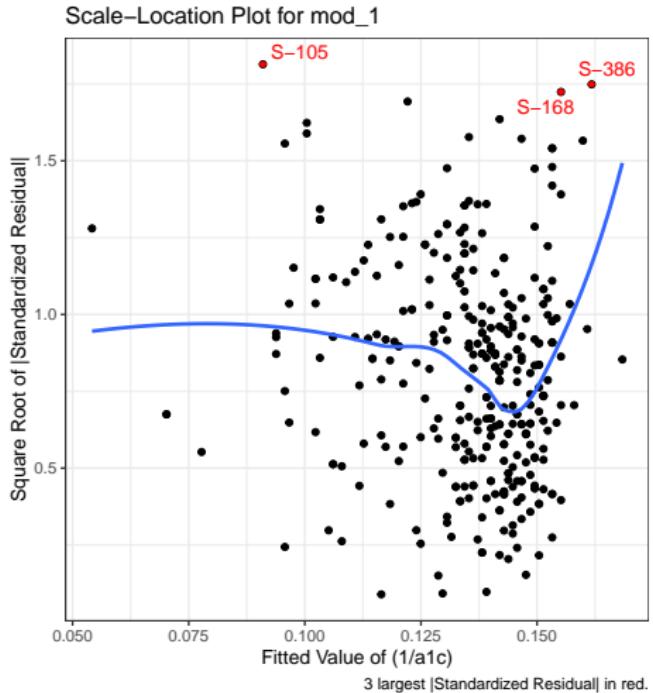
... continues on next slide

Code for Normality Checks (2 of 2)

```
p1 + p2 +
  plot_layout(widths = c(2, 1)) +
  plot_annotation(
    title = "Normality of Standardized Residuals from mod_1",
    caption = paste0("n = ",
                    nrow(aug1 %>% select(.std.resid)),
                    " residual values are plotted here."))

```

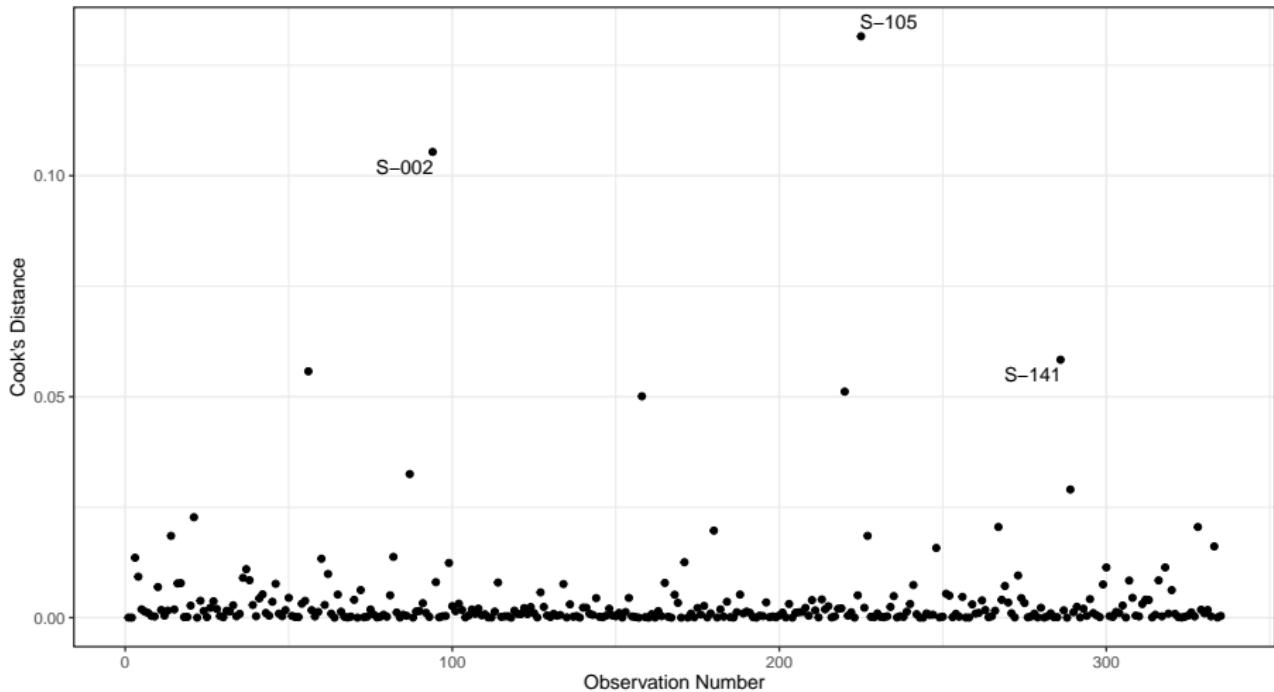
Scale-Location Plot via ggplot2



Code for Scale-Location Plot

```
ggplot(aug1, aes(x = .fitted, y = sqrt(abs(.std.resid)))) +  
  geom_point() +  
  geom_point(data = aug1 %>%  
              slice_max(sqrt(abs(.std.resid)), n = 3),  
              col = "red", size = 1) +  
  geom_text_repel(data = aug1 %>%  
                  slice_max(sqrt(abs(.std.resid)), n = 3),  
                  aes(label = subject), col = "red") +  
  geom_smooth(method = "loess", formula = y ~ x, se = F) +  
  labs(title = "Scale-Location Plot for mod_1",  
        caption = "3 largest |Standardized Residual| in red.",  
        x = "Fitted Value of (1/a1c)",  
        y = "Square Root of |Standardized Residual|") +  
  theme(aspect.ratio = 1)
```

Cook's Distance Index Plot via ggplot2

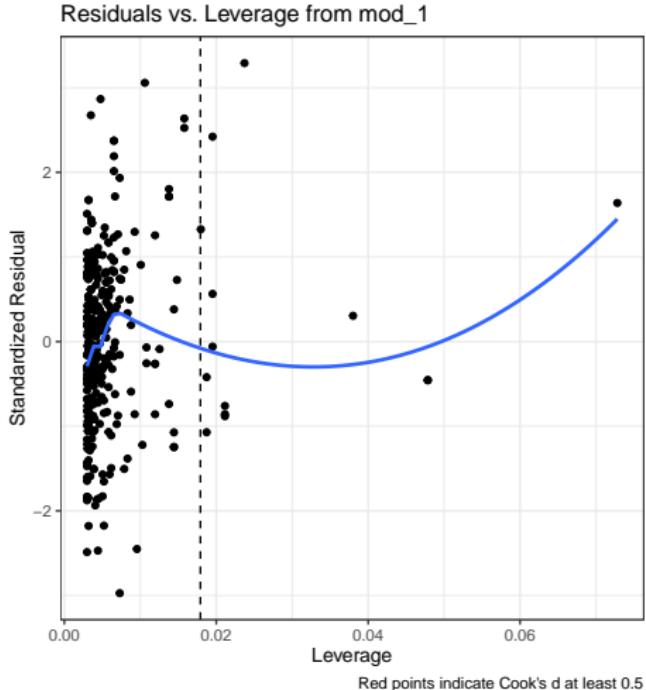


Code for Cook's Distance Index Plot

```
aug1_extra <- aug1 %>%
  mutate(obsnum = 1:nrow(aug1) %>% select(.cooksdi))

ggplot(aug1_extra, aes(x = obsnum, y = .cooksdi)) +
  geom_point() +
  geom_text_repel(data = aug1_extra %>%
    slice_max(.cooksdi, n = 3),
    aes(label = subject)) +
  labs(x = "Observation Number",
       y = "Cook's Distance")
```

Residuals vs. Leverage Plot via ggplot2



- Points with Cook's $d \geq 0.5$ would be highlighted and in red.
- Points right of the dashed line have high leverage, by one standard.

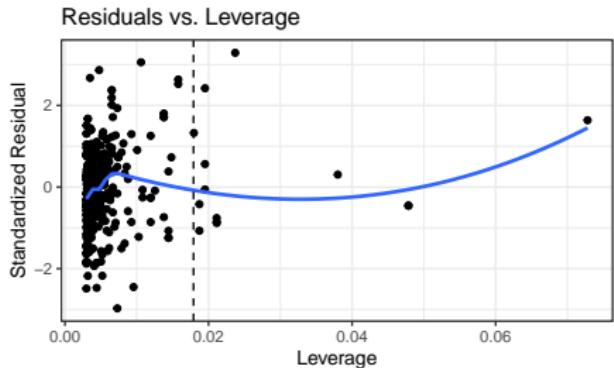
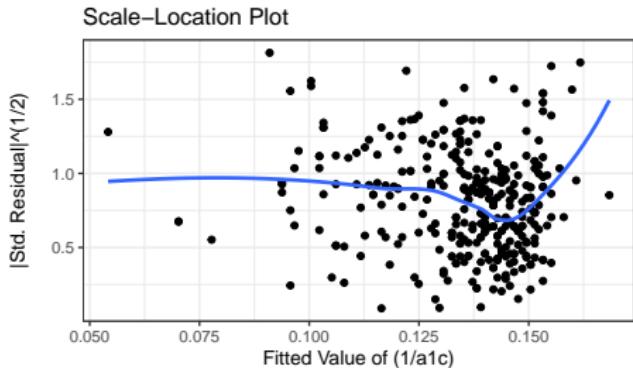
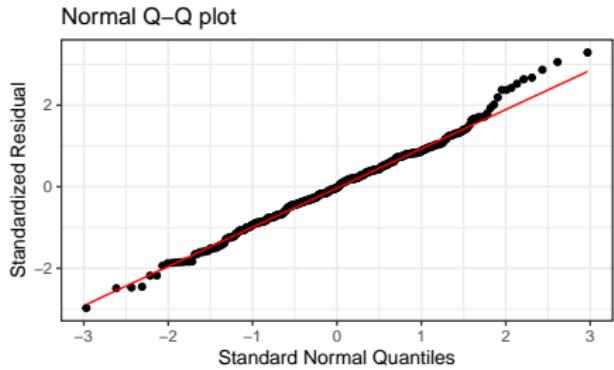
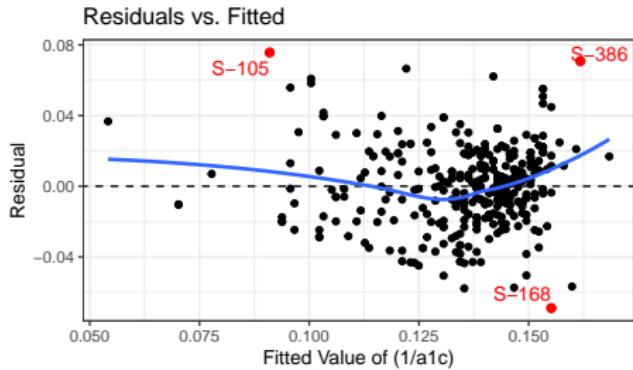
Code for Residuals vs. Leverage Plot

```
ggplot(aug1, aes(x = .hat, y = .std.resid)) +  
  geom_point() +  
  geom_point(data = aug1 %>% filter(.cooksdl >= 0.5),  
             col = "red", size = 2) +  
  geom_text_repel(data = aug1 %>% filter(.cooksdl >= 0.5),  
                  aes(label = subject), col = "red") +  
  geom_smooth(method = "loess", formula = y ~ x, se = F) +  
  geom_vline(aes(xintercept = 3*mean(.hat)), lty = "dashed") +  
  labs(title = "Residuals vs. Leverage from mod_1",  
        caption = "Red points indicate Cook's d at least 0.5",  
        x = "Leverage", y = "Standardized Residual") +  
  theme(aspect.ratio = 1)
```

- Points with more than 3 times the average leverage are identified as highly leveraged by some people, hence my dashed vertical line.

Main 4 Residual Plots for mod_1 (via ggplot2)

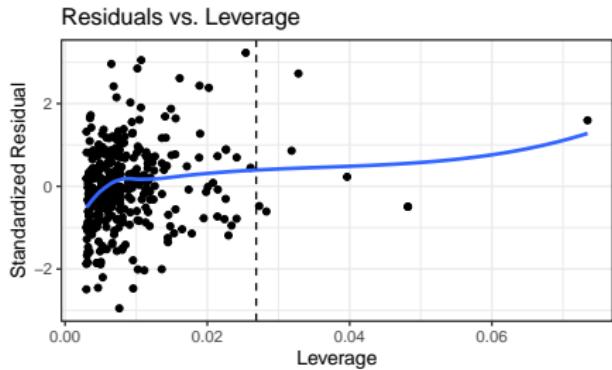
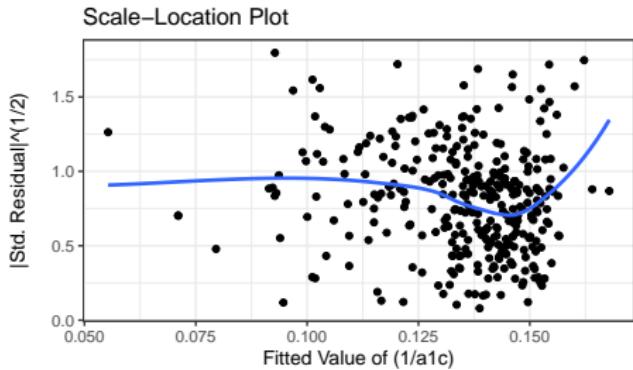
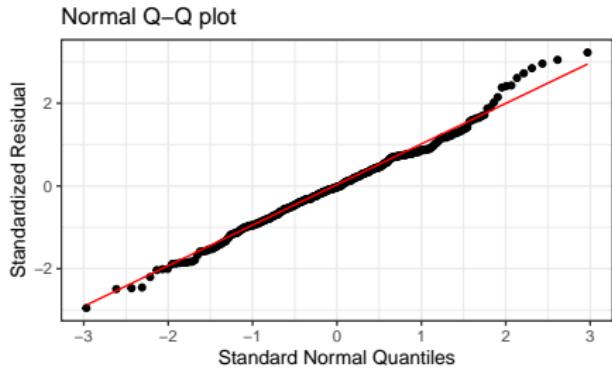
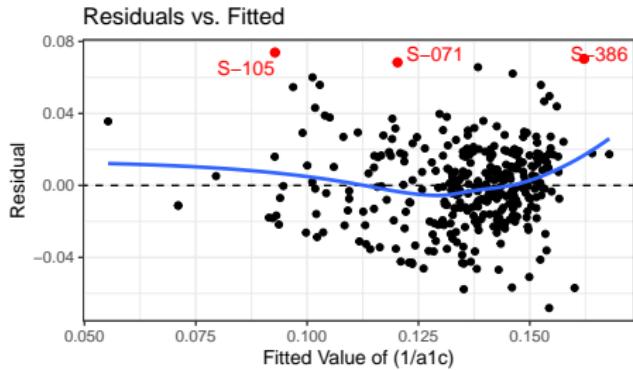
Assessing Residuals for mod_1



If applicable, Cook's d >= 0.5 shown in red in bottom right plot.

Main 4 Residual Plots for mod_2 (via ggplot2)

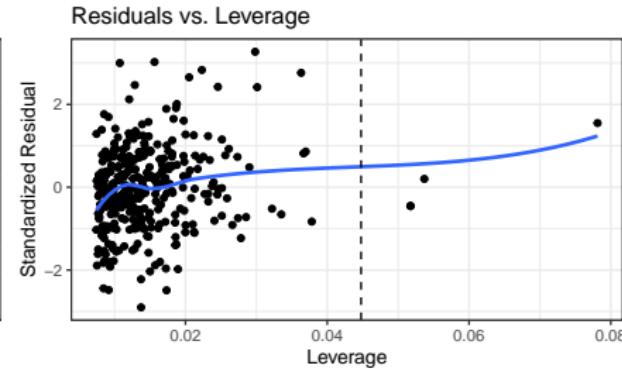
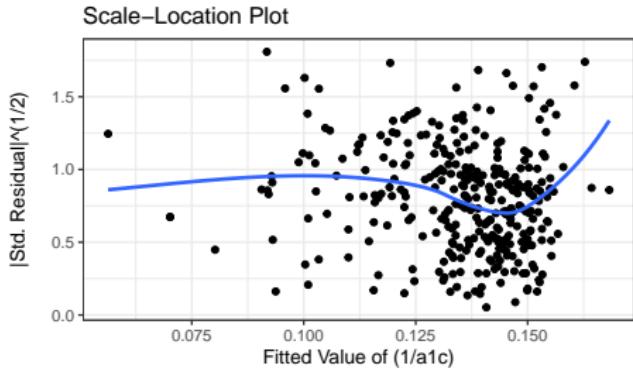
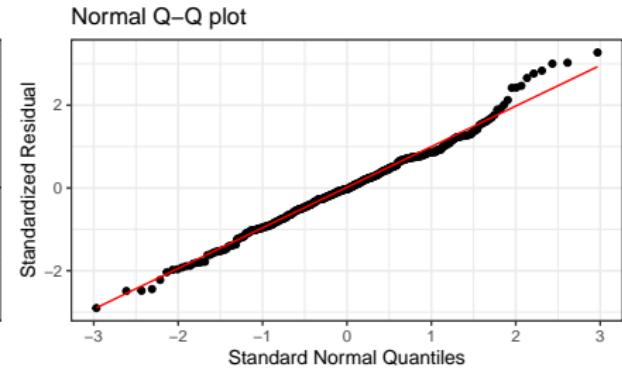
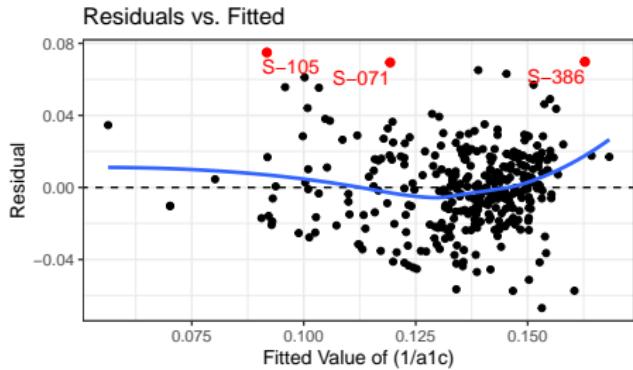
Assessing Residuals for mod_2



If applicable, Cook's d >= 0.5 shown in red in bottom right plot.

Main 4 Residual Plots for mod_3 (via ggplot2)

Assessing Residuals for mod_3



If applicable, Cook's $d \geq 0.5$ shown in red in bottom right plot.

Conclusions so far? (repeating what we said earlier)

- ① In-sample model predictions are about equally accurate for each of the three models. Model 2 looks better in terms of adjusted R^2 and AIC, but model 1 looks better on BIC. There's really not much to choose from there.
- ② Residual plots look similarly reasonable for linearity, Normality and constant variance in all three models.

**When you have candidates, assess them based
on the accuracy of the predictions they make
for the data held out (and thus not used in
building the models.)**

Calculate prediction errors for mod_1 in test sample

The augment function in the broom package will create predictions within our new sample, but we want to back-transform these predictions so that they are on the original scale (a1c, rather than our transformed regression outcome 1/a1c). Since the way to back out of the inverse transformation is to take the inverse again, we will take the inverse of the fitted values provided by augment and then calculate residuals on the original scale, as follows...

```
test_m1 <- augment(mod_1, newdata = dm1_cc_test) %>%
  mutate(name = "mod_1", fit_a1c = 1 / .fitted,
        res_a1c = a1c - fit_a1c)
```

What does test_m1 now include?

```
test_m1 %>%
  select(subject, a1c, fit_a1c, res_a1c, a1c_old,
         age, income) %>%
  head() %>%
  knitr::kable(digits = c(0, 1, 2, 2, 1, 0, 0))
```

subject	a1c	fit_a1c	res_a1c	a1c_old	age	income
S-004	6.5	6.52	-0.02	5.8	53	Below_30K
S-005	6.7	6.73	-0.03	6.3	64	Between_30-50K
S-012	12.2	9.87	2.33	11.3	52	Below_30K
S-014	8.4	7.88	0.52	8.6	44	Between_30-50K
S-015	5.7	6.48	-0.78	5.7	52	Between_30-50K
S-021	11.4	7.60	3.80	8.1	51	Higher_than_50K

Gather test-sample prediction errors for models 2, 3

```
test_m2 <- augment(mod_2, newdata = dm1_cc_test) %>%
  mutate(name = "mod_2", fit_a1c = 1 / .fitted,
        res_a1c = a1c - fit_a1c)

test_m3 <- augment(mod_3, newdata = dm1_cc_test) %>%
  mutate(name = "mod_3", fit_a1c = 1 / .fitted,
        res_a1c = a1c - fit_a1c)
```

Combine test sample results from the three models

```
test_comp <- bind_rows(test_m1, test_m2, test_m3) %>%  
  arrange(subject, name)  
  
test_comp %>% select(name, subject, a1c, fit_a1c, res_a1c,  
                      a1c_old, age, income) %>%  
  slice(1:3, 7:9) %>%  
  knitr::kable(digits = c(0, 0, 1, 2, 2, 1, 0, 0))
```

name	subject	a1c	fit_a1c	res_a1c	a1c_old	age	income
mod_1	S-004	6.5	6.52	-0.02	5.8	53	Below_30K
mod_2	S-004	6.5	6.57	-0.07	5.8	53	Below_30K
mod_3	S-004	6.5	6.62	-0.12	5.8	53	Below_30K
mod_1	S-012	12.2	9.87	2.33	11.3	52	Below_30K
mod_2	S-012	12.2	9.90	2.30	11.3	52	Below_30K
mod_3	S-012	12.2	9.99	2.21	11.3	52	Below_30K

What do we do to compare the test-sample errors?

Given this tibble, including predictions and residuals from the three models on our test data, we can now:

- ① Visualize the prediction errors from each model.
- ② Summarize those errors across each model.
- ③ Identify the “worst fitting” subject for each model in the test sample.

Visualize the prediction errors

```
ggplot(test_comp, aes(x = res_a1c, fill = name)) +  
  geom_histogram(bins = 20, col = "white") +  
  facet_grid (name ~ .) + guides(fill = "none")
```

or maybe

```
ggplot(test_comp, aes(x = name, y = res_a1c, fill = name)) +  
  geom_violin(alpha = 0.3) +  
  geom_boxplot(width = 0.3, outlier.shape = NA) +  
  geom_jitter(height = 0, width = 0.1) +  
  guides(fill = "none")
```

Test-Sample Prediction Errors

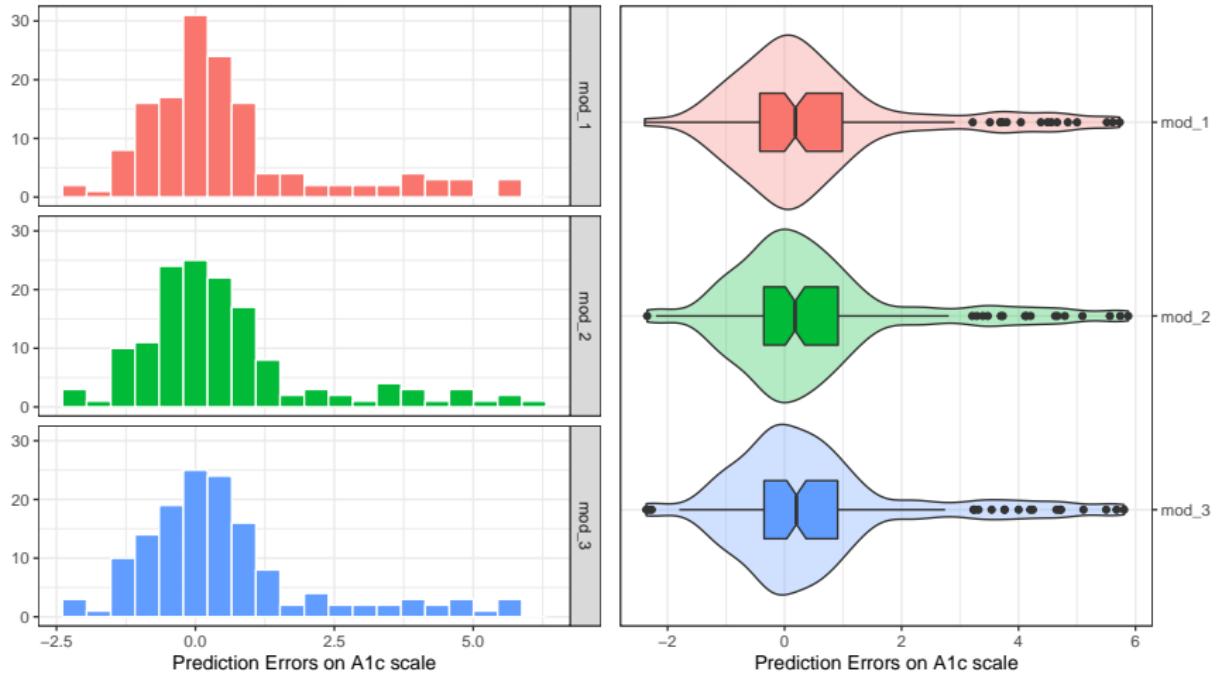


Table Comparing Model Prediction Errors

Calculate the mean absolute prediction error (MAPE), the square root of the mean squared prediction error (RMSPE) and the maximum absolute error across the predictions made by each model. Let's add the median absolute prediction error, too.

```
test_comp %>%
  group_by(name) %>%
  summarize(n = n(),
            MAPE = mean(abs(res_a1c)),
            RMSPE = sqrt(mean(res_a1c^2)),
            max_error = max(abs(res_a1c)),
            median_APE = median(abs(res_a1c))) %>%
  kable(digits = c(0, 0, 4, 3, 2, 3))
```

Table moved to the next slide.

Conclusions from Table of Errors

name	n	mean_APE	RMSPE	max_error	median_APE
mod_1	144	1.1153	1.731	5.73	0.651
mod_2	144	1.1201	1.723	5.88	0.658
mod_3	144	1.1242	1.722	5.81	0.699

- Model `mod_1` has the smallest MAPE (mean APE) and maximum error and median absolute prediction error.
- Model `mod_3` has the smallest root mean squared prediction error (RMSPE).

Identify the largest errors

Identify the subject(s) where that maximum prediction error was made by each model, and the observed and model-fitted values of a1c in each case.

```
temp1 <- test_m1 %>%
  filter(abs(res_a1c) == max(abs(res_a1c)))
```

```
temp2 <- test_m2 %>%
  filter(abs(res_a1c) == max(abs(res_a1c)))
```

```
temp3 <- test_m3 %>%
  filter(abs(res_a1c) == max(abs(res_a1c)))
```

Identify the largest errors (Results)

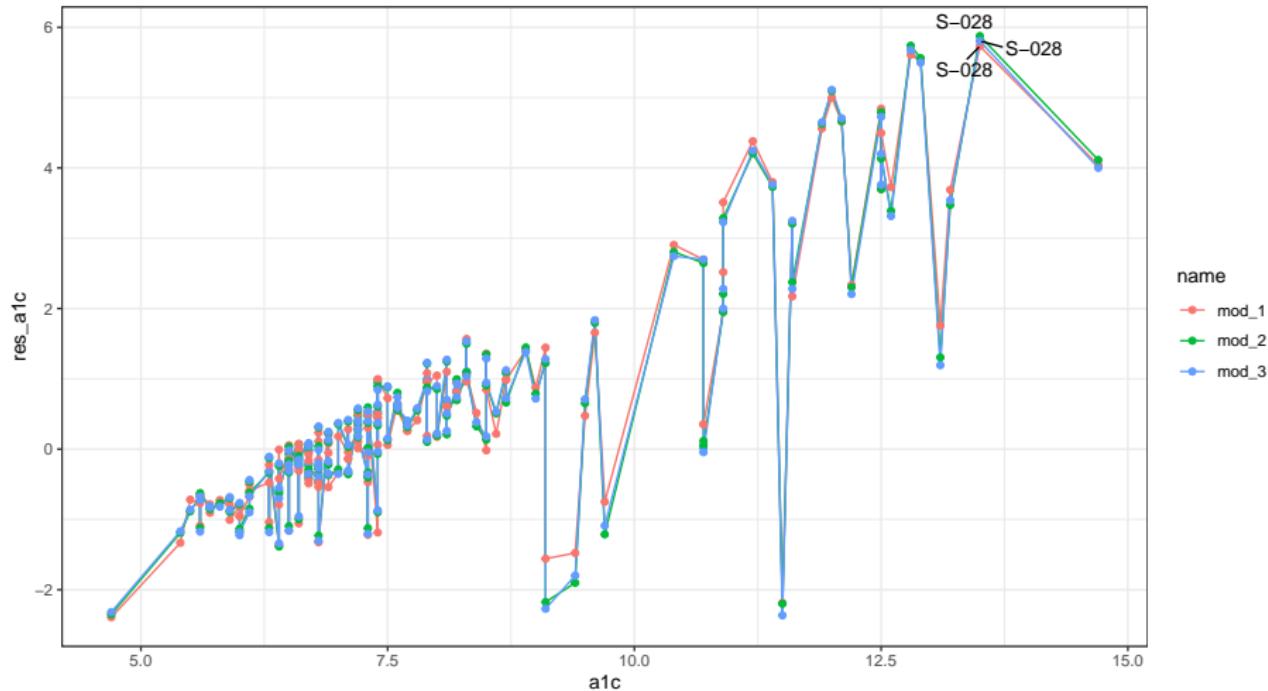
Identify the subject(s) where that maximum prediction error was made by each model, and the observed and model-fitted values of a1c in each case.

```
bind_rows(temp1, temp2, temp3) %>%
  select(subject, name, a1c, fit_a1c, res_a1c)
```

```
# A tibble: 3 x 5
  subject name     a1c fit_a1c res_a1c
  <chr>   <chr> <dbl>    <dbl>    <dbl>
1 S-028   mod_1   13.5     7.77    5.73
2 S-028   mod_2   13.5     7.62    5.88
3 S-028   mod_3   13.5     7.69    5.81
```

Line Plot of the Errors?

Compare the errors that are made at each level of observed A1c?



Code for the Line Plot of the Prediction Errors

```
ggplot(test_comp, aes(x = a1c, y = res_a1c,  
                      group = name)) +  
  geom_line(aes(col = name)) +  
  geom_point(aes(col = name)) +  
  geom_text_repel(data = test_comp %>%  
                  filter(subject == "S-028"),  
                  aes(label = subject))
```

What if we ignored S-028 for a moment?

All three miss this subject substantially, but without S-028, we have:

```
test_comp %>% filter(subject != "S-028") %>%
  group_by(name) %>%
  summarize(n = n(),
            MAPE = mean(abs(res_a1c)),
            RMSPE = sqrt(mean(res_a1c^2)),
            max_error = max(abs(res_a1c))) %>%
  kable(digits = c(0, 0, 3, 4, 2))
```

name	n	MAPE	RMSPE	max_error
mod_1	143	1.083	1.6694	5.61
mod_2	143	1.087	1.6577	5.74
mod_3	143	1.092	1.6583	5.68

Excluding subject S-028, mod_1 wins MAPE and maxE, but mod_2 wins RMSPE

Conclusions based on complete case analysis?

- ① In-sample model predictions are about equally accurate for each of the three models. Model 2 looks better in terms of adjusted R^2 and AIC, but model 1 looks better on BIC. There's really not much to choose from there.
- ② Residual plots look similarly reasonable for linearity, Normality and constant variance in all three models.
- ③ In our holdout sample, model `mod_1` has the smallest MAPE (mean APE) and RMSPE and maximum error, while model `mod_2` has the smallest median absolute prediction error, although again all three models are pretty comparable. Excluding a bad miss on one subject in the test sample yields similar comparisons. Again, the three models do about equally well on these measures.

So, what should our “most useful” model be?

OK. Let's do all of that again, using the (singly) imputed data.

Partition imputed data from dm1_imp

This time, we'll build an 80% development, 20% holdout partition of the dm1_imp data, and we'll also change our random seed, just for fun.

```
set.seed(20212021)

dm1_imp_train <- dm1_imp %>%
  slice_sample(prop = 0.8, replace = FALSE)

dm1_imp_test <-
  anti_join(dm1_imp, dm1_imp_train, by = "subject")

dim(dm1_imp_train); dim(dm1_imp_test)
```

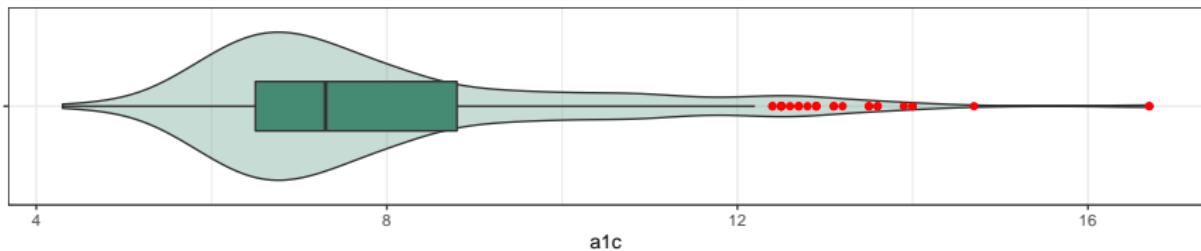
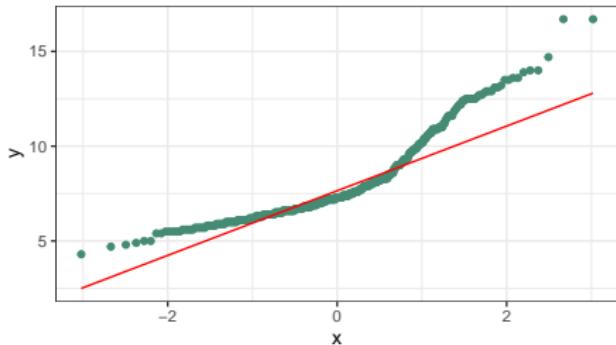
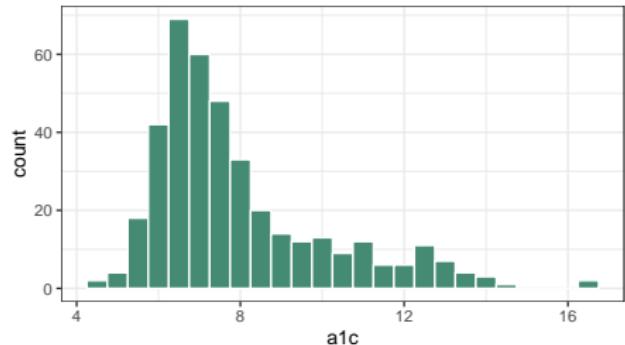
```
[1] 396    5
```

```
[1] 100    5
```

Distribution of a1c in training sample

Hemoglobin A1c values (%)

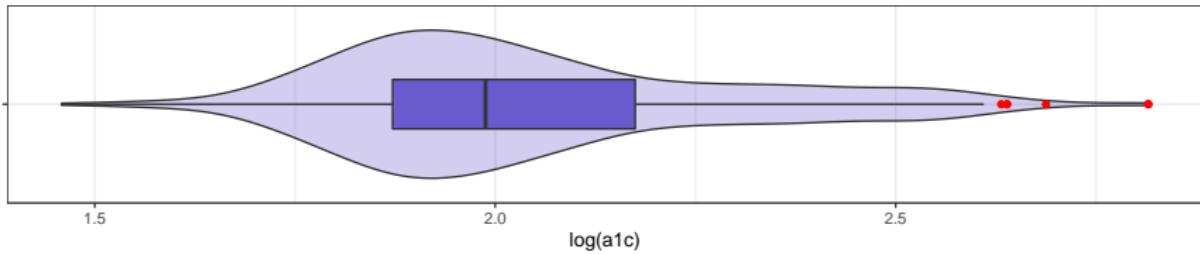
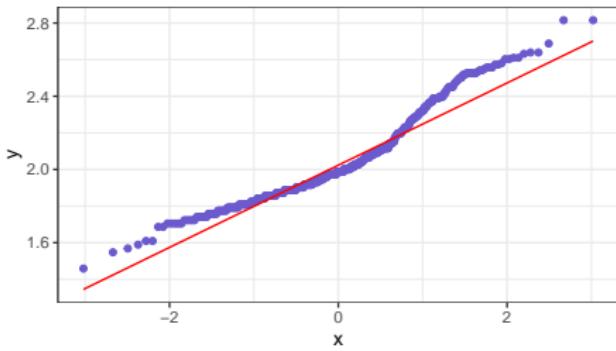
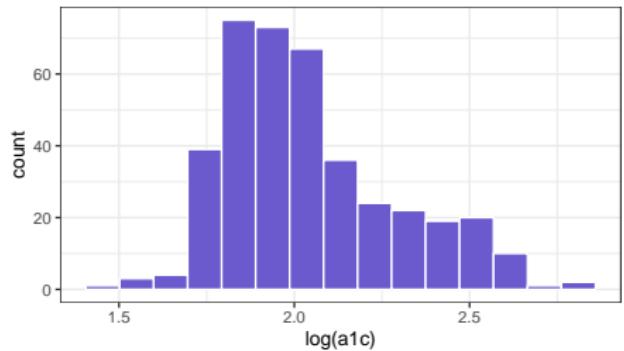
Model Development Sample after imputation: 396 adults with diabetes



Consider a log transformation?

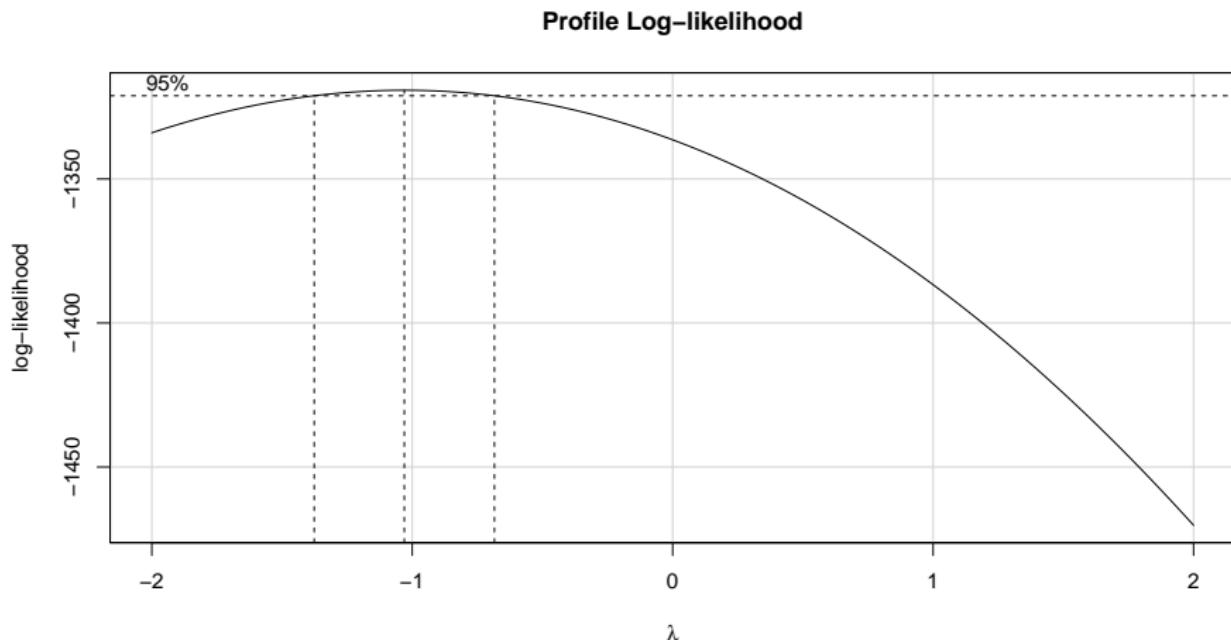
Natural Logarithm of Hemoglobin A1c

Model Development Sample: 396 adults with diabetes



What does Box-Cox suggest?

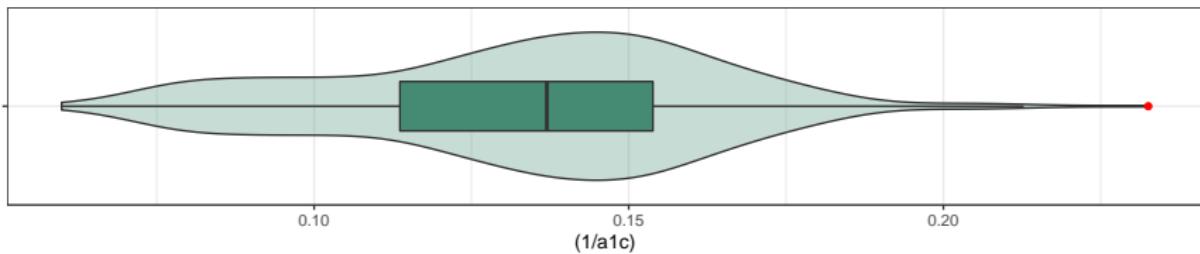
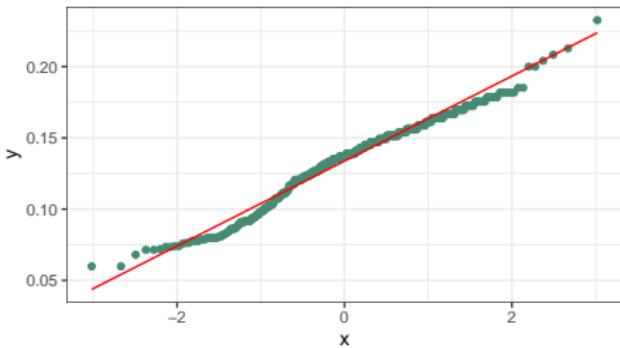
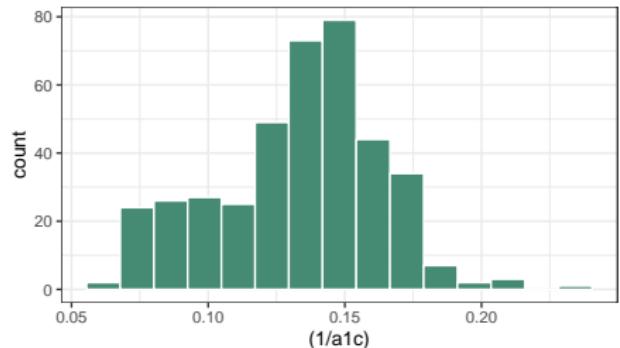
```
imod_0 <- lm(a1c ~ a1c_old + age + income,  
              data = dm1_imp_train)  
boxCox(imod_0)
```



Inverse of A1c again?

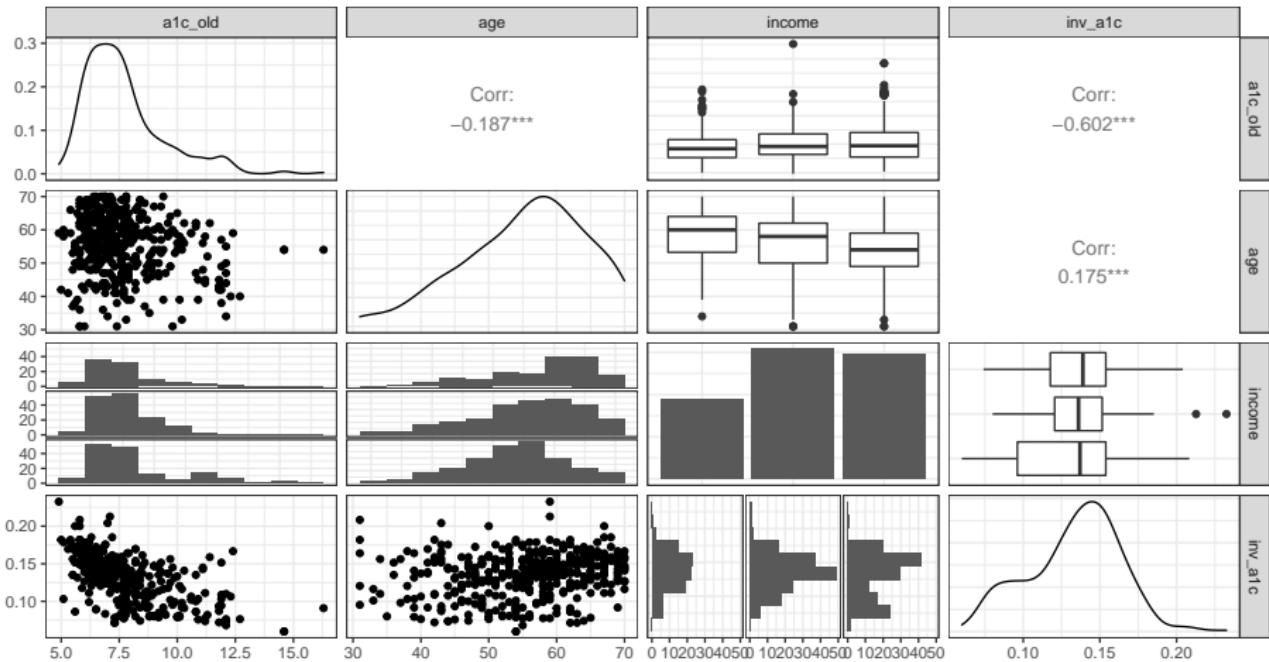
Inverse of Hemoglobin A1c

Model Development Sample after Imputation: 396 adults with diabetes



Scatterplot Matrix

Scatterplots: Model Development Imputed Sample



Fitting the Same Three Models

- Remember we're using the model development sample here.

```
imod_1 <- lm((1/a1c) ~ a1c_old, data = dm1_imp_train)
```

```
imod_2 <- lm((1/a1c) ~ a1c_old + age, data = dm1_imp_train)
```

```
imod_3 <- lm((1/a1c) ~ a1c_old + age + income,  
               data = dm1_imp_train)
```

**Assess the quality of fit for candidate models
within the development sample.**

Tidied coefficients (imod_1)

```
tidy_im1 <- tidy(imod_1, conf.int = TRUE, conf.level = 0.95)

tidy_im1 %>%
  select(term, estimate, std.error, p.value,
         conf.low, conf.high) %>%
knitr::kable(digits = 4)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	0.2126	0.0054	0	0.2019	0.2233
a1c_old	-0.0103	0.0007	0	-0.0117	-0.0090

The Regression Equation (imod_1)

Again, we'll use the equationatic package.

```
extract_eq(imod_1, use_coefs = TRUE, coef_digits = 4,  
          italic_vars = TRUE, wrap = TRUE, terms_per_line = 3)
```

$$\widehat{(\text{1}/\text{a1c})} = 0.2126 - 0.0103(\text{a1c_old}) \quad (4)$$

Summary of Fit Quality (imod_1)

```
glance(imod_1) %>%
  mutate(name = "imod_1") %>%
  select(name, r.squared, adj.r.squared,
         sigma, AIC, BIC) %>%
knitr::kable(digits = c(0, 3, 3, 3, 0, 0))
```

name	r.squared	adj.r.squared	sigma	AIC	BIC
imod_1	0.362	0.361	0.024	-1833	-1821

Tidied coefficients (imod_2)

```
tidy_im2 <- tidy(imod_2, conf.int = TRUE, conf.level = 0.95)

tidy_im2 %>%
  select(term, estimate, std.error, p.value,
        conf.low, conf.high) %>%
knitr::kable(digits = 4)
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	0.1991	0.0101	0.0000	0.1792	0.2189
a1c_old	-0.0101	0.0007	0.0000	-0.0115	-0.0087
age	0.0002	0.0001	0.1131	-0.0001	0.0005

The Regression Equation (imod_2)

Again, we'll use the `equatiomatic` package, and `results = 'asis'`.

```
extract_eq(imod_2, use_coefs = TRUE, coef_digits = 4,  
          italic_vars = TRUE)
```

$$\widehat{(\text{1}/\text{a1c})} = 0.1991 - 0.0101(\text{a1c_old}) + 2e - 04(\text{age}) \quad (5)$$

Summary of Fit Quality (imod_2)

```
glance(imod_2) %>%
  mutate(name = "imod_2") %>%
  select(name, r.squared, adj.r.squared,
         sigma, AIC, BIC) %>%
knitr::kable(digits = c(0, 3, 3, 3, 0, 0))
```

name	r.squared	adj.r.squared	sigma	AIC	BIC
imod_2	0.367	0.363	0.024	-1833	-1817

Tidied coefficients (imod_3)

```
tidy_im3 <- tidy(imod_3, conf.int = TRUE, conf.level = 0.95)

tidy_im3 %>%
  select(term, estimate, se = std.error,
         low = conf.low, high = conf.high, p = p.value) %>%
  knitr::kable(digits = c(4,4,4,4,3))
```

term	estimate	se	low	high	p
(Intercept)	0.2002	0.0106	0.1795	0.221	0.0000
a1c_old	-0.0101	0.0007	-0.0115	-0.009	0.0000
age	0.0002	0.0001	-0.0001	0.000	0.1530
incomeBetween_30-50K	0.0010	0.0031	-0.0052	0.007	0.7590
incomeBelow_30K	-0.0023	0.0032	-0.0086	0.004	0.4764

The Regression Equation (imod_3)

Again, we'll use the `equatiomatic` package.

```
extract_eq(imod_3, use_coefs = TRUE, coef_digits = 4,  
          italic_vars = TRUE, wrap = TRUE, terms_per_line = 2)
```

$$\widehat{(\text{1}/\text{a1c})} = 0.2002 - 0.0101(\text{a1c_old}) + \\ 2e - 04(\text{age}) + 0.001(\text{income}_{\text{Between_}30-50K}) - \quad (6) \\ 0.0023(\text{income}_{\text{Below_}30K})$$

Summary of Fit Quality (imod_3)

```
glance(imod_3) %>%
  mutate(name = "imod_3") %>%
  select(name, r.squared, adj.r.squared,
         sigma, AIC, BIC) %>%
knitr::kable(digits = c(0, 3, 3, 3, 0, 0))
```

name	r.squared	adj.r.squared	sigma	AIC	BIC
imod_3	0.369	0.362	0.024	-1831	-1807

I checked stepwise regression again

- Even though variable selection **never** works, it is seductive.

What if we do forward selection in this situation?

```
min.model <- lm(a1c ~ 1, data = dm1_imp_train)
fwd.model <- step(min.model, direction = "forward",
                    scope = ~ a1c_old + age + income)
```

Start: AIC=606.99

a1c ~ 1

	Df	Sum of Sq	RSS	AIC
+ a1c_old	1	694.77	1129.9	419.20
+ age	1	64.26	1760.4	594.79
+ income	2	48.20	1776.5	600.39
<none>			1824.7	606.99

Step: AIC=419.2

Stepwise Regression Results

We wind up back at the model with all three predictors in this case (mod_3).

```
fwd.model$coefficients
```

(Intercept)	a1c_old
3.05112418	0.74107516
incomeBetween_30-50K	incomeBelow_30K
-0.16321016	0.34246995
age	
-0.01520655	

- As we'll discuss in 432, there is an immense amount of evidence that variable selection causes severe problems in estimation and inference.

Which Model Looks Best In-Sample?

For each of these summaries, which model looks best in the training sample?

model	vars	r2	adj_r2	sigma	AIC	BIC
imod_1	a1c_old	0.362	0.361	0.02380	-1832.9	-1821
imod_2	+ age	0.367	0.363	0.02375	-1833.4	-1817
imod_3	+ income	0.369	0.362	0.02377	-1830.9	-1807

- imod_3 (as it must, here) has the best R-square.
- imod_2 wins on adjusted R-square and σ and AIC
- imod_1 has the best BIC

Using augment to add fits, residuals, etc.

```
augi1 <- augment(imod_1, data = dm1_imp_train) %>%
  mutate(inv_a1c = 1/a1c) # add in our model's outcome

augi2 <- augment(imod_2, data = dm1_imp_train) %>%
  mutate(inv_a1c = 1/a1c) # add in our model's outcome

augi3 <- augment(imod_3, data = dm1_imp_train) %>%
  mutate(inv_a1c = 1/a1c) # add in our model's outcome
```

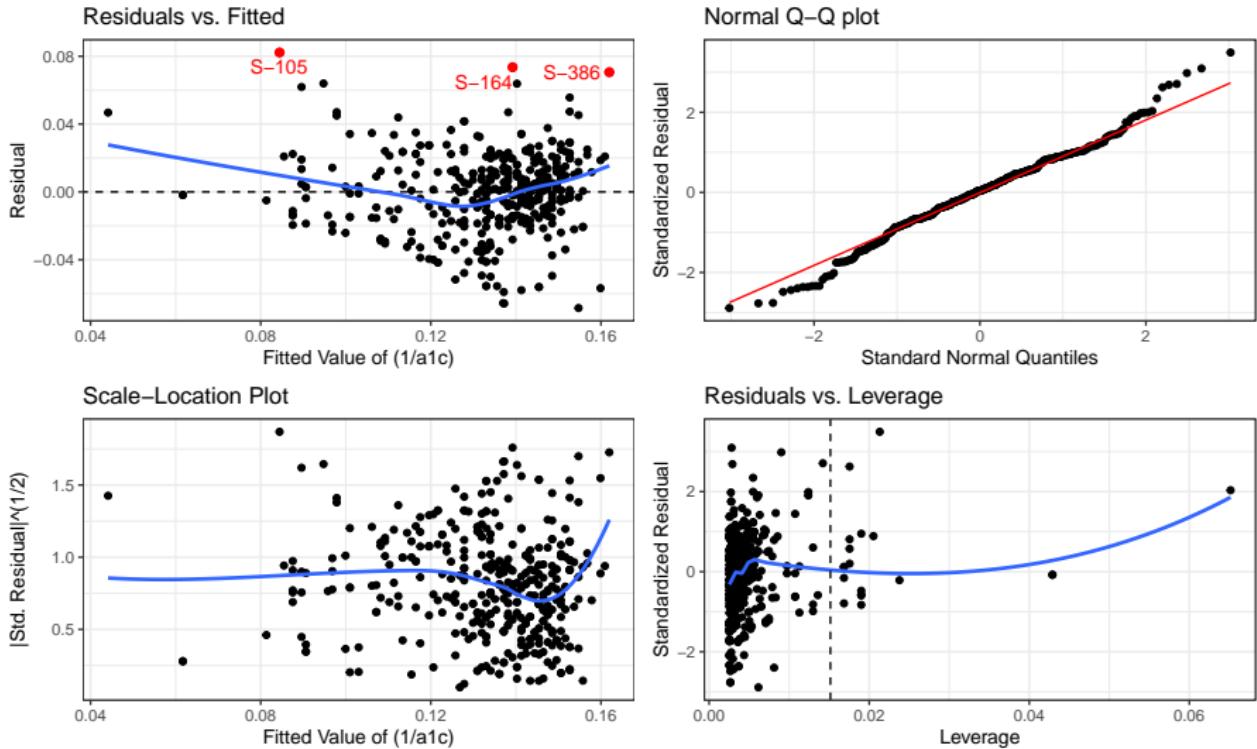
Checking Regression Assumptions

Four key assumptions we need to think about:

- ① Linearity
- ② Constant Variance (Homoscedasticity)
- ③ Normality
- ④ Independence

Main 4 Residual Plots for imod_1 (via ggplot2)

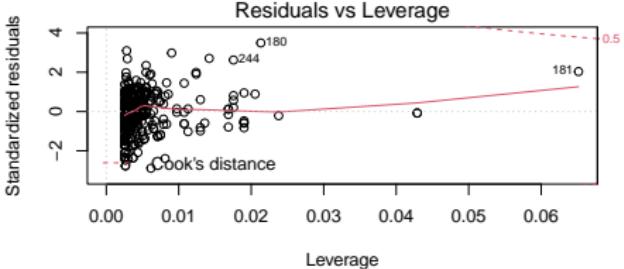
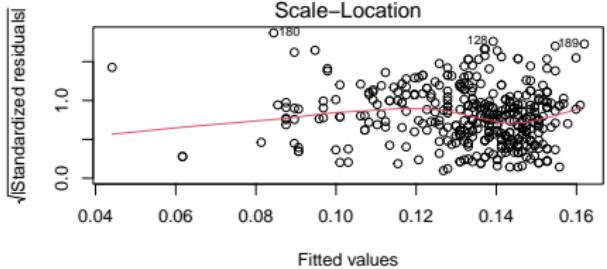
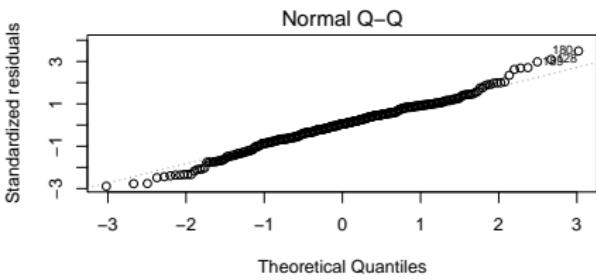
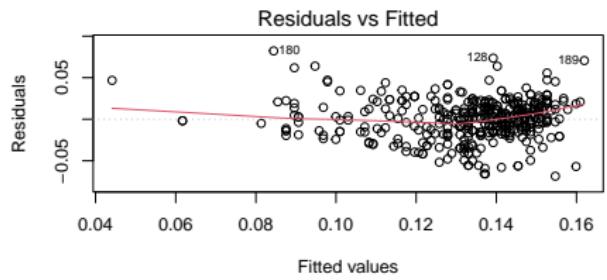
Assessing Residuals for imod_1



If applicable, Cook's d ≥ 0.5 shown in red in bottom right plot.

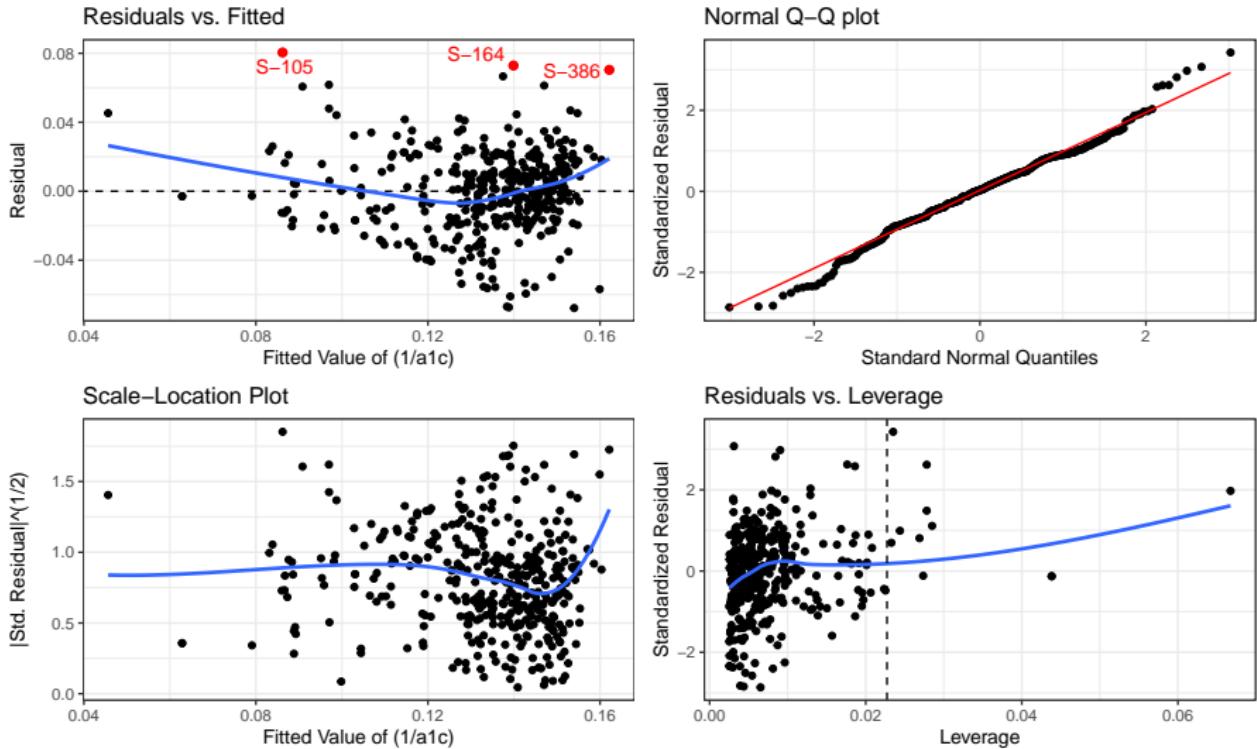
Base R Residual Plots for Model imod_1

```
par(mfrow = c(2,2)); plot(imod_1); par(mfrow = c(1,1))
```



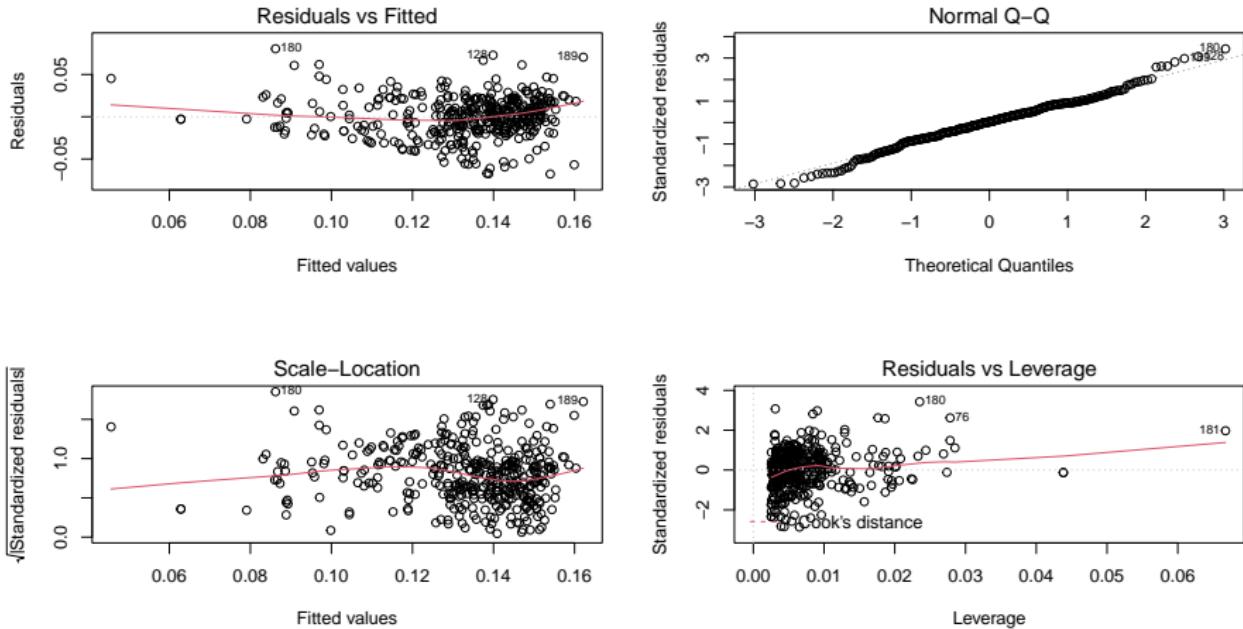
Main 4 Residual Plots for imod_2 (via ggplot2)

Assessing Residuals for imod_2



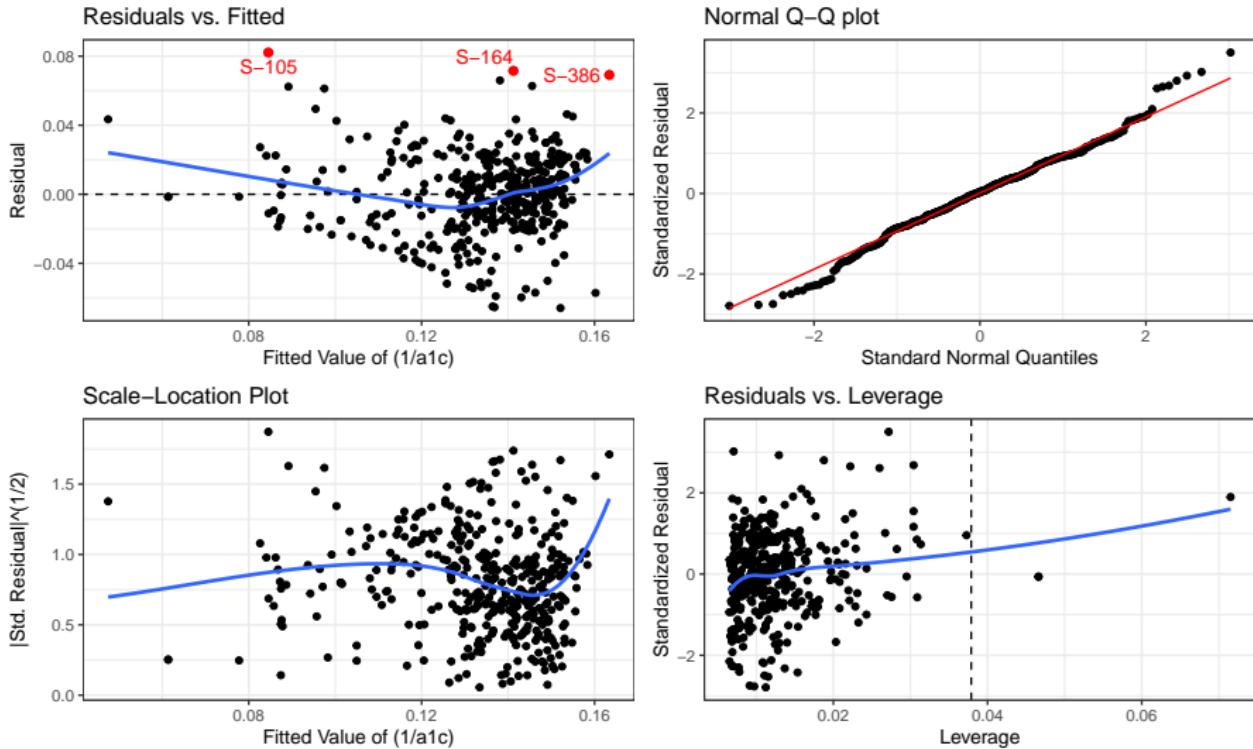
If applicable, Cook's $d \geq 0.5$ shown in red in bottom right plot.

Base R Residual Plots for Model imod_2



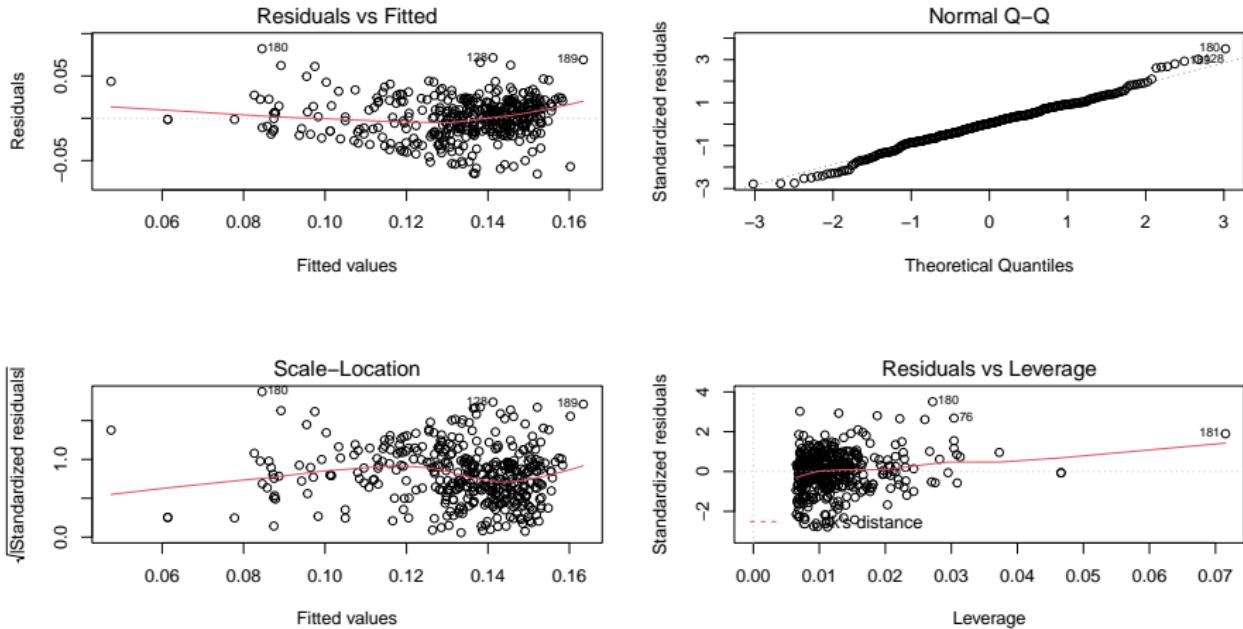
Main 4 Residual Plots for imod_3 (via ggplot2)

Assessing Residuals for imod_3



If applicable, Cook's d ≥ 0.5 shown in red in bottom right plot.

Base R Residual Plots for Model imod_3



Is collinearity a serious issue here?

```
car::vif(imod_3)
```

	GVIF	Df	GVIF ^{(1/(2*Df))}
a1c_old	1.041113	1	1.020350
age	1.069426	1	1.034131
income	1.042549	2	1.010472

None of these values exceed 5, so it doesn't seem like there's any problem.

```
car::vif(imod_2)
```

a1c_old	age
1.03632	1.03632

Conclusions so far (in-sample)?

- ① In-sample model predictions are not wildly different in terms of accuracy across the three models.
 - Model imod_3 has the best R^2 , while
 - Model imod_2 wins on adjusted R^2 , σ and AIC, and
 - Model imod_1 has the best BIC.
- ② Residual plots look similarly reasonable for linearity, Normality and constant variance in all three models after imputation.

Calculate prediction errors in test samples

```
test_im1 <- augment(imod_1, newdata = dm1_imp_test) %>%
  mutate(name = "imod_1", fit_a1c = 1 / .fitted,
        res_a1c = a1c - fit_a1c)

test_im2 <- augment(imod_2, newdata = dm1_imp_test) %>%
  mutate(name = "imod_2", fit_a1c = 1 / .fitted,
        res_a1c = a1c - fit_a1c)

test_im3 <- augment(imod_3, newdata = dm1_imp_test) %>%
  mutate(name = "imod_3", fit_a1c = 1 / .fitted,
        res_a1c = a1c - fit_a1c)

test_icomp <- bind_rows(test_im1, test_im2, test_im3) %>%
  arrange(subject, name)
```

Visualize Test-Sample Prediction Errors

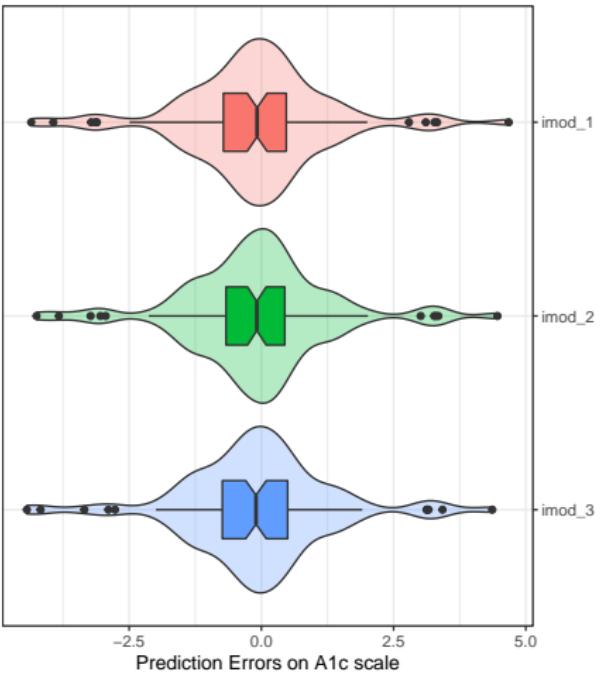
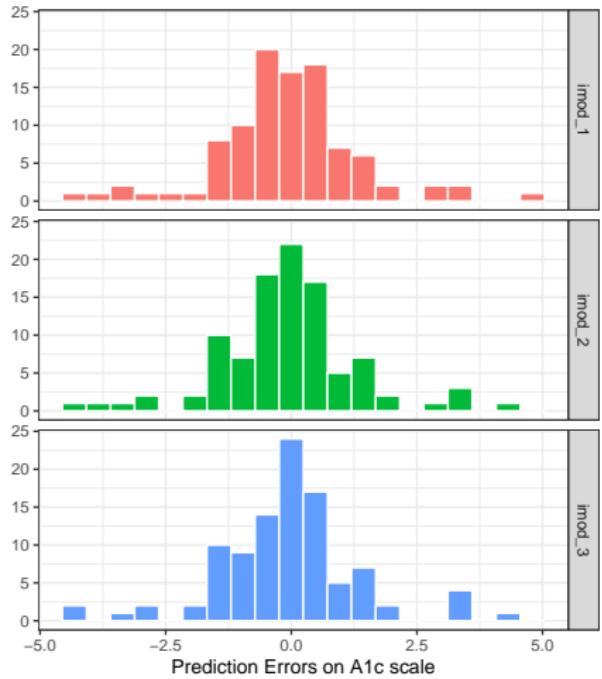


Table Comparing Model Prediction Errors

- Model imod_2 has the best mean APE (MAPE) and RMSPE, while imod_3 has the smallest maximum predictive error.

```
test_icomp %>%
  group_by(name) %>%
  summarize(n = n(),
            MAPE = mean(abs(res_a1c)),
            RMSPE = sqrt(mean(res_a1c^2)),
            max_error = max(abs(res_a1c))) %>%
  kable(digits = c(0, 0, 3, 3, 2))
```

name	n	MAPE	RMSPE	max_error
imod_1	100	0.971	1.396	4.68
imod_2	100	0.958	1.377	4.46
imod_3	100	0.964	1.384	4.43

Identify the largest errors (Results)

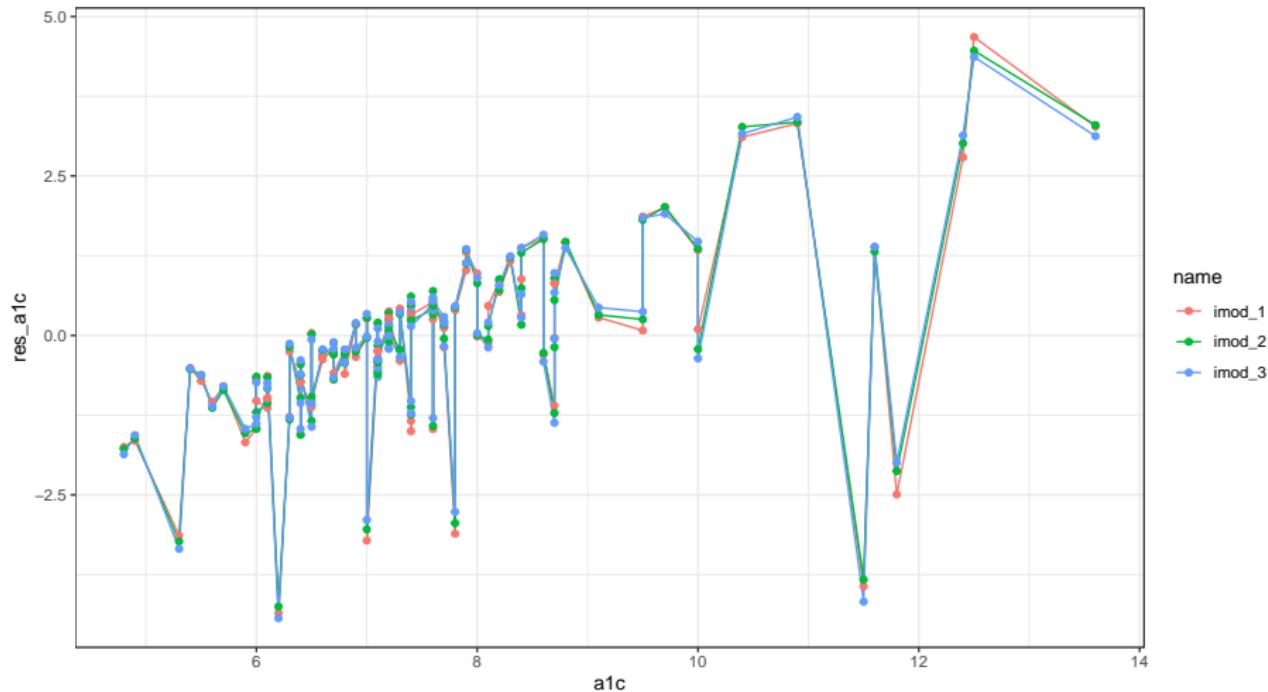
Identify the subject(s) where that maximum prediction error was made by each model, and the observed and model-fitted values of a1c in each case.

```
bind_rows(tempi1, tempi2, tempi3) %>%  
  select(subject, name, a1c, fit_a1c, res_a1c)
```

```
# A tibble: 3 x 5  
  subject name      a1c fit_a1c res_a1c  
  <chr>   <chr>    <dbl>    <dbl>    <dbl>  
1 S-471   imod_1   12.5     7.82    4.68  
2 S-471   imod_2   12.5     8.04    4.46  
3 S-341   imod_3   6.2      10.6    -4.43
```

Line Plot of the Errors?

Compare the errors that are made at each level of observed A1c?



Key Summaries

With complete cases,

- in-sample: all three models look OK on assumptions in residual plots, model 2 looks like it fits a little better by Adjusted R^2 and AIC, model 1 looks slightly better by BIC.
- out-of-sample: distributions of errors are similar. Model 1 has smallest MAPE, RMPSE and maximum error, while Model 2 has the smallest median error, but all three models are pretty similar.

With imputation,

- in-sample: nothing disastrous in residual plots, model 3 has the best R^2 , Model 2 wins on adjusted R^2 , σ , and AIC, and Model 1 has the best BIC.
- out-of-sample: Model 2 has the smallest MAPE, RMSE, but model 3 has the smallest maximum predictive error.

So what can we conclude? Does this particular imputation strategy have a big impact?

Again, this is our 431 Strategy

Which model is “most useful” in a prediction context?

- ① Split the data into a model development (training) sample of about 70-80% of the observations, and a model test (holdout) sample, containing the remaining observations.
- ② Develop candidate models using the development sample.
- ③ Assess the quality of fit for candidate models within the development sample.
- ④ Check adherence to regression assumptions in the development sample.
- ⑤ When you have candidates, assess them based on the accuracy of the predictions they make for the data held out (and thus not used in building the models.)
- ⑥ Select a “final” model for use based on the evidence in steps 3, 4 and especially 5.

431 Class 21

thomaselove.github.io/431

2021-11-09

Today's R Setup

```
library(janitor); library(knitr)
library(magrittr); library(naniar)
library(simputation) # to permit imputation work
library(broom); library(car)
library(equatiomatic); library(patchwork)
library(GGally) # to build scatterplot matrices
library(tidyverse) # always load tidyverse last

theme_set(theme_bw())

opts_chunk$set(comment=NA) # already loaded knitr
options(width = 55) # for the slides
options(dplyr.summarise.inform = FALSE)
```

Today's Agenda

- Regression Analysis: Today we focus on...
 - Data Management
 - Single Imputation
 - How well will retain our R^2 in new data?
 - Interpreting our Coefficients
 - Comparisons In-Sample via ANOVA, AIC, BIC, σ
 - “Uncertainty” Intervals around a model’s coefficients
- A closer look at assumptions, and at collinearity
 - Calibrating yourself

Today's Main Example: 192 adults with diabetes in NE Ohio

Load the dm192.csv Data

```
dm192_raw <- read_csv("data/dm192.csv") %>%  
  clean_names() %>%  
  mutate(across(where(is.character), ~ as.factor(.))) %>%  
  mutate(pt_id = as.character(pt_id)) %>%  
  select(-practice)
```

Anything wrong here?

```
glimpse(dm192_raw)
```

Rows: 192

Columns: 13

```
$ pt_id      <chr> "1", "2", "3", "4", "5", "6", "7", ~  
$ sbp        <dbl> 108, 162, 135, 133, 128, 153, 134, ~  
$ dbp        <dbl> 71, 92, 84, 87, 72, 71, 69, 70, 71,~  
$ a1c        <dbl> 5.8, 11.6, NA, 12.7, 6.8, 5.8, 6.4,~  
$ ldl        <dbl> 58, 54, NA, 112, 105, NA, 151, 98, ~  
$ age        <dbl> 44, 28, 58, 56, 54, 67, 46, 62, 54,~  
$ sex        <fct> male, female, female, male, female,~  
$ race       <fct> black, black, black, black, white, ~  
$ hisp       <fct> no, no, no, no, no, no, no, no, no,~  
$ insurance  <fct> medicaid, medicaid, medicare, medic~  
$ statin     <dbl> 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1,~  
$ sbp_old    <dbl> 110, 158, 142, 145, 140, 152, 134, ~  
$ a1c_old   <fct> 7.6, 12.1, 8.8, 10.9, 6.4, 6.3, 8.7~
```

Why is a1c_old a factor variable?

```
dm192_raw %>% count(a1c_old)
```

```
# A tibble: 75 x 2
  a1c_old     n
  <fct>    <int>
1 #VALUE!      2
2 10            3
3 10.1         4
4 10.2         1
5 10.3         1
6 10.4         2
7 10.5         1
8 10.6         1
9 10.7         2
10 10.9        3
# ... with 65 more rows
```

Let's try importing that again...

```
dm192_raw <- read_csv("data/dm192.csv") %>%  
  clean_names() %>%  
  mutate(a1c_old =  
         ifelse(a1c_old == "#VALUE!", NA, a1c_old)) %>%  
  mutate(a1c_old = as.numeric(a1c_old)) %>%  
  mutate(across(where(is.character), ~ as.factor(.))) %>%  
  mutate(pt_id = as.character(pt_id)) %>%  
  select(-practice)
```

Now, how do things look?

```
head(dm192_raw)
```

```
# A tibble: 6 x 13
  pt_id    sbp    dbp    a1c    ldl    age sex   race  hisp
  <chr>  <dbl> <dbl> <dbl>  <dbl> <dbl> <fct> <fct> <fct>
1 1        108     71    5.8     58     44 male  black no 
2 2        162     92   11.6     54     28 fema~ black no 
3 3        135     84    NA      NA     58 fema~ black no 
4 4        133     87   12.7    112     56 male  black no 
5 5        128     72    6.8     105     54 fema~ white no 
6 6        153     71    5.8     NA     67 male  black no 
# ... with 4 more variables: insurance <fct>,
#   statin <dbl>, sbp_old <dbl>, a1c_old <dbl>
```

```
miss_var_summary(dm192_raw)
```

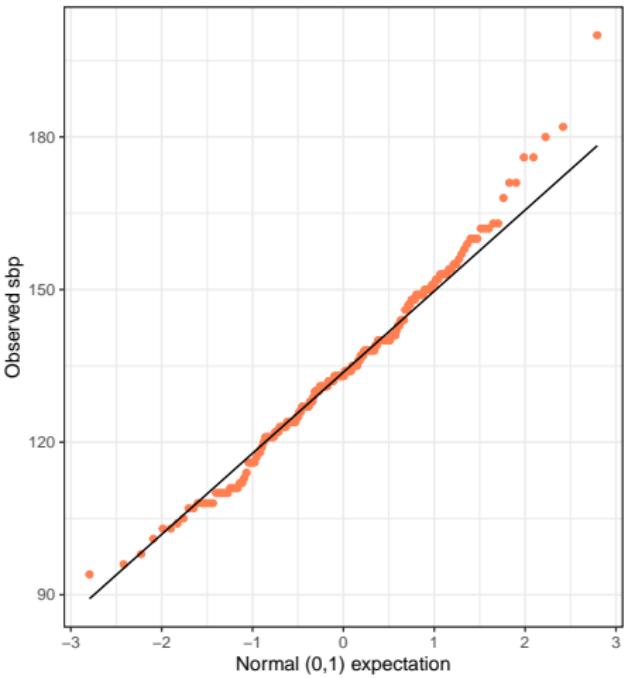
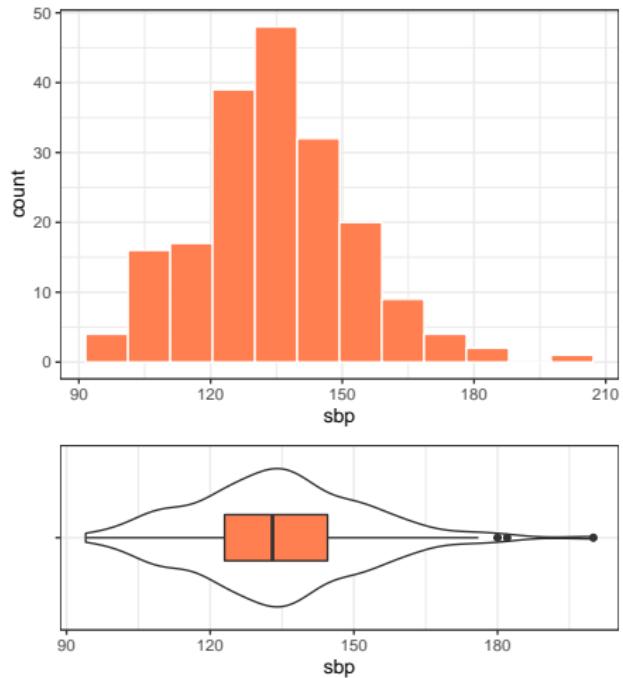
```
# A tibble: 13 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 ldl         43     22.4
2 a1c         4      2.08
3 a1c_old     4      2.08
4 hisp        2      1.04
5 pt_id       0       0
6 sbp         0       0
7 dbp         0       0
8 age         0       0
9 sex         0       0
10 race        0       0
11 insurance   0       0
12 statin      0       0
13 sbp_old     0       0
```

Single Imputation into dm192

```
dm192 <- dm192_raw %>%
  impute_cart(hisp ~ race) %>%
  impute_rlm(ldl ~ age + statin) %>%
  impute_pmm(a1c ~ ldl + age) %>%
  impute_rlm(a1c_old ~ a1c)
```

- Why do I not need to set a seed here?

Systolic BP values in dm192 (our outcome)



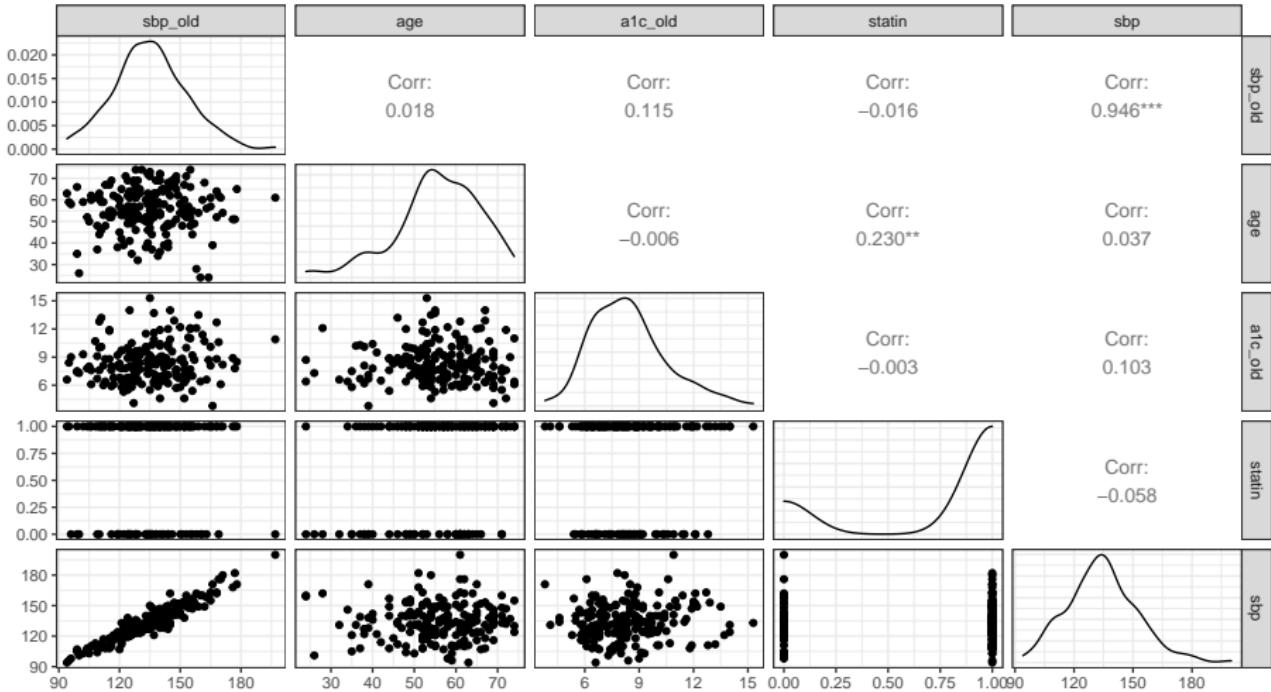
Scatterplot Matrix?

Too many predictors to fit on one slide, so I'll split them...

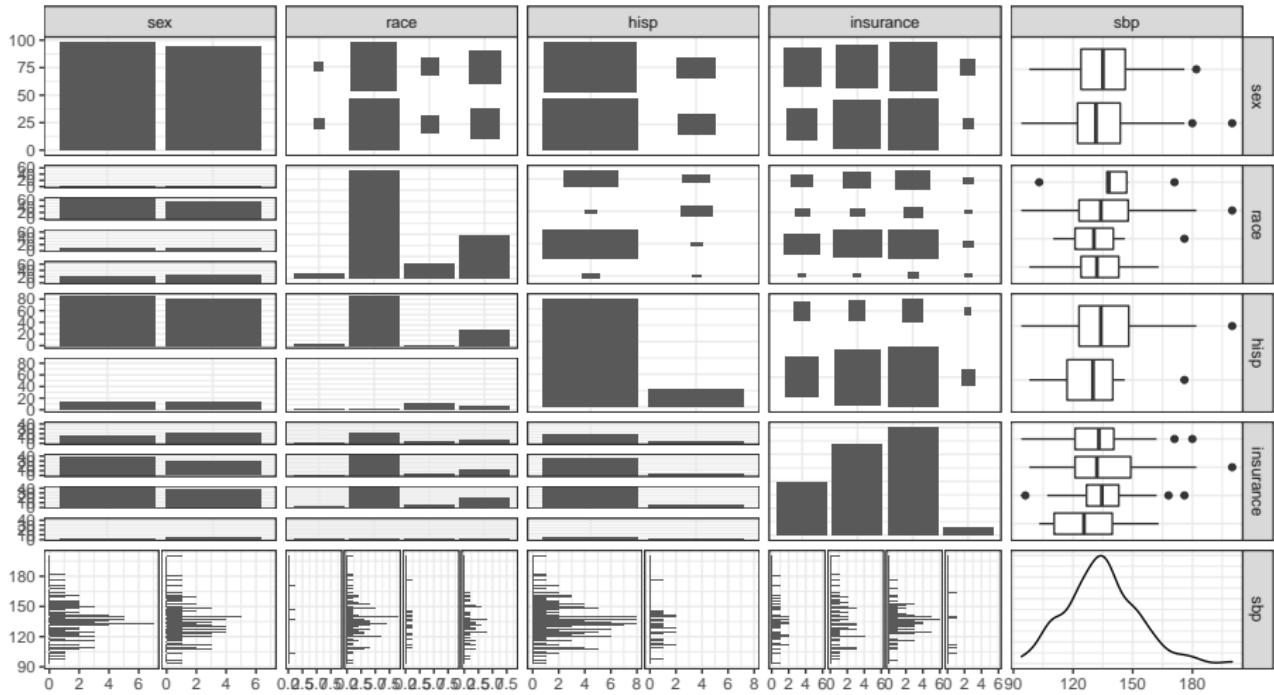
```
dm192 %>% select(sbp_old, age, a1c_old, statin, sbp) %>%  
  ggpairs(.,  
    lower=list(combo=wrap("facethist", binwidth=0.8)))  
  
dm192 %>% select(sex, race, hisp, insurance, sbp) %>%  
  ggpairs(.,  
    lower=list(combo=wrap("facethist", binwidth=0.8)))
```

- Adding the `lower = ...` stuff eliminates the binwidth message.
- Note that I list my outcome `sbp` last in each plot. Why?

Scatterplot Matrix 1 (numerical predictors)



Scatterplot Matrix 2 (categorical predictors)



Consider a transformation of our outcome sbp?

- What are the predictors we'll consider?

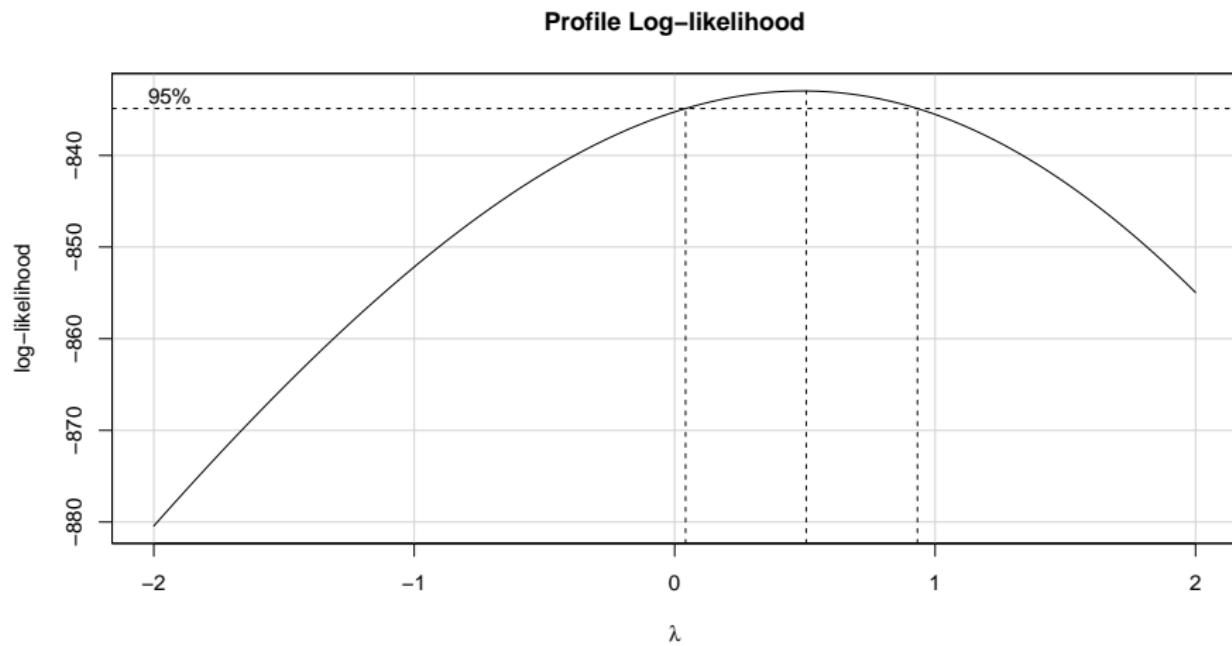
- sbp_old
- age
- sex
- race
- hisp
- insurance
- statin
- a1c_old

So we could fit that model and look at the Box-Cox results...

```
m0 <- lm(sbp ~ sbp_old + age + sex + race + hisp +  
          insurance + statin + a1c_old, data = dm192)
```

Box-Cox results (requires car package)

```
boxCox(m0)
```



Power Transformation results (for m0)

```
summary(powerTransform(m0))
```

bcPower Transformation to Normality

	Est	Power	Rounded	Pwr	Wald Lwr	Bnd Wald	Upr Bnd
Y1	0.491		0.5		0.0471		0.935

Likelihood ratio test that transformation parameter is equal to
(log transformation)

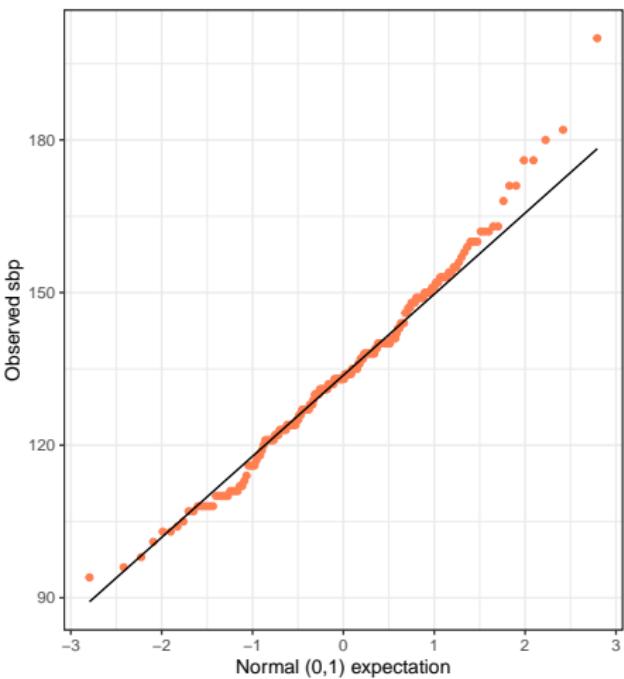
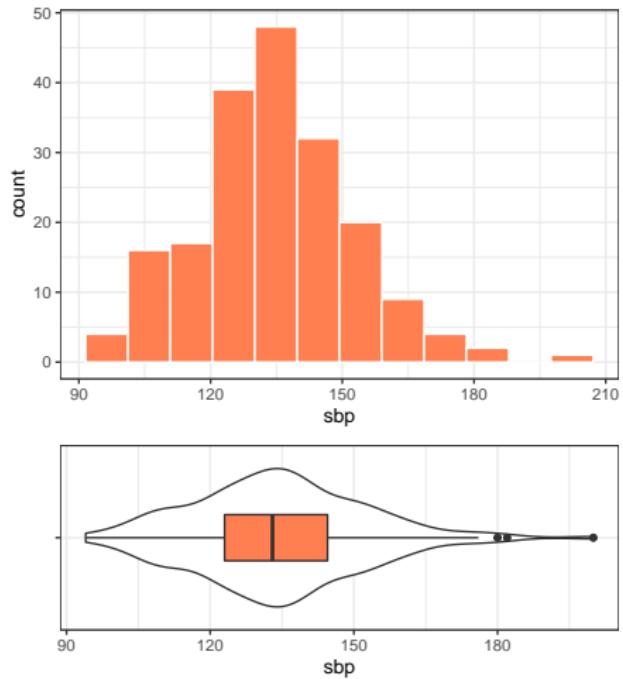
	LRT	df	pval
LR test, lambda = (0)	4.571498	1	0.032508

Likelihood ratio test that no transformation is needed

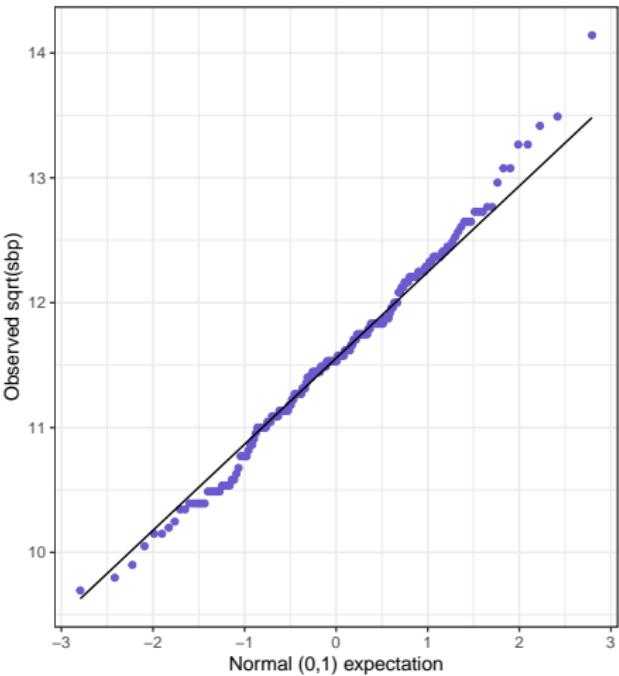
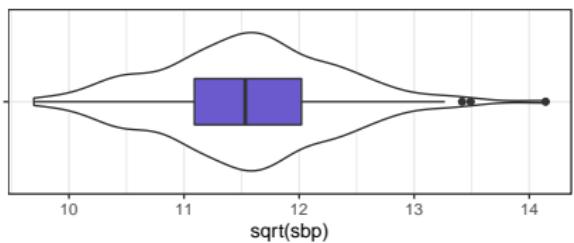
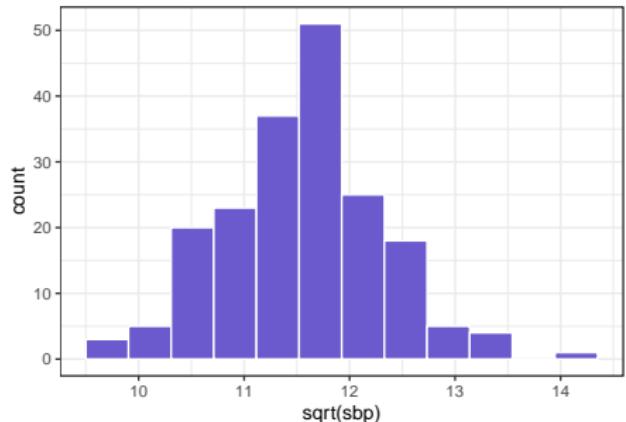
	LRT	df	pval
LR test, lambda = (1)	5.113835	1	0.023736

Note: I suppressed a warning here. What to do?

Systolic BP values in dm192 (untransformed)



Square Root of Systolic BP values in dm192



What if I was going to fit a model to \sqrt{sbp} ?

Before anything else, I'd create a new variable containing the transformed outcome.

```
dm192 <- dm192 %>%
  mutate(sbp_sqrt = sqrt(sbp))
```

and I'd split the data after this, build a scatterplot matrix, run models, all based on this, I think.

So, for instance, if the choice is to run:

```
modelX <- lm(sqrt(sbp) ~ statin + sbp_old, data = dm192)
```

or

```
modelX <- lm(sbp_sqrt ~ statin + sbp_old, data = dm192)
```

it doesn't matter, actually, but for anything else (like a scatterplot matrix, for instance), I'd use `sbp_sqrt`.

Two models for sbp in the dm192 data

For now, let's hold off on a square root transformation - we'll return to this later.

```
m1 <- lm(sbp ~ sbp_old + statin, data = dm192)
m2 <- lm(sbp ~ sbp_old + age + sex + race + hisp +
          insurance + statin + a1c_old, data = dm192)
```

Stepwise Variable Selection?

I'll just note here that if you start with m2, and run

```
step(m2)
```

you wind up with m1. That's how I came up with them as candidate models.

Model Equations with equatiomatic

Remember to include 'results = 'asis'' in the code chunk.

```
extract_eq(m1, use_coefs = TRUE, coef_digits = 3)
```

$$\widehat{\text{sbp}} = 8.63 + 0.94(\text{sbp_old}) - 1.775(\text{statin}) \quad (1)$$

```
extract_eq(m2, use_coefs = TRUE,  
          wrap = TRUE, terms_per_line = 3)
```

$$\begin{aligned}\widehat{\text{sbp}} = & 3.01 + 0.94(\text{sbp_old}) + 0.09(\text{age}) + \\& 0.06(\text{sex}_{\text{male}}) + 1.5(\text{race}_{\text{black}}) + 2.78(\text{race}_{\text{other}}) + \\& 0.67(\text{race}_{\text{white}}) - 0.13(\text{hisp}_{\text{yes}}) + 0.86(\text{insurance}_{\text{medicaid}}) - \\& 0.38(\text{insurance}_{\text{medicare}}) + 1.6(\text{insurance}_{\text{uninsured}}) - 2.27(\text{statin}) - \\& 0.1(\text{a1c_old})\end{aligned} \quad (2)$$

Which model looks better, by R^2 and Adjusted R^2 ?

```
g1 <- glance(m1) %>% mutate(model = "m1")
g2 <- glance(m2) %>% mutate(model = "m2")
comp <- bind_rows(g1, g2)

comp %>% select(model, r.squared, adj.r.squared)
```

```
# A tibble: 2 x 3
  model r.squared adj.r.squared
  <chr>     <dbl>        <dbl>
1 m1         0.897       0.896
2 m2         0.900       0.894
```

- Which of these two models is more likely to **retain its nominal R^2 value** in new data?

Which model looks better, by σ , AIC or BIC?

```
comp %>% select(model, sigma, AIC, BIC)
```

```
# A tibble: 2 x 4
  model sigma    AIC    BIC
  <chr> <dbl> <dbl> <dbl>
1 m1     5.72 1220. 1233.
2 m2     5.80 1234. 1280.
```

Is one model detectably better than the other in-sample?

```
anova(m1, m2)
```

Analysis of Variance Table

Model 1: sbp ~ sbp_old + statin

Model 2: sbp ~ sbp_old + age + sex + race + hisp + insurance + a1c_old

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	189	6192.1				
2	179	6022.8	10	169.28	0.5031	0.8863

Interpreting the slopes and their CIs

```
tidy(m1, 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)	8.63	3.25	3.25	14.01
sbp_old	0.94	0.02	0.90	0.98
statin	-1.78	0.98	-3.39	-0.16

- Averaging over repeated samples, we can be 90% confident that the true slope of `sbp_old` lies between 0.90 and 0.98, assuming that `statin` is unchanged.

Model m1, and 90% uncertainty intervals

Let's describe these as *uncertainty intervals*, since they are meant to help you understand how much uncertainty you have.

```
tidy(m1, 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)	8.63	3.25	3.25	14.01
sbp_old	0.94	0.02	0.90	0.98
statin	-1.78	0.98	-3.39	-0.16

“Uncertainty Interval” is kind of nice because it fights the ambiguity between confidence intervals and predictive intervals. Also notice that confidence intervals are smaller when you have more confidence, which can confuse people. (references in a few slides)

Model m2, and 50% uncertainty intervals (Code)

```
tidy(m2, conf.int = TRUE, conf.level = 0.50) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

50% intervals have some potential advantages over 95% intervals...

- Computational Stability
- More intuitive (half the 50% intervals should contain the true value)
- Sometimes it's best to get a sense of where the parameters will be, not to attempt an unrealistic near-certainty.

Model m2, and 50% uncertainty intervals (Results)

term	estimate	std.error	conf.low	conf.high
(Intercept)	3.01	5.77	-0.89	6.91
sbp_old	0.94	0.02	0.93	0.96
age	0.09	0.06	0.05	0.13
sexmale	0.06	0.86	-0.52	0.64
raceblack	1.50	2.75	-0.36	3.36
raceother	2.78	3.23	0.60	4.96
racewhite	0.67	2.84	-1.26	2.59
hispyes	-0.13	1.66	-1.26	0.99
insurancemedicaid	0.86	1.25	0.01	1.70
insurancemedicare	-0.38	1.21	-1.20	0.43
insuranceuninsured	1.60	2.66	-0.19	3.40
statin	-2.27	1.03	-2.96	-1.57
a1c_old	-0.10	0.21	-0.25	0.04

Model m1, and 50% uncertainty intervals

```
tidy(m1, conf.int = TRUE, conf.level = 0.50) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	8.63	3.25	6.43	10.83
sbp_old	0.94	0.02	0.92	0.96
statin	-1.78	0.98	-2.43	-1.12

Under repeated sampling, the true slope of sbp_old will be inside (0.92, 0.96) half of the time, assuming statin is held constant.

Andrew Gelman Blog Posts Worth a Little Time

- Instead of “confidence interval,” let’s say “uncertainty interval” at https://andrewgelman.com/2010/12/21/lets_say_uncert/
- “Why I prefer 50% rather than 95% intervals” at <https://andrewgelman.com/2016/11/05/why-i-prefer-50-to-95-intervals/>
- “Abraham Lincoln and confidence intervals” at <https://andrewgelman.com/2016/11/23/abraham-lincoln-confidence-intervals/>

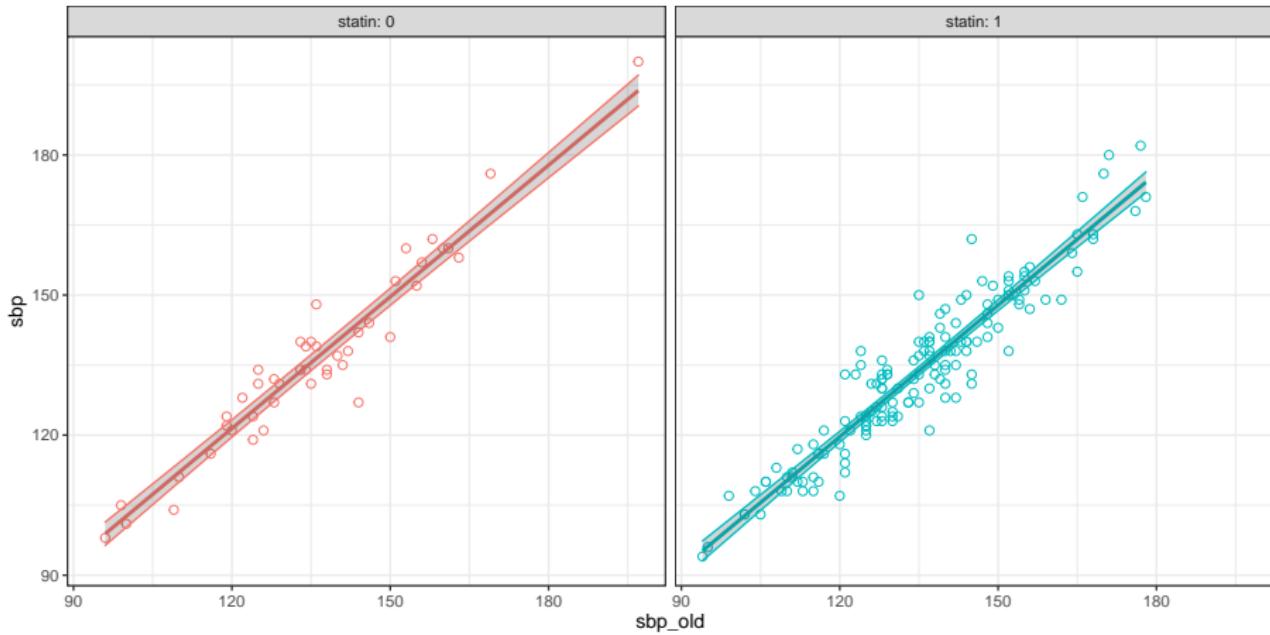
Plotting Model m1 (Code)

```
m1_aug <- augment(m1, data = dm192, se_fit = TRUE)

ggplot(m1_aug, aes(x = sbp_old, y = sbp,
                    col = factor(statin))) +
  geom_point(pch = 1, size = 2) +
  geom_line(aes(y = .fitted), size = 1) +
  geom_ribbon(aes(ymin = .fitted - .se.fit*2,
                  ymax = .fitted + .se.fit*2),
              alpha = 0.2) +
  facet_wrap(~ statin, labeller = "label_both") +
  guides(col = "none") +
  labs(title = "Model m1",
       subtitle = "with approximate 95% uncertainty intervals")
```

Plotting Model m1 (Result)

Model m1
with approximate 95% uncertainty intervals



Residual Plots and Regression Assumptions

Multivariate Regression: Checking Assumptions

Assumptions (see Course Notes, Section 27)

- Linearity
- Normality
- Homoscedasticity
- Independence

Available Residual Plots

```
plot(model, which = c(1:3,5))
```

- ① Residuals vs. Fitted Values
- ② Normal Q-Q Plot of Standardized Residuals
- ③ Scale-Location Plot
- ④ Index Plot of Cook's Distance
- ⑤ Residuals, Leverage and Influence

An Idealized Model (by Simulation)

```
set.seed(431122)

x1 <- rnorm(200, 20, 5)
x2 <- rnorm(200, 20, 12)
x3 <- rnorm(200, 20, 10)
er <- rnorm(200, 0, 1)
y <- .3*x1 - .2*x2 + .4*x3 + er

sim0 <- tibble(y, x1, x2, x3)

mod0 <- lm(y ~ x1 + x2 + x3, data = sim0)

summary(mod0) # appears on next slide
```

Call:

```
lm(formula = y ~ x1 + x2 + x3, data = sim0)
```

An Idealized Model (by Simulation)

```
Call: lm(formula = y ~ x1 + x2 + x3, data = sim0)
```

```
Residuals: Min 1Q Median 3Q Max  
-3.14553 -0.68079 0.08096 0.69216 2.65265
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.122852 0.348584 0.352 0.725  
x1 0.285539 0.014211 20.093 <2e-16 ***  
x2 -0.204908 0.005828 -35.159 <2e-16 ***  
x3 0.413308 0.007172 57.631 <2e-16 ***
```

```
---
```

```
Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.007 on 196 degrees of freedom
```

```
Multiple R-squared: 0.9589, Adjusted R-squared: 0.9583
```

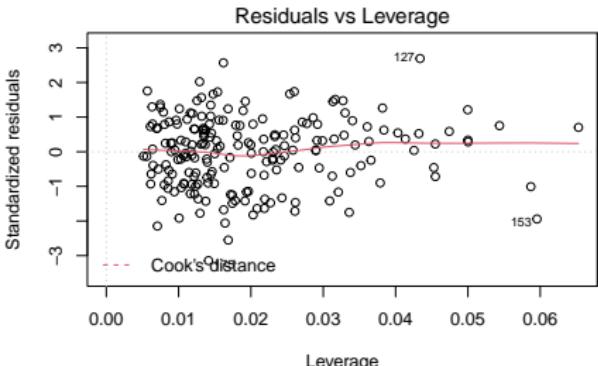
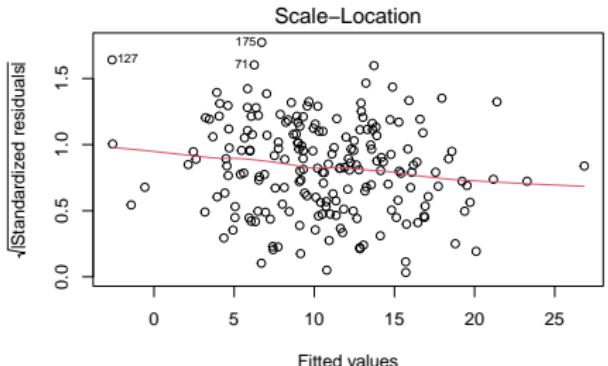
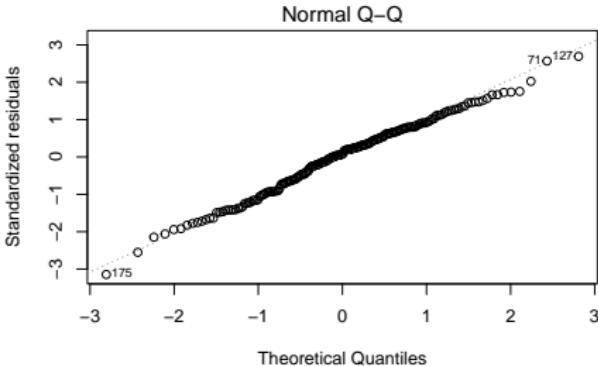
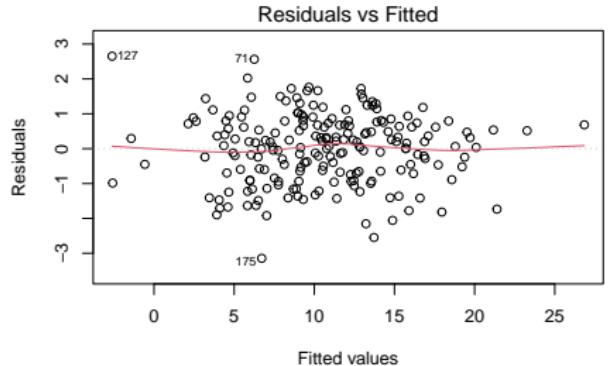
```
F-statistic: 1524 on 3 and 196 DF, p-value: < 2.2e-16
```

Building Residual Plots for Idealized Model

```
par(mfrow=c(2,2))
plot(mod0)
par(mfrow=c(1,1))
```

- Residuals vs. Fitted values (Top Left)
- Normal Q-Q plot of Standardized Residuals (Top Right)
- Scale-Location plot (Bottom Left)
- Residuals vs. Leverage, Cook's Distance contours (Bottom Right)

Residual Analysis (Idealized Model: n = 200)



Is one of the regression assumptions violated?

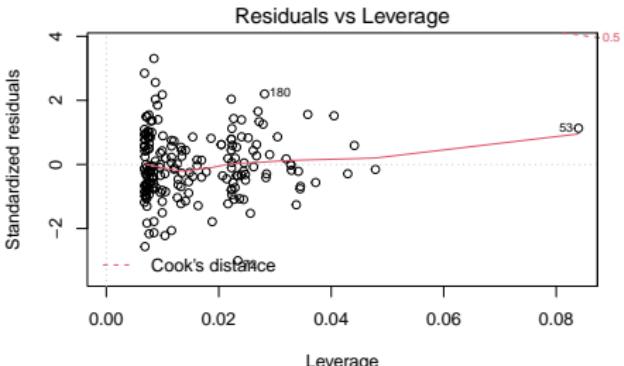
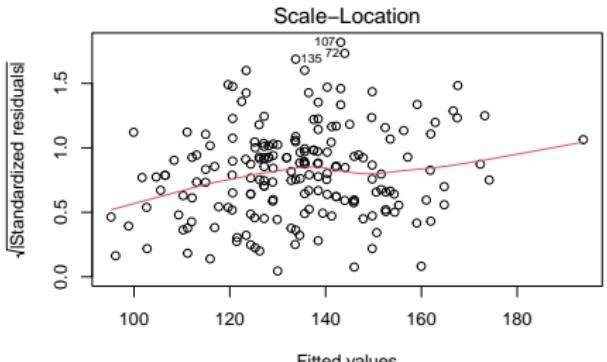
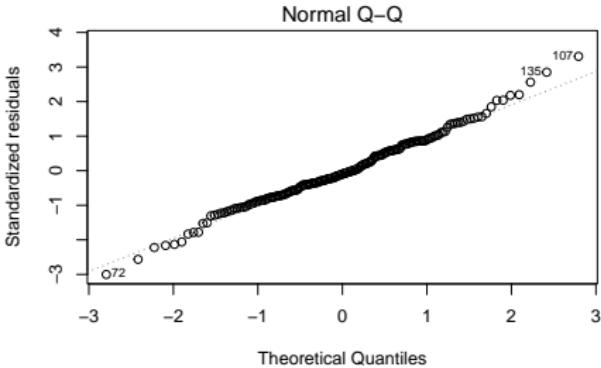
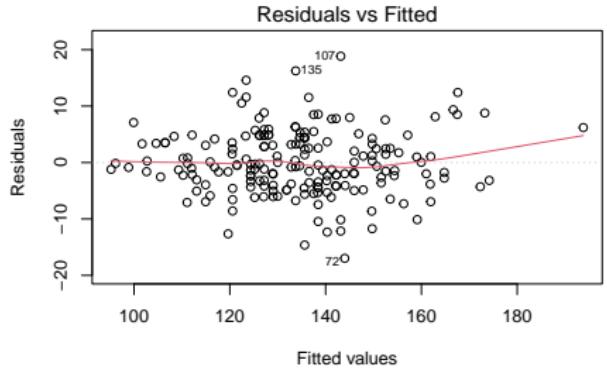
- Non-linearity problems
 - show up as a curve in the Top Left plot (Residuals vs. Fitted)
- Heteroscedasticity problems
 - show up as a fan in the Top Left plot
 - show up as a trend (up or down) in the Scale-Location plot
- Non-Normality problems
 - shows up as individual outliers in all plots
 - Normal Q-Q plot describes skew / many outliers / a few big outliers
 - Bottom Right plot shows each point's residual, leverage and influence

What to Do?

Importance of Assumptions:

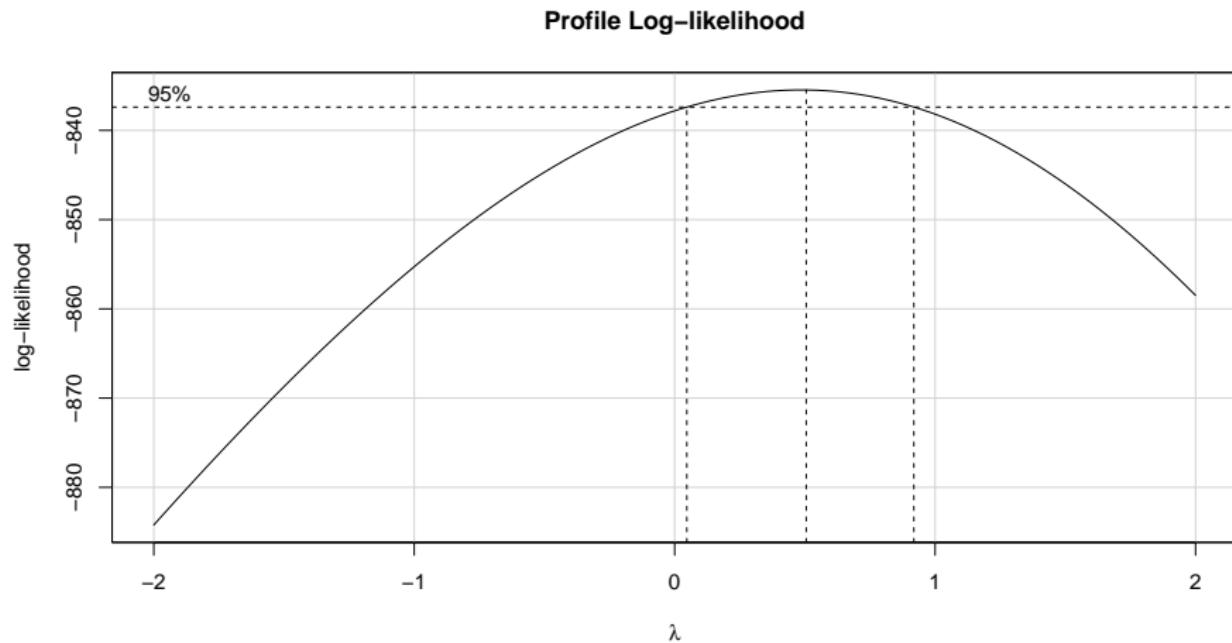
- ① Linearity (critical, but amenable to transformations, often)
- ② Independence (critical, not relevant if data are a cross-section with no meaningful ordering in space or time, but vitally important if space/time play a meaningful role - longitudinal data analysis required)
- ③ Homoscedasticity (constant variance: important, sometimes amenable to transformation)
- ④ Normality due to skew (usually amenable to transformation)
- ⑤ Normality due to many more outliers than we would expect (heavy-tailed - inference is problematic unless you account for this, sometimes a transformation can help)
- ⑥ Normality due to a severe outlier (or a small number of severely poorly fitted points - can consider setting those points away from modeling, but requires a meaningful external explanation)

Residual Plots for Model m1 for our dm192 data?



Recall the Box-Cox plot's suggestion, earlier...

```
boxCox(m1)
```



Adjusted model m1 predicting \sqrt{sbp}

```
m1_adj <- lm(sqrt(sbp) ~ sbp_old + statin, data = dm192)
summary(m1_adj)
```

Call:

```
lm(formula = sqrt(sbp) ~ sbp_old + statin, data = dm192)
```

Residuals:

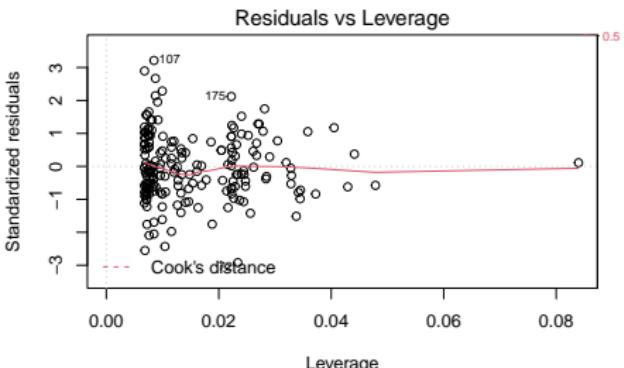
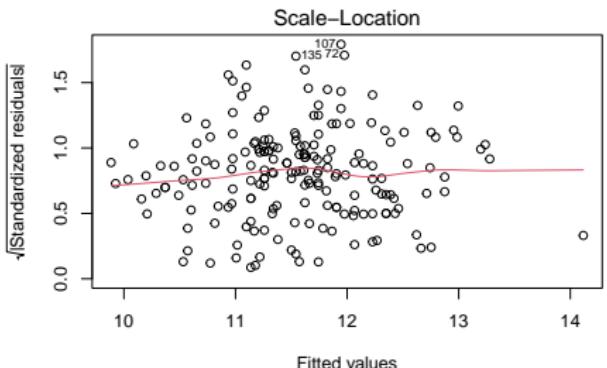
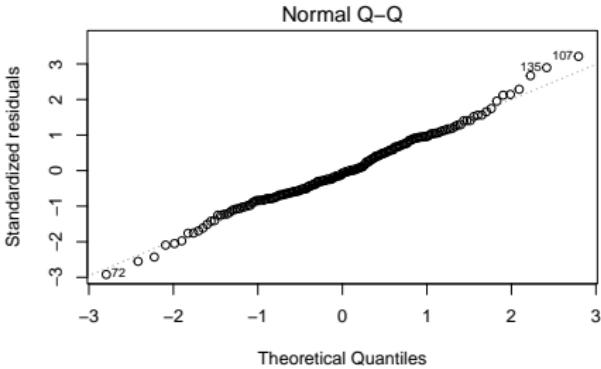
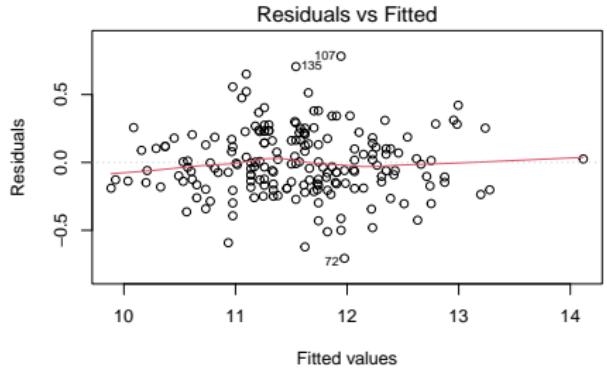
Min	1Q	Median	3Q	Max
-0.70616	-0.15787	-0.01856	0.16667	0.78269

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.1590466	0.1390482	44.294	<2e-16 ***
sbp_old	0.0403926	0.0009897	40.813	<2e-16 ***
statin	-0.0707505	0.0416825	-1.697	0.0913 .

Signif. codes:

Residuals for m1 predicting square root of sbp



Resolving Assumption Violations

Options include:

- transform the Y variable, likely with one of our key power transformations (use Box-Cox to help)
- transform one or more of the X variables if it seems particularly problematic, or perhaps combine them (rather than height and weight, perhaps look at BMI, or BMI and height to help reduce collinearity)
- remove a point only if you have a good explanation for the point that can be provided outside of the modeling, and this is especially important if the point is influential
- consider other methods for establishing a non-linear model (432: splines, loess smoothers, non-linear modeling)
- consider other methods for longitudinal data with substantial dependence (432)

Six Simulations To Help You Calibrate Yourself

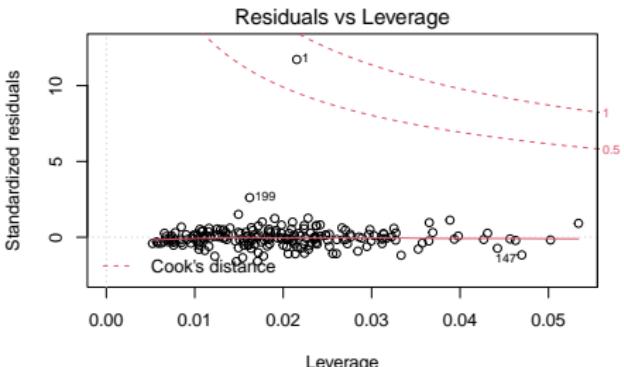
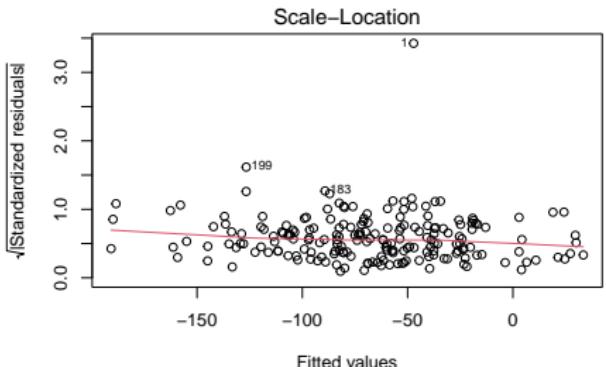
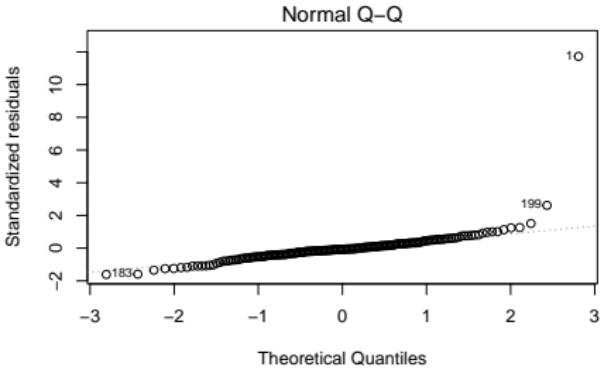
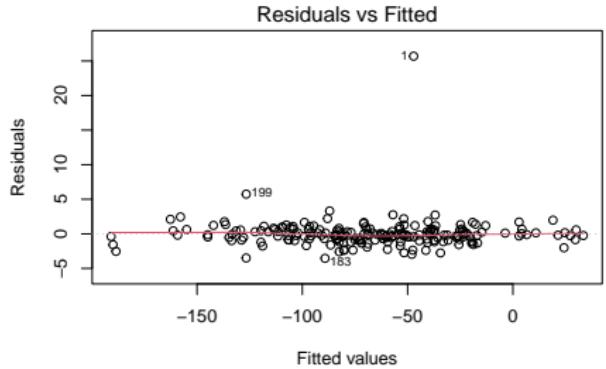
For each simulation, decide on the following:

Is one of the regression assumptions violated?

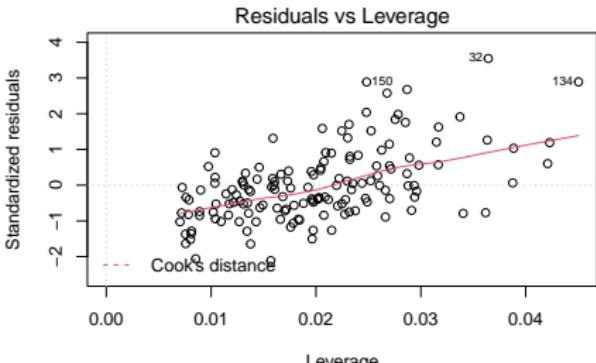
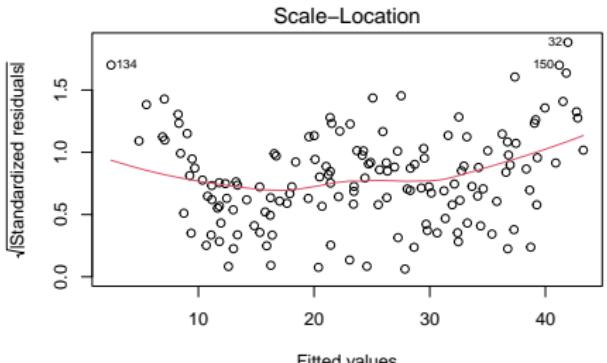
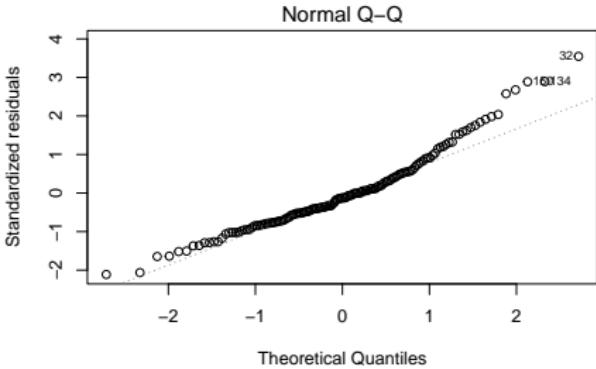
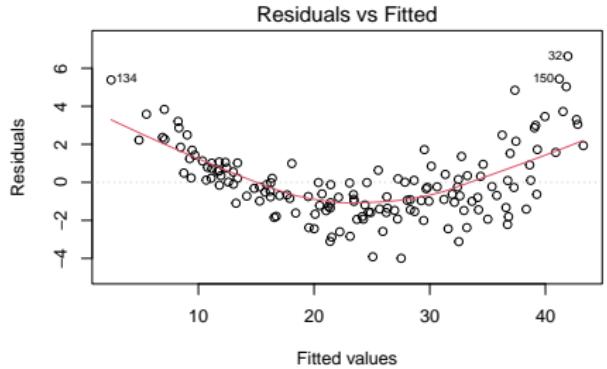
- Linearity, Homoscedasticity, Normality, or multiple problems?
 - All of these simulations describe cross-sectional data, with no importance to the order of the observations, so the assumption of independence isn't a concern.
- In which of the four plot(s) shown do you see the problem?
 - Top Left: Residuals vs. Fitted values (in R: plot 1)
 - Top Right: Normal Q-Q plot of Standardized Residuals (plot 2)
 - Bottom Left: Scale-Location plot (plot 3)
 - Bottom Right: Residuals vs. Leverage, Cook's Distance contours (plot 5)
- If you see a point that is problematic, then:
 - is it poorly fit?
 - is it highly leveraged?
 - is it influential?
- What might you try to do about the assumption problem you see (if you see one), to resolve it?

This **isn't** easy. We'll do three, and then regroup.

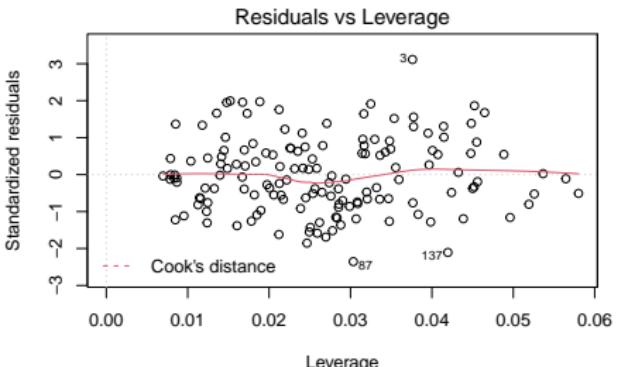
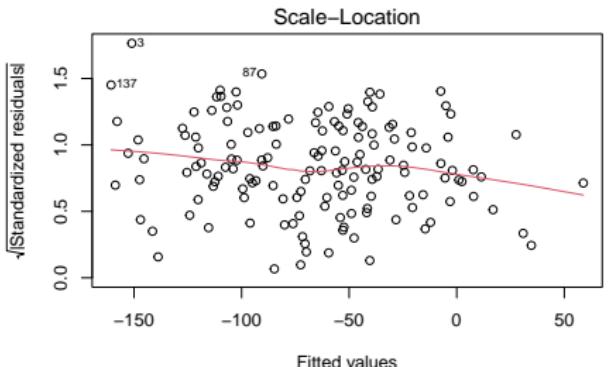
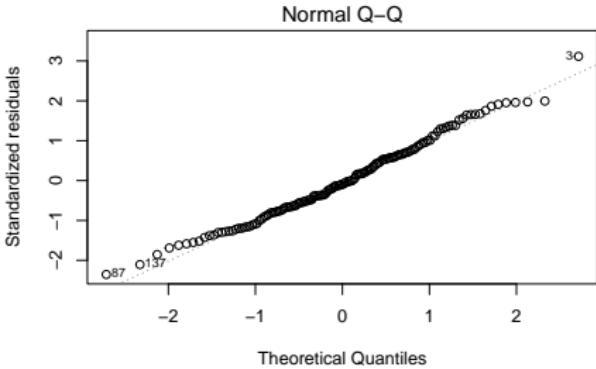
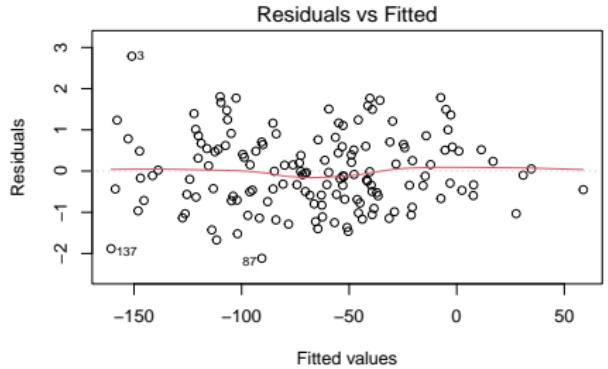
Simulation 1 ($n = 200$ subjects)



Simulation 2 ($n = 150$)



Simulation 3 ($n = 150$)



OK. How are we doing so far?

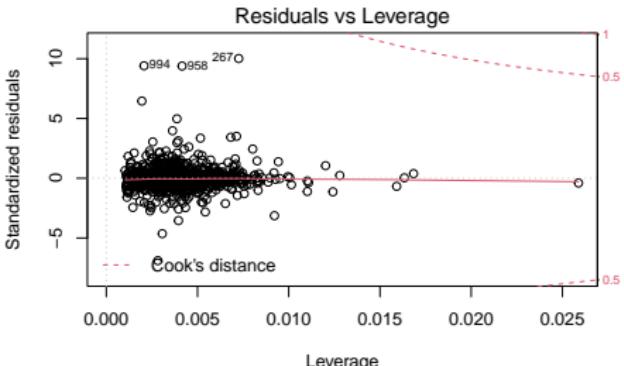
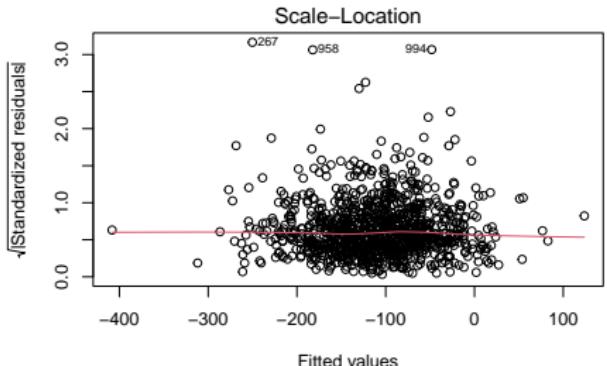
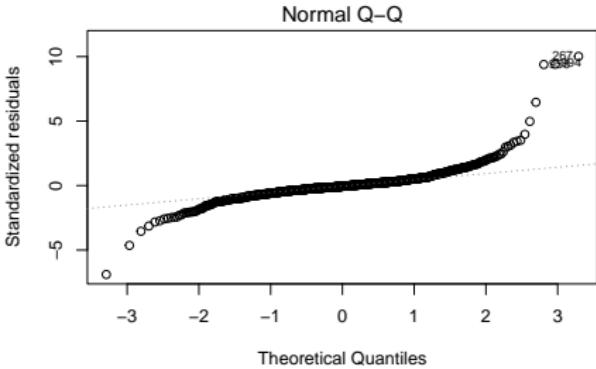
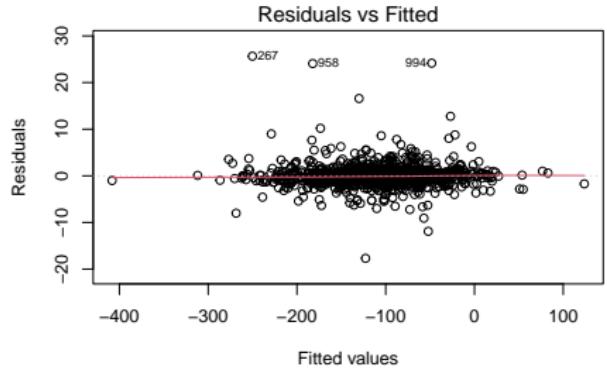
The First Three Simulations

For those of you playing along at home...

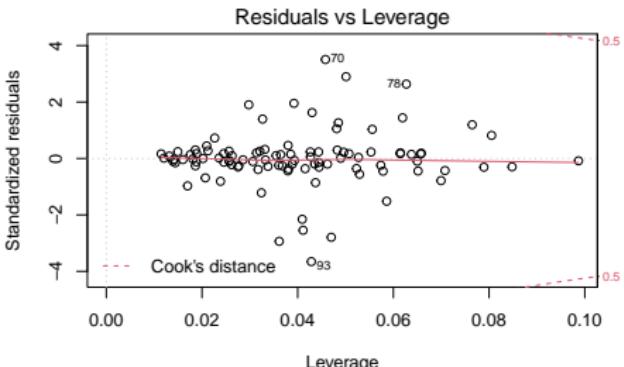
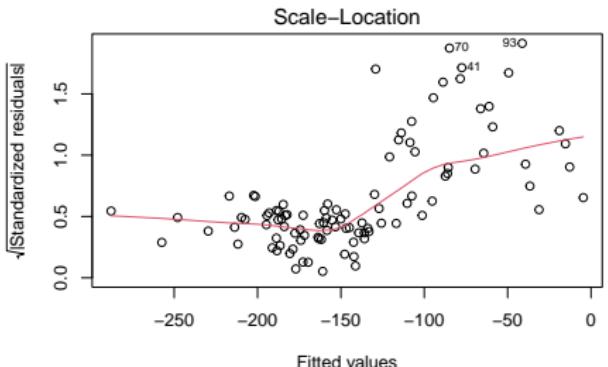
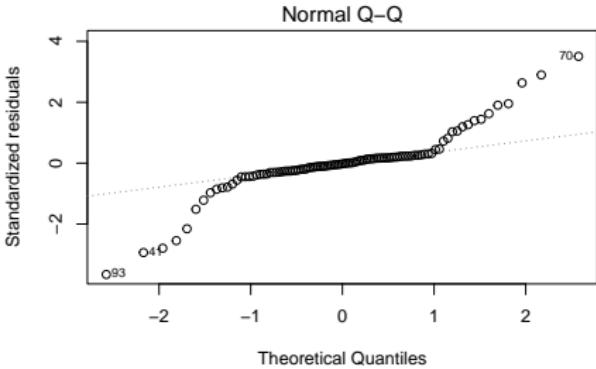
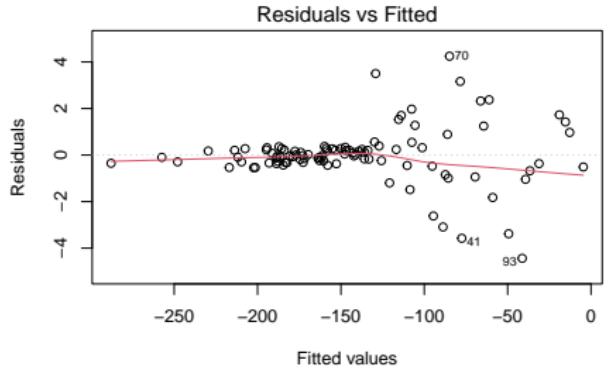
- ① Observation 1 has an impossibly large standardized residual (Z score is close to 12), of substantial influence (Cook's distance around 0.7).
- Probably need to remove the point, and explain it separately.
- ② Curve in residuals vs. fitted values plot suggests potential non-linearity.
 - Natural choice would be a transformation of the outcome.
- ③ No substantial problems, although there's a little bit of heteroscedasticity.
 - I'd probably just go with the model as is.

Let's try three more...

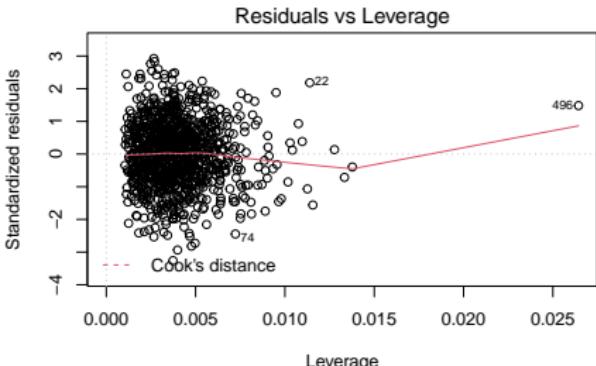
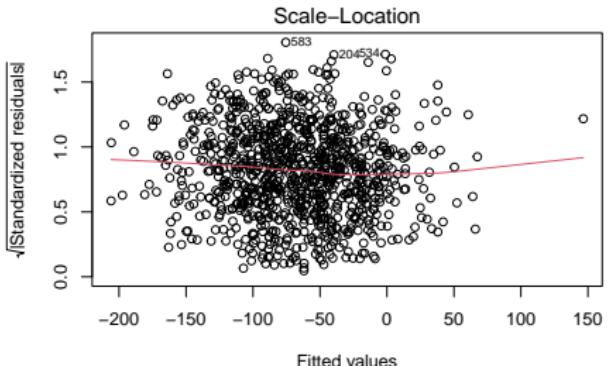
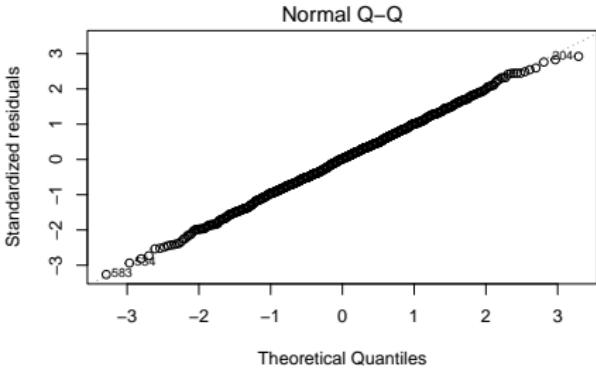
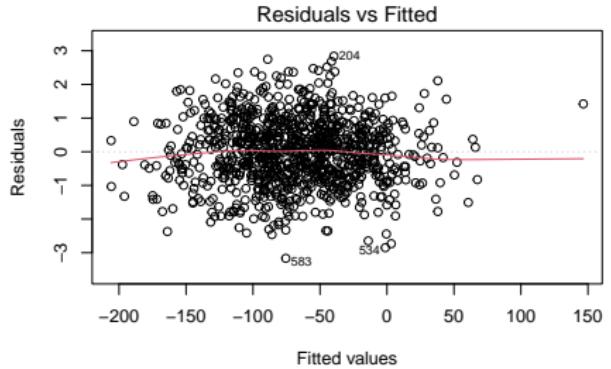
Simulation 4 ($n = 1000$)



Simulation 5 ($n = 100$)



Simulation 6 ($n = 1000$)



OK. How did this go?

The Last Three Simulations

For those of you playing along at home . . .

- ④ Normality issues - outlier-prone even with 1000 observations.
- Transform Y? Consider transforming the Xs?
- ⑤ Serious heteroscedasticity - residuals much more varied for larger fitted values.
- Look at Residuals vs. each individual X to see if this is connected to a specific predictor, which might be skewed or something?
- ⑥ No serious violations - point 496 has very substantial leverage, though.
- I'd probably just go with the model as is, after making sure that point 496's X values aren't incorrect.

What about Collinearity?

Do we have collinearity in our dm192 models?

```
# requires library(car)
```

```
vif(m1)
```

```
sbp_old statin  
1.000271 1.000271
```

```
car::vif(m2)
```

	GVIF	Df	GVIF^(1/(2*Df))
sbp_old	1.053618	1	1.026459
age	1.790969	1	1.338271
sex	1.047137	1	1.023297
race	2.155455	3	1.136553
hisp	1.909531	1	1.381858
insurance	1.921500	3	1.114996
statin	1.084574	1	1.041429
a1c_old	1.094353	1	1.046113

What about collinearity?

“No collinearity” is not a regression assumption, but if we see substantial collinearity, we are inclined to consider dropping some of the variables, or combining them (height and weight may be highly correlated, height and BMI may be less so).

The variance inflation factor (or VIF), if it exceeds 5, is a clear indication of collinearity. We'd like to see the variances inflated only slightly (that is, VIF not much larger than 1) by correlation between the predictors, to facilitate interpretation.

The best way to tell if you've improved the situation by fitting an alternative model is to actually compare and fit the two models, looking in particular at:

- the standard errors of their coefficients, and
- their VIFs.

What's the Goal Here?

Develop an effective model. (?) (!)

- Models can do many different things. What you're using the model for matters, a lot.
- Don't fall into the trap of making binary decisions (this model isn't perfect, no matter what you do, and so your assessment of residuals will also have shades of gray).
- The tools we have provided (scatterplots, mostly) are well designed for rather modest sample sizes. When you have truly large samples, they don't scale very well.
- Just because R chooses four plots for you to study doesn't mean they provide the only relevant information.
- Embrace the uncertainty. Look at it as an opportunity to study your data more effectively.

431 Class 22

thomaselove.github.io/431

2021-11-11

Today's Agenda: Some Loose Ends

- Working with the Favorite Movies data
- When is a complete-case analysis reasonable? MCAR!
- ANOVA
 - Assessing Assumptions with Data Visualizations
 - Dealing with Multiple Comparisons via the Holm approach
 - Kruskal-Wallis rank-based alternative
- Wilcoxon rank-based procedures for comparing pseudo-medians
 - in paired samples (Wilcoxon signed rank)
 - in independent samples (Wilcoxon rank sum)

Today's R Setup, Part 1

```
library(knitr); library(magrittr); library(naniar)
library(broom); library(patchwork)

library(googlesheets4)
source("data/Love-boost.R")

library(janitor)
```

Attaching package: 'janitor'

The following objects are masked from 'package:stats':

chisq.test, fisher.test

Today's R Setup, Part 2

```
library(tidyverse)

-- Attaching packages ----- tidyverse 1.3.1 --
v ggplot2 3.3.5      v purrr    0.3.4
v tibble   3.1.5      v dplyr    1.0.7
v tidyr    1.1.4      v stringr  1.4.0
v readr    2.0.2      vforcats  0.5.1

-- Conflicts ----- tidyverse_conflicts() --
x tidyr::extract()  masks magrittr::extract()
x dplyr::filter()   masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

- The 8 core tidyverse packages are listed every time you load the tidyverse (unless you tell R not to show messages.)

```
theme_set(theme_bw())
```

Favorite Movies

```
url_raw <- "https://docs.google.com/spreadsheets/d/1t4668vGN-9  
  
movie_raw <- read_sheet(url_raw, na = c("", "NA")) %>%  
  arrange(film_id) %>%  
  clean_names()  
  
dim(movie_raw)  
  
[1] 115 69  
  
write_rds(movie_raw, "data/movie_raw.Rds")
```

Or, more simply, since we have the .Rds file

```
movie_raw <- read_rds("data/movie_raw.Rds")  
  
dim(movie_raw)
```

```
[1] 115 69
```

Missing Data?

```
miss_var_summary(movie_raw)
```

```
# A tibble: 69 x 3
  variable      n_miss  pct_miss
  <chr>        <int>     <dbl>
1 country2       66     57.4
2 imdb_cat3      33     28.7
3 locations       14     12.2
4 budget_est      13     11.3
5 prod_co_2       12     10.4
6 imdb_cat2       10      8.70
7 domestic_gross     6      5.22
8 worldwide_gross    3      2.61
9 rt_critics       2      1.74
10 bom_link        2      1.74
# ... with 59 more rows
```

Picking Out A Few Variables

```
movie1 <- movie_raw %>%
  select(film_id, film, imdb_stars, mpa)

head(movie1) %>% kable()
```

film_id	film	imdb_stars	mpa
1	8 1/2	8.0	NR
2	2001: A Space Odyssey	8.3	G
3	About Time	7.8	R
4	Avatar	7.8	PG-13
5	Avengers: Endgame	8.4	PG-13
6	Avengers: Infinity War	8.4	PG-13

How many films in each mpa category?

```
movie1 %>% tabyl(mpa)
```

```
mpa    n      percent
      G  3  0.02608696
      NR 7  0.06086957
      PG 30 0.26086957
PG-13 33 0.28695652
      R 42 0.36521739
```

Let's look at the three categories with at least 30 films.

```
movie1 <- movie1 %>%
  filter(mpa %in% c("PG", "PG-13", "R"))
```

```
nrow(movie1)
```

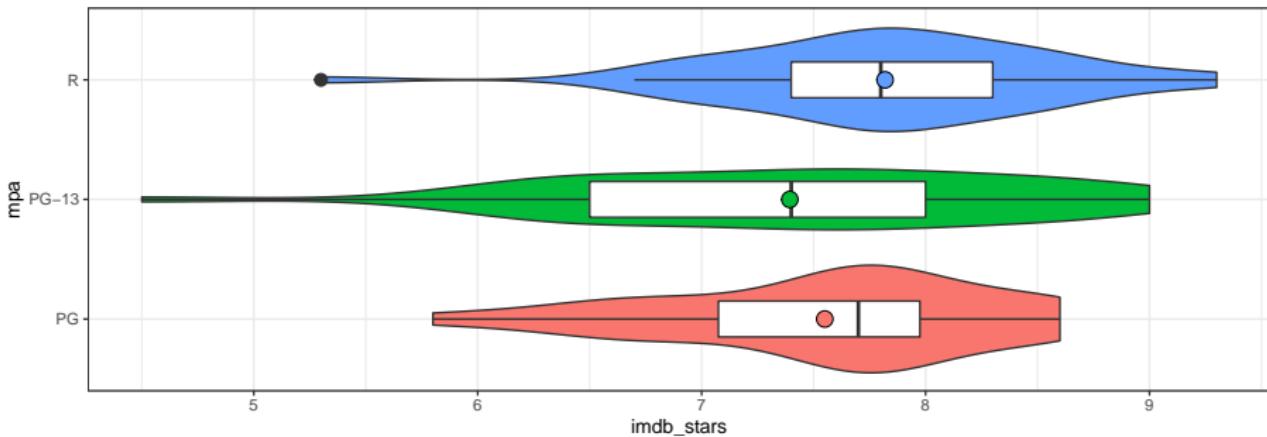
```
[1] 105
```

ANOVA assumptions

- ① Approximately Normal distribution of outcome within each group
 - Can use a rank-based alternative (Kruskal-Wallis) if there is a serious concern, or consider a transformation of the outcome
- ② Equal population variance of outcome within each group (extension of pooled t)
- ③ Independently drawn samples of the outcome within each group

Plot IMDB Stars by MPA Rating

```
ggplot(movie1, aes(x = mpa, y = imdb_stars)) +  
  geom_violin(aes(fill = mpa)) +  
  geom_boxplot(width = 0.3, outlier.size = 3) +  
  stat_summary(aes(fill = mpa), fun = mean,  
               geom="point", pch = 21, size = 4) +  
  guides(fill = "none") + coord_flip()
```



Numerical Summary of IMDB Stars by MPA Rating

```
mosaic::favstats(imdb_stars ~ mpa, data = movie1) %>%
  kable(digits = 2)
```

mpa	min	Q1	median	Q3	max	mean	sd	n	missing
PG	5.8	7.08	7.7	7.97	8.6	7.55	0.72	30	0
PG-13	4.5	6.50	7.4	8.00	9.0	7.39	1.01	33	0
R	5.3	7.40	7.8	8.30	9.3	7.82	0.74	42	0

Analysis of Variance of IMDB Stars by MPA

```
mod1 <- lm(imdb_stars ~ mpa, data = movie1)
anova(mod1) %>% kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mpa	2	3.484	1.742	2.529	0.085
Residuals	102	70.259	0.689	NA	NA

```
tidy(mod1, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  kable(digits = 3)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	7.550	0.152	7.298	7.802
mpaPG-13	-0.156	0.209	-0.504	0.191
mpaR	0.269	0.198	-0.060	0.598

Bonferroni pairwise comparisons for mpa groups

```
movie1 %$%
pairwise.t.test(imdb_stars, mpa,
                 p.adjust.method = "bonferroni")
```

Pairwise comparisons using t tests with pooled SD

data: imdb_stars and mpa

PG PG-13

PG-13 1.00 -

R 0.53 0.09

P value adjustment method: bonferroni

Holm pairwise comparisons across mpa groups

- Works well even with an unbalanced design so long as ANOVA assumptions hold.
- Not as conservative as Bonferroni, but uniformly more powerful.

```
movie1 %$%
pairwise.t.test(imdb_stars, mpa,
                 p.adjust.method = "holm")
```

Pairwise comparisons using t tests with pooled SD

data: imdb_stars and mpa

	PG	PG-13
PG-13	0.46	-
R	0.36	0.09

P value adjustment method: holm

Tukey HSD set of Pairwise Comparisons

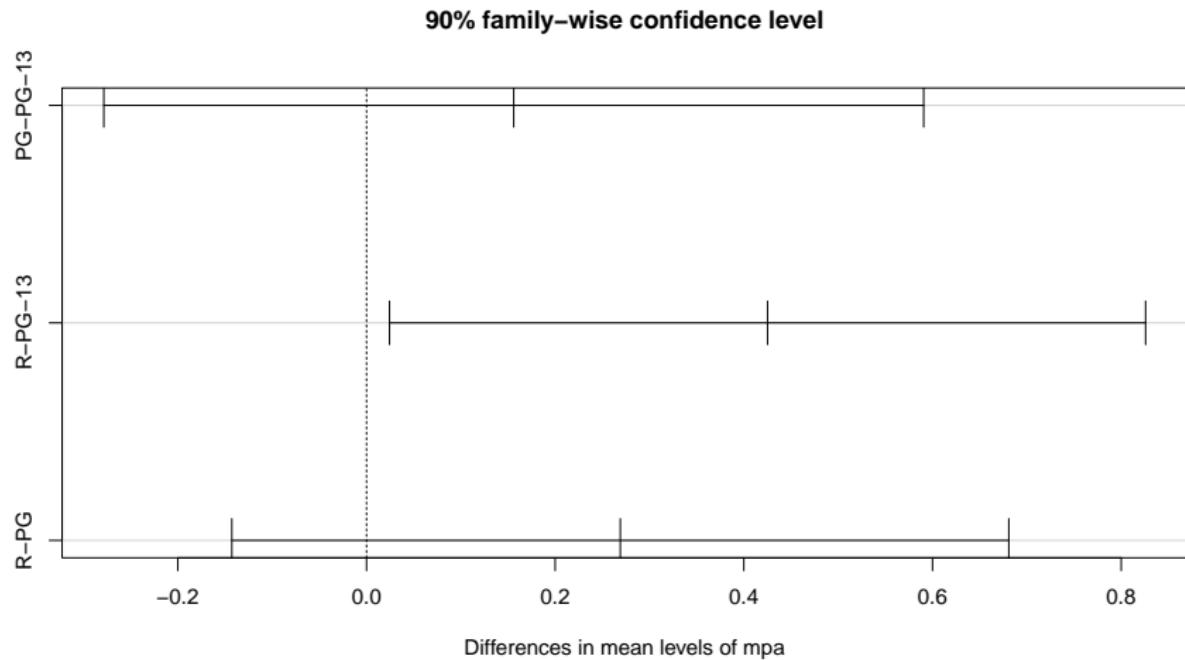
```
tukey1 <- movie1 %$%
  TukeyHSD(aov(imdb_stars ~ mpa),
            ordered = TRUE, conf.level = 0.90)

tidy(tukey1) %>% kable(digits = 3)
```

term	contrast	null.value	estimate	conf.low	conf.high	adj.p.value
mpa	PG-PG-13	0	0.156	-0.279	0.591	0.737
mpa	R-PG-13	0	0.425	0.024	0.826	0.076
mpa	R-PG	0	0.269	-0.143	0.681	0.368

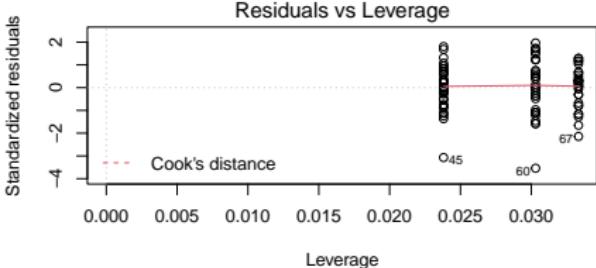
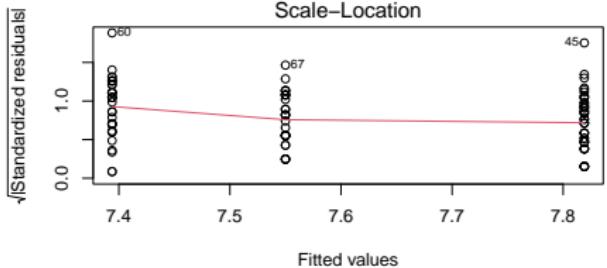
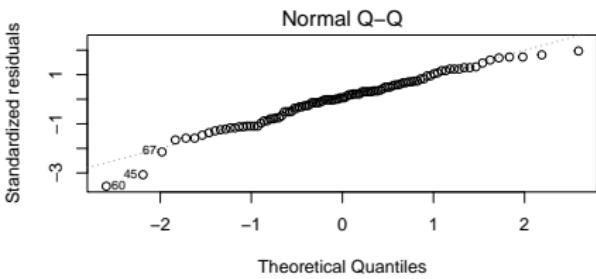
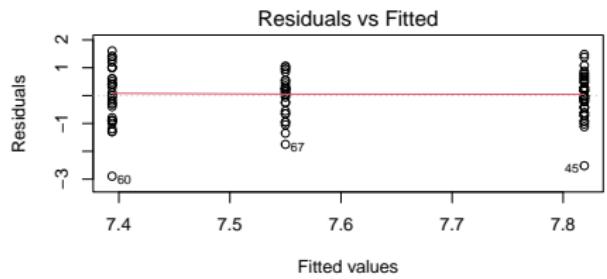
Tukey HSD set of Pairwise Comparisons

```
plot(tukey1)
```



Residual Analysis for our ANOVA model

```
par(mfrow = c(2,2)); plot(mod1); par(mfrow = c(1,1))
```



Kruskal-Wallis Rank-Based ANOVA Approach

```
kruskal.test(imdb_stars ~ mpa, data = movie1)
```

Kruskal-Wallis rank sum test

```
data: imdb_stars by mpa
Kruskal-Wallis chi-squared = 4.1131, df = 2,
p-value = 0.1279
```

- No longer comparing means, and no confidence intervals here.
- No straightforward decision about what to do about pairwise comparisons, other than Holm-based comparisons based on Wilcoxon rank sum tests.
- Speaking of Wilcoxon rank-based tests...

A New Question...

Are students in 431 more likely to have seen movies that were nominated for Academy Awards?

- How many of the 115 movies received Oscar nominations?

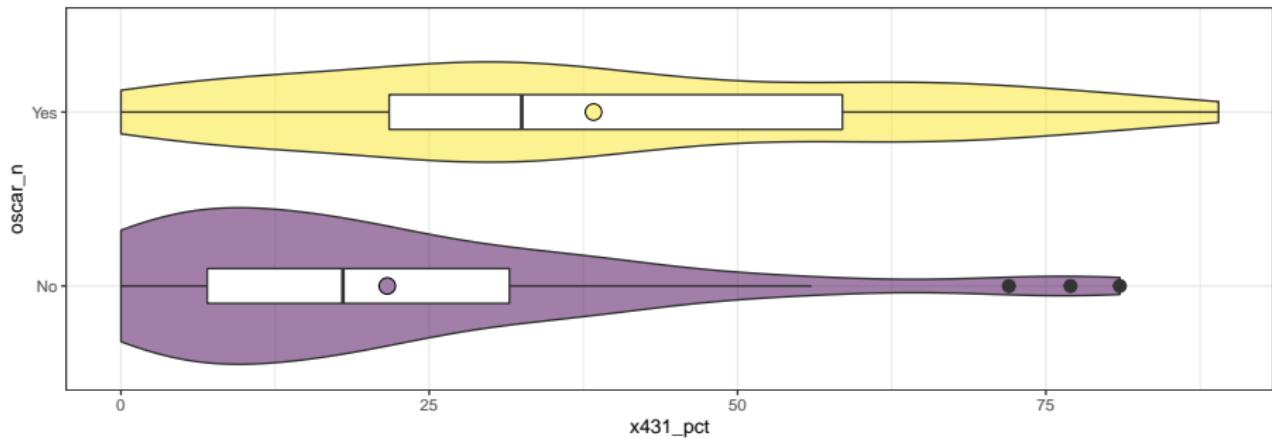
```
movie2 <- movie_raw %>%
  select(film_id, film, x431_pct, oscar_noms) %>%
  mutate(oscar_n = factor(
    ifelse(oscar_noms > 0, "Yes", "No")))
```

```
movie2 %>% tabyl(oscar_n)
```

oscar_n	n	percent
No	47	0.4086957
Yes	68	0.5913043

% of students who've seen film, by Oscar Status?

```
ggplot(movie2, aes(x = oscar_n, y = x431_pct)) +  
  geom_violin(aes(fill = oscar_n)) +  
  geom_boxplot(width = 0.2, outlier.size = 3) +  
  stat_summary(aes(fill = oscar_n), fun = mean,  
               geom="point", pch = 21, size = 4) +  
  guides(fill = "none") + coord_flip() +  
  scale_fill_viridis_d(alpha = 0.5)
```



Comparing Two Population Means

Does this design involve paired samples, or independent samples? Why?

```
mosaic::favstats(x431_pct ~ oscar_n, data = movie2) %>%  
  kable(digits = 2)
```

oscar_n	min	Q1	median	Q3	max	mean	sd	n	missing
No	0	7.00	18.0	31.5	81	21.60	20.38	47	0
Yes	0	21.75	32.5	58.5	89	38.32	24.22	68	0

Comparing 431 % by Oscar Nomination Status

We have four different approaches. Two are based on the t distribution.

```
# pooled t
compA <- t.test(x431_pct ~ oscar_n, data = movie2,
                  var.equal = TRUE, conf.level = 0.90)

tidy(compA) %>%
  select(estimate1, estimate2, estimate,
         conf.low, conf.high) %>% kable(dig = 2)
```

estimate1	estimate2	estimate	conf.low	conf.high
21.6	38.32	-16.73	-23.88	-9.58

Comparing 431 % by Oscar Nomination Status

Here's the second one based on the t distribution.

```
# Welch's t
compB <- t.test(x431_pct ~ oscar_n, data = movie2,
                  conf.level = 0.90)

tidy(compB) %>%
  select(estimate1, estimate2, estimate,
         conf.low, conf.high) %>% kable(dig = 2)
```

estimate1	estimate2	estimate	conf.low	conf.high
21.6	38.32	-16.73	-23.66	-9.79

Comparing 431 % by Oscar Nomination Status

This strategy doesn't use the t distribution.

```
# bootstrap
set.seed(431)
compC <- movie2 %$% bootdif(x431_pct, oscar_n,
                           conf.level = 0.90)

compC
```

Mean Difference	0.05	0.95
16.727785	9.971558	23.421621

Comparing 431 % by Oscar Nomination Status

What if we rank the observations from low to high in each group, then compare the results?

```
# Wilcoxon-Mann-Whitney rank sum with continuity correction  
compD <- wilcox.test(x431_pct ~ oscar_n, data = movie2,  
                      conf.int = TRUE, conf.level = 0.90)
```

```
tidy(compD) %>% select(estimate, conf.low, conf.high) %>%  
  kable(digits = 2)
```

estimate	conf.low	conf.high
-17	-24	-10

Interpreting the Rank Sum Test

The Wilcoxon-Mann-Whitney rank sum test (also called the Mann-Whitney U test and lots of other things) tests the null hypothesis that if we randomly select X and Y from the two populations of interest, the probability of X being greater than Y is the same as the probability of Y being greater than X.

- The “Pseudomedian” is sometimes referred to as the Hodges-Lehmann estimate. It is the median of all possible differences between an observation in the first sample and an observation in the second sample.

To write up a Wilcoxon rank sum result, we usually suggest specifying the two medians directly, and then describing the p value or (less commonly) a confidence interval.

The Wilcoxon-Mann-Whitney test requires only that the data be ordinal, and this reduces the influence of outliers. It doesn't test the same thing as the t test (or bootstrap) however.

Comparing 431 % by Oscar Nomination Status

Four Comparisons for these Independent Samples

Method	Yes - No Est.	90% CI	Statistic
Pooled t	16.73	(9.58, 23.88)	Mean
Welch's t	16.73	(9.79, 23.66)	Mean
Bootstrap	16.73	(9.97, 23.42)	Mean
Rank Sum	17	(10, 24)	"Pseudo-median"

Numerical Summaries from the Data

oscar_n	n	mean	median	sd
No	47	21.596	18.0	20.378
Yes	68	38.324	32.5	24.223

Comparing A New Outcome under 2 Conditions

Suppose we want to compare the percentage of **critics** who recommend a film to the percentage of **the general audience** who recommend the film. Each film has:

- `rt_critics` = from Rotten Tomatoes: percentage of critics who recommend the film (sample mean across 113 films was 80.6)
- `rt_audience` = from Rotten Tomatoes: percentage of audience who recommend the film (sample mean across 114 films was 82.6)

Can we compare the difference between the means (at least for the 113 films with data on each variable) appropriately?

What's the design we have here?

Creating our Third Data Set

```
movie3 <- movie_raw %>%
  filter(complete.cases(rt_critics, rt_audience)) %>%
  select(film_id, film, rt_critics, rt_audience)
```

```
dim(movie3)
```

```
[1] 113    4
```

We lost two films (Farewell My Concubine and Jab We Met) which didn't have information on each of our variables.

The movie3 data

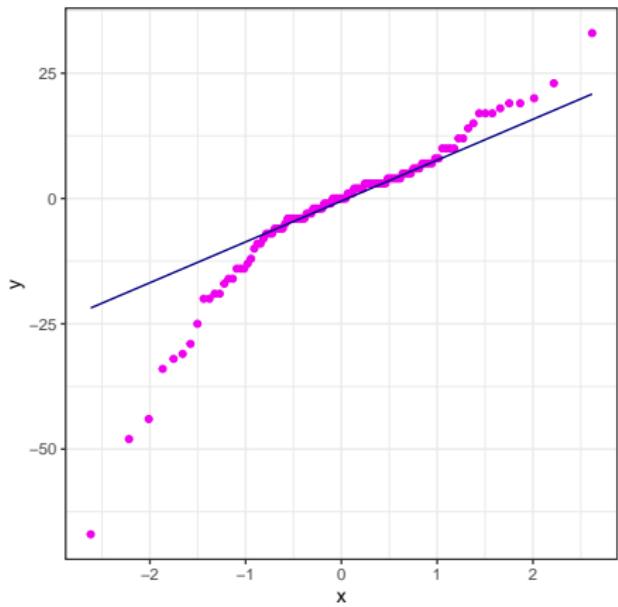
```
movie3 <- movie3 %>%
  mutate(diff = rt_critics - rt_audience)

head(movie3)

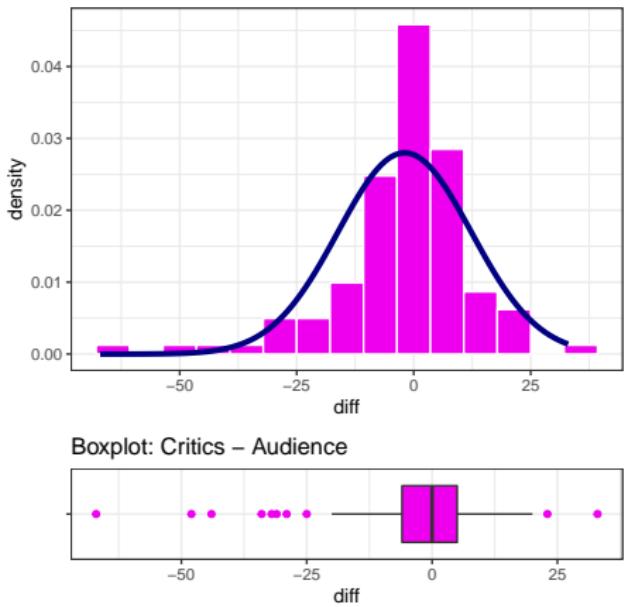
# A tibble: 6 x 5
  film_id film                rt_critics rt_audience diff
  <dbl>   <chr>                  <dbl>        <dbl>    <dbl>
1     1 8 1/2                      98          92       6
2     2 2001: A Space O~           92          89       3
3     3 About Time                 69          81      -12
4     4 Avatar                     81          82      -1
5     5 Avengers: Endga~          94          90       4
6     6 Avengers: Infin~          85          91      -6
```

We can develop paired differences in this paired samples setting. What do those differences look like?

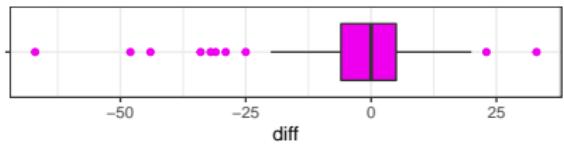
Normal Q–Q: Critics – Audience



Density + Normal: Critics – Audience



Boxplot: Critics – Audience



min	Q1	median	Q3	max	mean	sd	n	missing
-67	-6	0	5	33	-1.9	14.2	113	0

Paired T approach

```
compX <- movie3 %$% t.test(diff, conf.level = 0.90)

tidy(compX) %>%
  select(estimate, conf.low, conf.high, p.value) %>%
  kable(dig = 2)
```

estimate	conf.low	conf.high	p.value
-1.95	-4.17	0.28	0.15

Bootstrap approach

```
set.seed(431431)
compY <- movie3 %$%
  Hmisc::smean.cl.boot(diff, conf.int = 0.90)
```

compY

Mean	Lower	Upper
-1.9469027	-4.2491150	0.1712389

Wilcoxon Signed Rank approach

```
compZ <- movie3 %$%
  wilcox.test(diff, conf.int = 0.90)

tidy(compZ) %>%
  select(estimate, conf.low, conf.high, p.value) %>%
  kable(dig = 2)
```

estimate	conf.low	conf.high	p.value
-0.5	-3	1.5	0.62

Comparing % Recommend by Critics vs. Audiences

Three Comparisons for these Paired Samples

Method	Crit - Aud Est.	90% CI	Statistic
t	-1.95	(-4.17, 0.28)	Mean difference
Bootstrap	-1.95	(-4.25, 0.17)	Mean difference
Signed Rank	-0.5	(-3, 1.5)	"Pseudo-median" difference

Summarizing the Data

	group	mean	sd	median	n	min	max
...1	critics	80.58	17.61	86	113	21	100
...2	audience	82.53	12.24	86	113	42	98
...3	c-a diffs	-1.95	14.25	0	113	-67	33

431 Class 23

thomaselove.github.io/431

2021-11-16

Today's Agenda

On Contingency Tables (Chapter 24)

- Building a $J \times K$ Table
- χ^2 Tests of Independence
 - Cochran Conditions and Checking Assumptions

Replicable Research and the Crisis in Science

- ASA 2016 Statement on P values (Context, Process, Purpose)
- Is changing the p value cutoff the right strategy?
- Second-generation p values: A next step?
- ASA 2019 Statement on Statistical Inference in the 21st Century

Today's Setup

```
library(janitor)
library(magrittr)
library(patchwork)
library(vcd)
library(tidyverse)

theme_set(theme_bw())
```

Working with Larger Cross-Tabulations

A 2×3 contingency table

This table displays the count of patients who show *complete*, *partial*, or *no response* after treatment with either **active** medication or a **placebo** in a study of 100 patients...

Group	None	Partial	Complete
Active	8	24	20
Placebo	12	26	10

Is there a statistically detectable association here, at $\alpha = 0.10$?

- H_0 : Response Distribution is the same, regardless of Treatment.
- H_A : There is an association between Treatment and Response.

The Pearson Chi-Square Test

The Pearson χ^2 test assumes the null hypothesis is true (rows and columns are independent.) That is a model for our data. How does it work? Here's the table, with marginal totals added.

-	None	Partial	Complete	TOTAL
Active	8	24	20	52
Placebo	12	26	10	48
TOTAL	20	50	30	100

The test needs to estimate the expected frequency in each of the six cells under the assumption of independence. If the rows and columns are in fact independent of each other, then what is the expected number of subjects that will fall in the Active/None cell?

The Independence Model

The independence model means the overall rate of “Response = None” or “Partial” or “Complete” applies to both “Active” and “Placebo” subjects.

-	None	Partial	Complete	TOTAL
Active	-	-	-	52
Placebo	-	-	-	48
TOTAL	20	50	30	100

If the rows and columns were independent, then:

- 20/100 or 20% of subjects would have the response “None”
 - That means 20% of the 52 Active, and 20% of the 48 Placebo subjects.
- 50% would have a “Partial” response in each exposure group, and
- 30% would have a “Complete” response in each group.

So, can we fill in the expected frequencies under our independence model?

Observed (*Expected*) Cell Counts

-	None	Partial	Complete	TOTAL
Active	8 (10.4)	24 (26.0)	20 (15.6)	52
Placebo	12 (9.6)	26 (24.0)	10 (14.4)	48
TOTAL	20	50	30	100

General Formula for Expected Frequencies under Independence

$$\text{Expected Frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand Total}}$$

This assumes that the independence model holds: the probability of being in a particular column is exactly the same in each row, and vice versa.

Chi-Square Assumptions

- Expected Frequencies: We assume that the expected frequency, under the null hypothesized model of independence, will be **at least 5** (and ideally at least 10) in each cell. If that is not the case, then the χ^2 test is likely to give unreliable results. The *Cochran conditions* require us to have no cells with zero and at least 80% of the cells in our table with expected counts of 5 or higher. That's what R uses to warn you of trouble.
- Don't meet the standards? Consider collapsing categories.

Observed (Expected) Cell Counts

-	None	Partial	Complete	TOTAL
Active	8 (10.4)	24 (26.0)	20 (15.6)	52
Placebo	12 (9.6)	26 (24.0)	10 (14.4)	48
TOTAL	20	50	30	100

Getting the Table into R

We'll put the table into a matrix in R. Here's one approach...

```
T1 <- matrix(c(8, 24, 20, 12, 26, 10),  
             ncol=3, nrow=2, byrow=TRUE)  
rownames(T1) <- c("Active", "Placebo")  
colnames(T1) <- c("None", "Partial", "Complete")  
T1
```

	None	Partial	Complete
Active	8	24	20
Placebo	12	26	10

Chi-Square Test Results in R

- H_0 : Response Distribution is the same, regardless of Treatment.
 - Rows and Columns of the table are *independent*
- H_A : There is an association between Treatment and Response.
 - Rows and Columns of the table are *associated*.

```
chisq.test(T1)
```

Pearson's Chi-squared test

```
data: T1  
X-squared = 4.0598, df = 2, p-value = 0.1313
```

What is the conclusion?

Does Sample Size Affect The χ^2 Test?

- T1 results were: $\chi^2 = 4.0598$, $df = 2$, $p = 0.1313$
- What if we had the same pattern, but twice as much data?

```
T1_doubled <- T1*2
```

```
T1_doubled
```

	None	Partial	Complete
Active	16	48	40
Placebo	24	52	20

```
chisq.test(T1_doubled)
```

Pearson's Chi-squared test

```
data: T1_doubled  
X-squared = 8.1197, df = 2, p-value = 0.01725
```

Can we run Fisher's exact test instead?

Yes, but . . . if the Pearson assumptions don't hold, then the Fisher's test is not generally an improvement.

```
fisher.test(T1)
```

Fisher's Exact Test for Count Data

```
data: T1  
p-value = 0.1358  
alternative hypothesis: two.sided
```

- It's also really meant more for square tables, with the same number of rows as columns, and relatively modest sample sizes.

OK. Back to dm1000

```
dm1000 <- read_rds("data/dm_1000.Rds") %>%  
  select(subject, tobacco, insurance) %>%  
  filter(complete.cases(.))
```

```
head(dm1000)
```

```
# A tibble: 6 x 3  
  subject tobacco insurance  
  <chr>    <fct>   <fct>  
1 M-0001   Current Medicaid  
2 M-0002   Never   Commercial  
3 M-0003   Former  Medicare  
4 M-0004   Never   Medicaid  
5 M-0005   Never   Medicare  
6 M-0006   Current Medicaid
```

Arrange the Factors in a Useful Order

```
dm1000 <- dm1000 %>%
  mutate(tobacco =
    fct_relevel(tobacco, "Current", "Former"),
  insurance =
    fct_relevel(insurance, "Medicare",
                "Commercial", "Medicaid"))

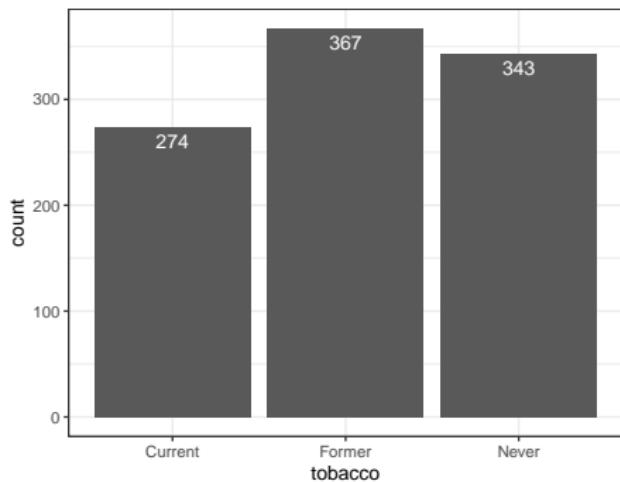
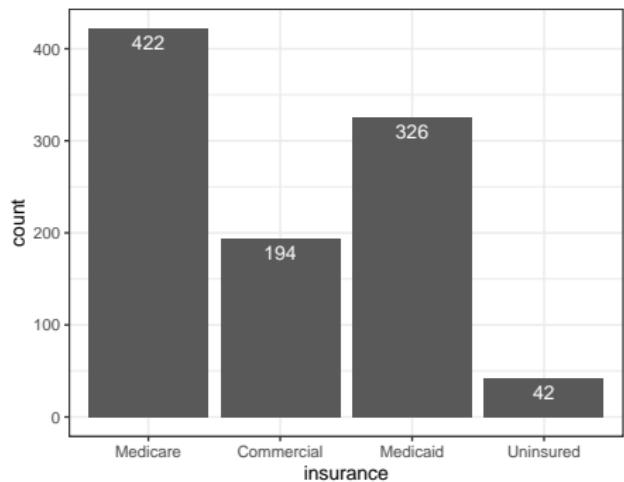
dm1000 %>% tabyl(tobacco, insurance) %>%
  adorn_totals(where = c("row", "col"))
```

	tobacco	Medicare	Commercial	Medicaid	Uninsured	Total
Current	99	44	118	13	274	
Former	183	70	103	11	367	
Never	140	80	105	18	343	
Total	422	194	326	42	984	

What am I plotting here?

```
p1 <- ggplot(dm1000, aes(x = insurance)) + geom_bar() +  
  theme_bw() +  
  geom_text(aes(label = ..count..), stat = "count",  
            vjust = 1.5, col = "white")  
  
p2 <- ggplot(dm1000, aes(x = tobacco)) + geom_bar() +  
  theme_bw() +  
  geom_text(aes(label = ..count..), stat = "count",  
            vjust = 1.5, col = "white")  
  
p1 + p2
```

dm1000: Two Categorical Variables of interest



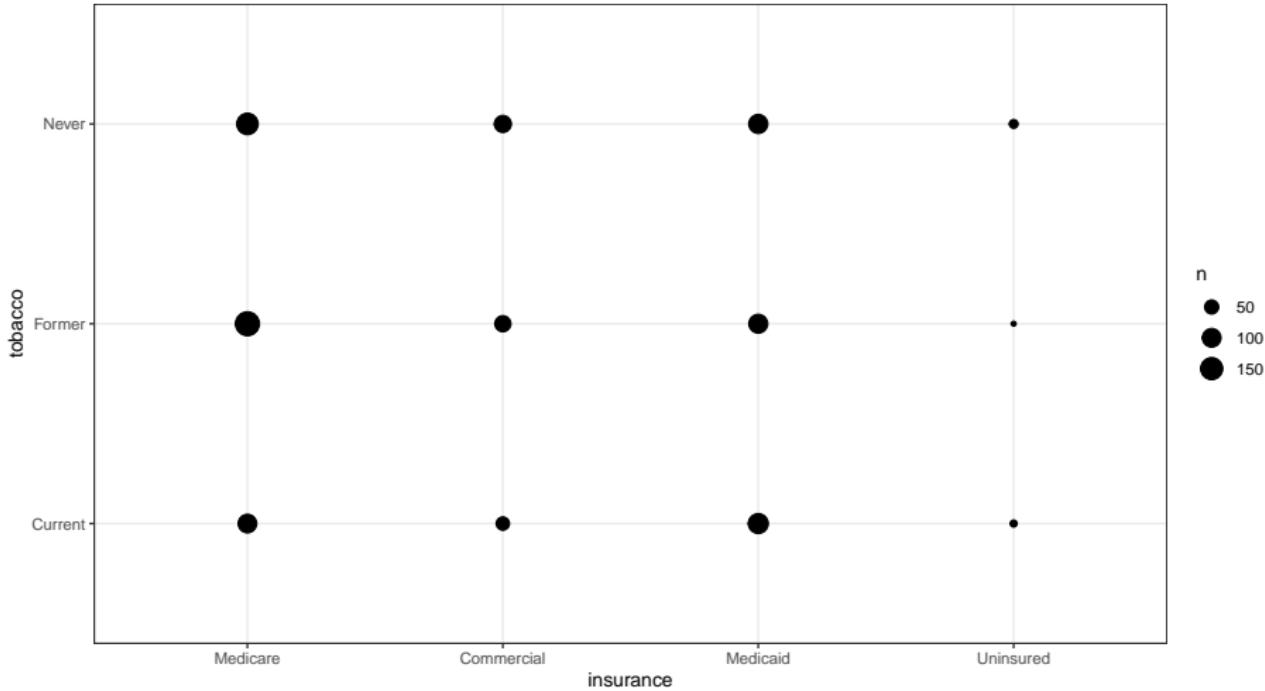
A 4×3 table with the dm1000 data

```
dm1000 %>%
  tabyl(insurance, tobacco) %>%
  adorn_totals(where = c("row", "col"))
```

insurance	Current	Former	Never	Total
Medicare	99	183	140	422
Commercial	44	70	80	194
Medicaid	118	103	105	326
Uninsured	13	11	18	42
Total	274	367	343	984

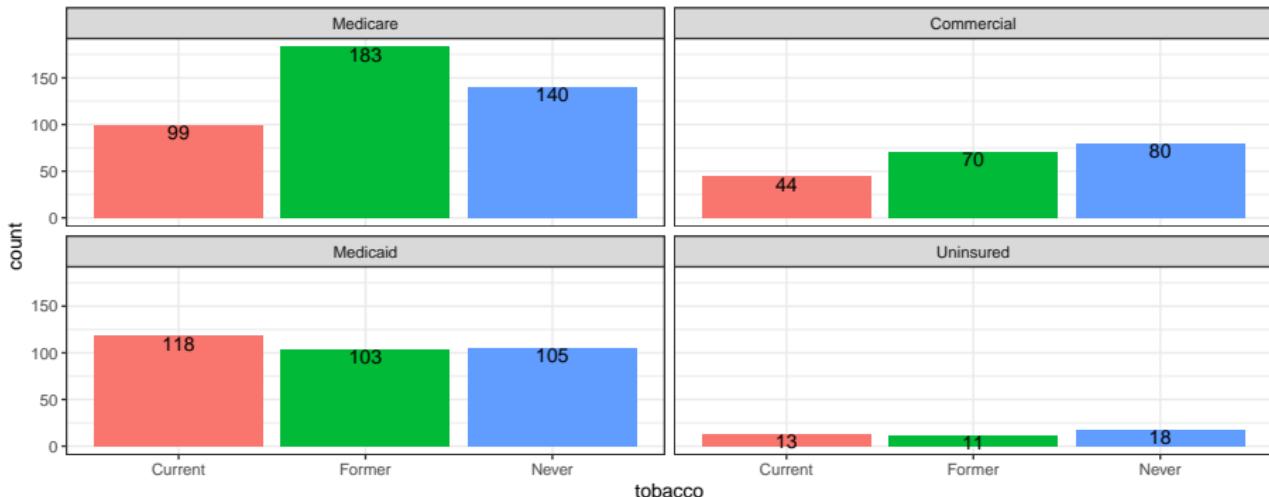
Plotting a Cross-Tabulation?

```
ggplot(dm1000, aes(x = insurance, y = tobacco)) +  
  geom_count() + theme_bw()
```



Tobacco Bar Chart faceted by Insurance

```
ggplot(dm1000, aes(x = tobacco, fill = tobacco)) +  
  geom_bar() + theme_bw() + facet_wrap(~ insurance) +  
  guides(fill = "none") +  
  geom_text(aes(label = ..count..), stat = "count",  
            vjust = 1, col = "black")
```



Tobacco Status and Insurance in dm1000

- H_0 : Insurance type and Tobacco status are independent
- H_A : Insurance type and Tobacco status have a detectable association

Pearson χ^2 results?

```
dm1000 %>% tabyl(insurance, tobacco) %>% chisq.test()
```

Pearson's Chi-squared test

```
data: .  
X-squared = 25.592, df = 6, p-value = 0.0002651
```

Can we check our expected frequencies?

Checking Expected Frequencies

```
res <- dm1000 %>% tabyl(insurance, tobacco) %>% chisq.test()
```

```
res$observed
```

insurance	Current	Former	Never
Medicare	99	183	140
Commercial	44	70	80
Medicaid	118	103	105
Uninsured	13	11	18

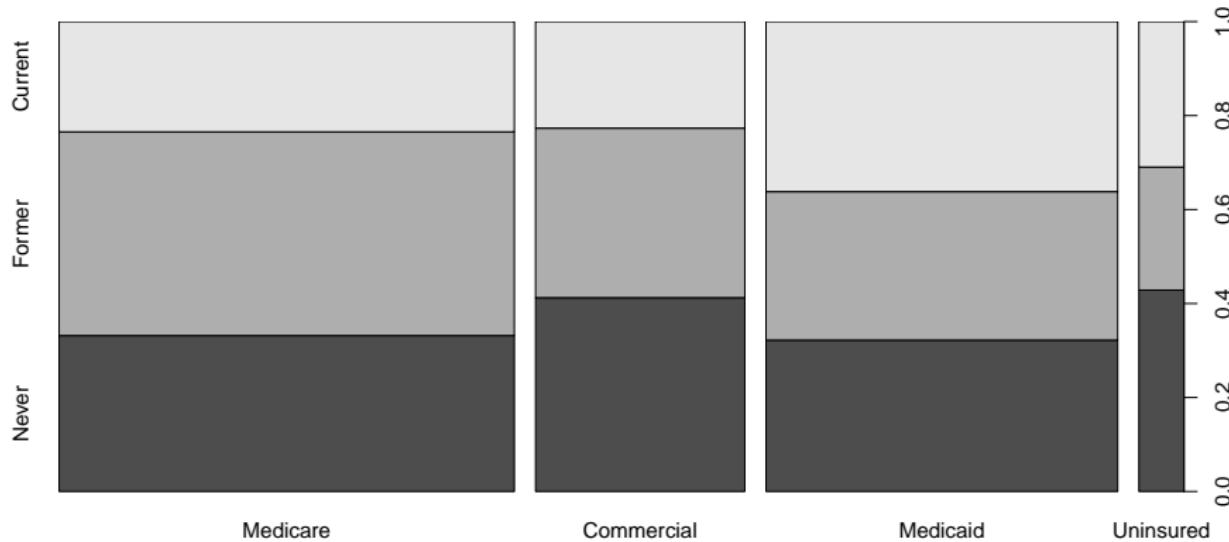
```
res$expected
```

insurance	Current	Former	Never
Medicare	117.50813	157.39228	147.09959
Commercial	54.02033	72.35569	67.62398
Medicaid	90.77642	121.58740	113.63618
Uninsured	11.69512	15.66463	14.64024

Mosaic Plot for Cross-Tabulation

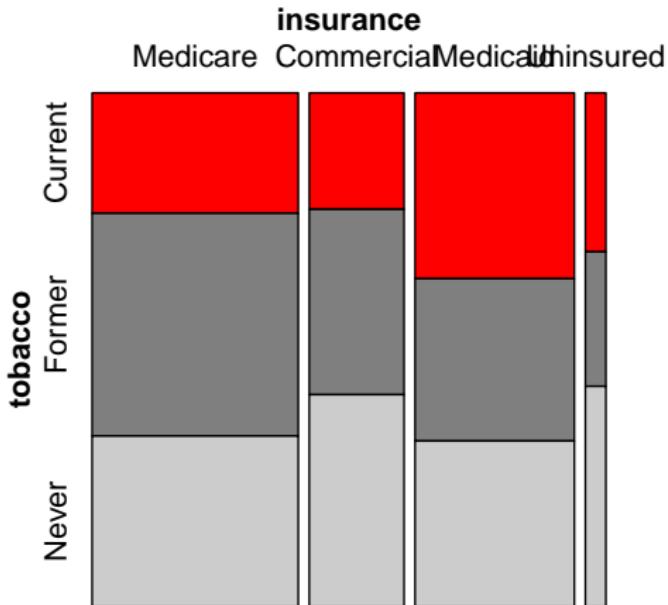
Each rectangle's area is proportional to the number of cases in that cell.

```
dm1000 %$% plot(insurance, tobacco, ylab = "", xlab = "")
```



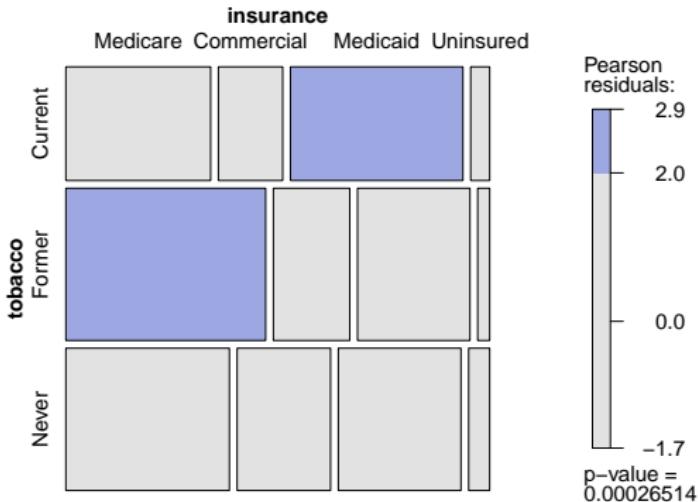
Mosaic Plot from the vcd package (highlighting)

```
mosaic(~ tobacco + insurance, data = dm1000,  
       highlighting = "tobacco",  
       highlighting_fill = c("red", "gray50", "gray80"))
```



Mosaic Plot from the vcd package (with χ^2 shading)

```
mosaic(~ tobacco + insurance, data = dm1000, shade = TRUE)
```



P values: What's the problem?

Replicable Research and the Crisis in Science

- ASA 2016 Statement on P values (Context, Process, Purpose)
- Is changing the p value cutoff the right strategy?
- Second-generation p values: A next step?
- ASA 2019 Statement on Statistical Inference in the 21st Century

JELLY BEANS
CAUSE ACNE!

SCIENTISTS!
INVESTIGATE!

BUT WE'RE
PLAYING
MINECRAFT!
... FINE.



WE FOUND NO
LINK BETWEEN
JELLY BEANS AND
ACNE ($P > 0.05$).



THAT SETTLES THAT.

I HEAR IT'S ONLY
A CERTAIN COLOR
THAT CAUSES IT.
SCIENTISTS!

BUT
MINECRAFT!



WE FOUND NO
LINK BETWEEN
PURPLE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BROWN JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
PINK JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BLUE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
TEAL JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
SALMON JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
RED JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
TURQUOISE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
MAGENTA JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
YELLOW JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
GREY JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
TAN JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
CYAN JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND A
LINK BETWEEN
GREEN JELLY
BEANS AND ACNE
($P < 0.05$).



WE FOUND NO
LINK BETWEEN
MAUVE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BEIGE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
LILAC JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BLACK JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
PEACH JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
ORANGE JELLY
BEANS AND ACNE
($P > 0.05$).



== News ==

GREEN JELLY BEANS LINKED TO ACNE!

95% CONFIDENCE

ONLY 5% CHANCE
OF COINCIDENCE!



SCIENTISTS...

Roger Peng's description of a successful data analysis

A data analysis is successful if the audience to which it is presented accepts the results.

- “What is a Successful Data Analysis?” simplystatistics.org (2018-04-17).

So what makes a data analysis more believable / more acceptable?

2016

- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

2019

- Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0.05$ ”, *The American Statistician*, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913.

Statistical Inference in the 21st Century

... a world learning to venture beyond “ $p < 0.05$ ”

This is a world where researchers are free to treat “ $p = 0.051$ ” and “ $p = 0.049$ ” as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number.

In this world, where studies with “ $p < 0.05$ ” and studies with “ $p > 0.05$ ” are not automatically in conflict, researchers will see their results more easily replicated – and, even when not, they will better understand why.

The 2016 ASA Statement on P-Values and Statistical Significance started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times—an average of about 11 citations per week since its release. Now we must go further.

The American Statistical Association Statement on P values and Statistical Significance

The ASA Statement (2016) was mostly about what **not** to do.

The 2019 effort represents an attempt to explain what to do.

Some of you exploring this special issue of The American Statistician might be wondering if it's a scolding from pedantic statisticians lecturing you about what not to do with p-values, without offering any real ideas of what to do about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

“Don’t” is not enough.

If you’re just arriving to the debate, here’s a sampling of what not to do.

- Don’t base your conclusions solely on whether an association or effect was found to be “statistically significant” (i.e., the p value passed some arbitrary threshold such as $p < 0.05$).
- Don’t believe that an association or effect exists just because it was statistically significant.
- Don’t believe that an association or effect is absent just because it was not statistically significant.
- Don’t believe that your p -value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don’t conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

One More Don't...

The ASA *Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of “statistical significance” be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term “statistically significant” entirely. Nor should variants such as “significantly different,” “ $p < 0.05$,” and “nonsignificant” survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of “statistical significance” has today become meaningless. Made

A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.

Problems with *P* Values

- ① *P* values are inherently unstable
- ② The *p* value, or statistical significance, does not measure the size of an effect or the importance of a result
- ③ Scientific conclusions should not be based only on whether a *p* value passes a specific threshold
- ④ Proper inference requires full reporting and transparency
- ⑤ By itself, a *p* value does not provide a good measure of evidence regarding a model or hypothesis

[Link](#)

Solutions to the P Value Problems

- ① Estimation of the Size of the Effect
- ② Precision of the Estimate (Confidence Intervals)
- ③ Inference About the Target Population
- ④ Determination of Whether the Results Are Compatible With a Clinically Meaningful Effect
- ⑤ Replication and Steady Accumulation of Knowledge

Link

Importance of Meta-Analytic Thinking

In JAMA Otolaryngology: Head & Neck Surgery, we look to publish original investigations where the investigators planned the study with sufficient sample size to have adequate power to detect a clinically meaningful effect and report the results with effect sizes and CIs. Authors should interpret the effect sizes in relation to previous research and use CIs to help determine whether the results are compatible with clinically meaningful effects. And finally, we acknowledge that no single study can define truth and that the advancement of medical knowledge and patient care depends on the steady accumulation of reliable clinical information.

[Link](#)

The Value of a *p*-Valueless Paper

Jason T. Connor (2004) *American J of Gastroenterology* 99(9): 1638-40.

Abstract: As is common in current bio-medical research, about 85% of original contributions in *The American Journal of Gastroenterology* in 2004 have reported *p*-values. However, none are reported in this issue's article by Abraham et al. who, instead, rely exclusively on effect size estimates and associated confidence intervals to summarize their findings. **Authors using confidence intervals communicate much more information in a clear and efficient manner than those using *p*-values. This strategy also prevents readers from drawing erroneous conclusions caused by common misunderstandings about *p*-values.** I outline how standard, two-sided confidence intervals can be used to measure whether two treatments differ or test whether they are clinically equivalent.

[Link](#)

Editor's Note

Do Not Over (*P*) Value Your Research Article

Laine E. Thomas, PhD; Michael J. Pencina, PhD

P value is by far the most prevalent statistic in the medical literature but also one attracting considerable controversy. Recently, the American Statistical Association¹ released a policy statement on *P* values, noting that misunderstanding and

misuse of *P* values is an important contributing factor to the common problem of scientific conclusions that fail to

be reproducible. Furthermore, reliance on *P* values may distract from the good scientific principles that are needed for high-quality research. Mark et al² delve deeper into the history and interpretation of the *P* value in this issue of *JAMA Cardiology*. Herein, we take the opportunity to state a few principles to help guide authors in the use and reporting of *P* values in the journal.

When the limitations surrounding *P* values are emphasized, a common question is, "What should we do instead?" Ron Wasserstein of the American Statistical Association explained: "In the post $p < 0.05$ era, scientific argumentation is not based on whether a *p*-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy.... Instead, journals [should evaluate] papers based on clear and detailed description of the study design, execution, and analysis, having conclusions that are based on valid

We suggest that researchers submitting manuscripts to *JAMA Cardiology* should also consider the following:

1. Data that are descriptive of the sample (ie, indicating imbalances between observed groups but not making inference to a population) should not be associated with *P* values. Appropriate language, in this case, would describe numerical differences and sample summary statistics and focus on differences of clinical importance.
2. In addition to summary statistics and confidence intervals, standardized differences (rather than *P* values) are a preferred way to exhibit imbalances between groups.
3. *P* values are most meaningful in the context of clear, *a priori* hypotheses that support the main conclusions of a manuscript.
4. Reporting stand-alone *P* values is discouraged, and preference should be given to presentation and interpretation of effect sizes and their uncertainty (confidence intervals) in the scientific context and in light of other evidence. Crossing a threshold (eg, $P < .05$) by itself constitutes only weak evidence.
5. Researchers should define and interpret effect measures that are clinically relevant. For example, clinical importance is often difficult to establish on the odds ratio scale but is clearer on the risk ratio or absolute risk difference scale.

In summary, following Mark et al,² we encourage research-



Related article

Abstract

P values and hypothesis testing methods are frequently misused in clinical research. Much of this misuse appears to be owing to the widespread, mistaken belief that they provide simple, reliable, and objective triage tools for separating the true and important from the untrue or unimportant. The primary focus in interpreting therapeutic clinical research data should be on the treatment ("oomph") effect, a metaphorical force that moves patients given an effective treatment to a different clinical state relative to their control counterparts. This effect is assessed using 2 complementary types of statistical measures calculated from the data, namely, effect magnitude or size and precision of the effect size. In a randomized trial, effect size is often summarized using constructs, such as odds ratios, hazard ratios, relative risks, or adverse event rate differences. How large a treatment effect has to be to be consequential is a matter for clinical judgment. The precision of the effect size (conceptually related to the amount of spread in the data) is usually addressed with confidence intervals. *P* values (significance tests) were first proposed as an informal heuristic to help assess how "unexpected" the observed effect size was if the true state of nature was no effect or no difference. Hypothesis testing was a modification of the significance test approach that envisioned controlling the false-positive rate of study results over many (hypothetical) repetitions of the experiment of interest. Both can be helpful but, by themselves, provide only a tunnel vision perspective on study results that ignores the clinical effects the study was conducted to measure.

Link

Dividing Data Comparisons into Categories based on p values

Regina Nuzzo in Nature on Statistical Errors

PROBABLE CAUSE

A P value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausible the hypothesis is in the first place.

- Chance of real effect
- Chance of no real effect

Before the experiment

The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

The measured P value

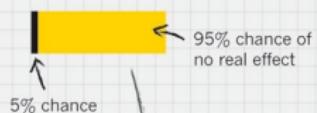
A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

After the experiment

A small P value can make a hypothesis more plausible, but the difference may not be dramatic.

THE LONG SHOT

19-to-1 odds against



THE TOSS-UP

1-to-1 odds



THE GOOD BET

9-to-1 odds in favour



$P = 0.05$

$P = 0.01$

$P = 0.05$

$P = 0.01$

$P = 0.05$

$P = 0.01$

11% chance of real effect

89% chance of no real effect

30%

70%

71%

29%

89%

11%

96%

4%

99%

1%

Gelman on p values, 1

The common practice of dividing data comparisons into categories based on significance levels is terrible, but it happens all the time. . . so it's worth examining the prevalence of this error.

Consider, for example, this division:

- “really significant” for $p < .01$,
- “significant” for $p < .05$,
- “marginally significant” for $p < .1$, and
- “not at all significant” otherwise.

Now consider some typical p -values in these ranges: say, $p = .005$, $p = .03$, $p = .08$, and $p = .2$.

Translate these two-sided p -values back into z-scores. . .

Gelman 2016-10-15

Gelman on p values, 2

Description	really sig.	sig.	marginally sig.	not at all sig.
p value	0.005	0.03	0.08	0.20
Z score	2.8	2.2	1.8	1.3

The seemingly yawning gap in p -values comparing the not at all significant p -value of .2 to the really significant p -value of .005, is only a z score of 1.5.

If you had two independent experiments with z-scores of 2.8 and 1.3 and with equal standard errors and you wanted to compare them, you'd get a difference of 1.5 with a standard error of 1.4, which is completely consistent with noise.

Gelman on *p* values, 3

From a **statistical** point of view, the trouble with using the p-value as a data summary is that the p-value can only be interpreted in the context of the null hypothesis of zero effect, and (much of the time), nobody's interested in the null hypothesis.

Indeed, once you see comparisons between large, marginal, and small effects, the null hypothesis is irrelevant, as you want to be comparing effect sizes.

From a **psychological** point of view, the trouble with using the p-value as a data summary is that this is a kind of deterministic thinking, an attempt to convert real uncertainty into firm statements that are just not possible (or, as we would say now, just not replicable).

The key point: The difference between statistically significant and NOT statistically significant is not, generally, statistically significant.

Are P values all that bad?



Grumpy Old Health Stats Dude

@healthstatsdude

Following



"If you never use another p-value, you will have improved medicine."

-me, to clinicians

#statstwitter #medtwitter #epitwitter

12:36 AM - 4 Mar 2019



Replies to @healthstatsdude @EugeneDayDSc and 2 others

My main reason for being overtly/in public anti p-values is this:

P values
of overall
analyses

partly
statistical
even if
group,
ever, due
al distri-
fference
specific
all death

mate of a 5% decrease in 10-year survival with
watchful waiting, 750 men might have died
prematurely as a result.

A mistake in the operating room can threaten
the life of one patient; a mistake in statistical
analysis or interpretation can lead to hundreds
of early deaths. So it is perhaps odd that, while we
allow a doctor to conduct surgery only after years
of training, we give SPSS® (SPSS, Chicago, IL) to
almost anyone. Moreover, whilst only a surgeon
would comment on surgical technique, it seems
that anybody, regardless of statistical training,

day); a
that on
risk of t
in many
that the
event is

Comp
The aut
no com

7:59 PM - 19 Apr 2019

Where to Go from Here?

- ① Be the change you want to see in the world.
- ② Frank Harrell's "A Litany of Problems with p-values" blog post
- ③ William Briggs' "Everything Wrong with P-values under One Roof" article.

These resources are linked on our Class 23 README.

431 Class 24

thomaselove.github.io/431

2021-11-18

Today's Agenda

- What I Taught for Many Years
- p-Hacking
- Do Confidence Intervals Solve the Problem?
- Borrowing from Bayesian Ideas
- Replicable Research and the Crisis in Science
- Retrospective Power and why most smart folks avoid it
 - Type S and Type M error: Saying something more useful

What I Taught for Many Years

- Null hypothesis significance testing is here to stay.
 - Learn how to present your p value so it looks like what everyone else does
 - Think about “statistically detectable” rather than “statistically significant”
 - Don’t accept a null hypothesis, just retain it.
- Use point **and** interval estimates
 - Try to get your statements about confidence intervals right (right = just like I said it)
- Use Bayesian approaches/simulation/hierarchical models when they seem appropriate or for “non-standard” designs
 - But look elsewhere for people to teach/do that stuff
- Power is basically a hurdle to overcome in a grant application

Conventions for Reporting *p* Values

- ① Use an italicized, lower-case *p* to specify the *p* value. Don't use *p* for anything else.
- ② For *p* values above 0.10, round to two decimal places, at most.
- ③ For *p* values near α , include only enough decimal places to clarify the reject/retain decision.
- ④ For very small *p* values, always report either $p < 0.0001$ or even just $p < 0.001$, rather than specifying the result in scientific notation, or, worse, as $p = 0$ which is glaringly inappropriate.
- ⑤ Report *p* values above 0.99 as $p > 0.99$, rather than $p = 1$.

American Statistical Association to the rescue!?

ASA Statement on *p* Values

ASA Statement: “Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.”

fivethirtyeight.com “Not Even Scientists Can Easily Explain *p* Values”

... Try to distill the p-value down to an intuitive concept and it loses all its nuances and complexity, said science journalist Regina Nuzzo, a statistics professor at Gallaudet University. “Then people get it wrong, and this is why statisticians are upset and scientists are confused.” **You can get it right, or you can make it intuitive, but it’s all but impossible to do both.**

fivethirtyeight.com “Statisticians found one thing they can agree on”

A Few Comments on Significance

- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always “significant” even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?
- **A non-significant effect is not necessarily the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.
- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.
- **“Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.”**

ASA *statement* on *p* values

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a *p*-value as one possible summary of evidence morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$,

which morphed into a

- **rule** for journals: reject all articles that report *p*-values¹

¹<http://www.nature.com/news/psychology-journal-bans-p-values-1.17001> describes the recent banning of null hypothesis significance testing by *Basic and Applied Psychology*.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$, which morphed into a
- **rule** for editors: reject the submitted article if $p > .05$, which morphed into a
- **rule** for journals: reject all articles that report p-values.

Bottom line: **Reject rules. Ideas matter.**

$P > 0.05$



GAME OVER, TRY AGAIN

imgflip.com

$p = 0.05?$

"For decades, the conventional p-value threshold has been 0.05," says Dr. Paul Wakim, chief of the biostatistics and clinical epidemiology service at the National Institutes of Health Clinical Center, "but it is extremely important to understand that this 0.05, there's nothing rigorous about it. It wasn't derived from statisticians who got together, calculated the best threshold, and then found that it is 0.05. No, it's Ronald Fisher, who basically said, 'Let's use 0.05,' and he admitted that it was arbitrary."

- NOVA “[Rethinking Science’s Magic Number](#)” by Tiffany Dill
2018-02-28. See especially the video labeled “Science’s most important (and controversial) number has its origins in a British experiment involving milk and tea.”

More from Dr. Wakim...

"People say, 'Ugh, it's above 0.05, I wasted my time.' No, you didn't waste your time." says Dr. Wakim. "If the research question is important, the result is important. Whatever it is."

- NOVA Season 45 Episode 6 [Prediction by the Numbers](#) 2018-02-28.

p values don't trend...



Randy Sweis, MD

@RandySweisMD

Follow



If a P value of 0.06 trends toward statistical significance, then doesn't a P value of 0.04 trend toward non-significance?

9:47 AM - 12 Jan 2018

George Cobb's Questions (with Answers)

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?

A: Because that's **still** what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

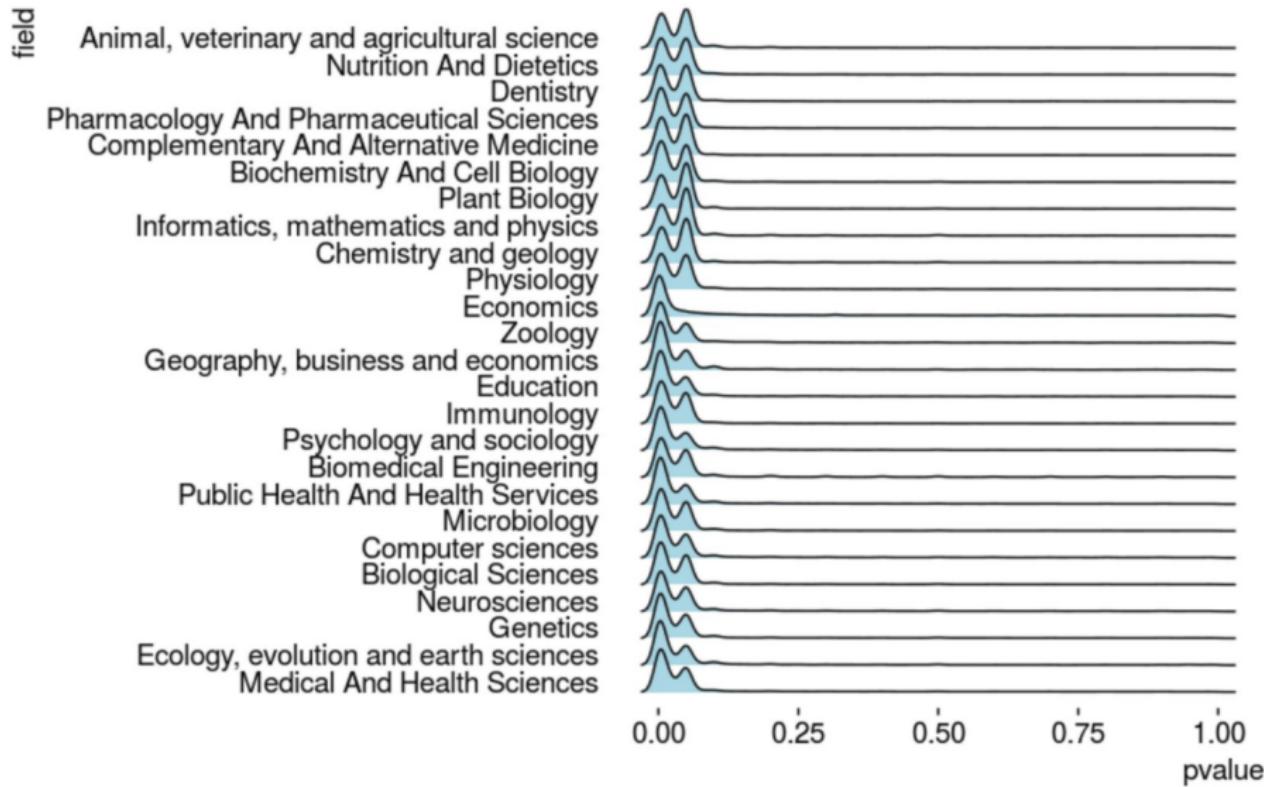
All the p values

The p-value is the most widely-known statistic. P-values are reported in a large majority of scientific publications that measure and report data. R.A. Fisher is widely credited with inventing the p-value. If he was cited every time a p-value was reported his paper would have, at the very least, 3 million citations - making it the most highly cited paper of all time.

- Visit Jeff Leek's [Github for tidypvals package](#)
 - 2.5 million p values in 25 scientific fields

What do you suppose the distribution of those p values is going to look like?

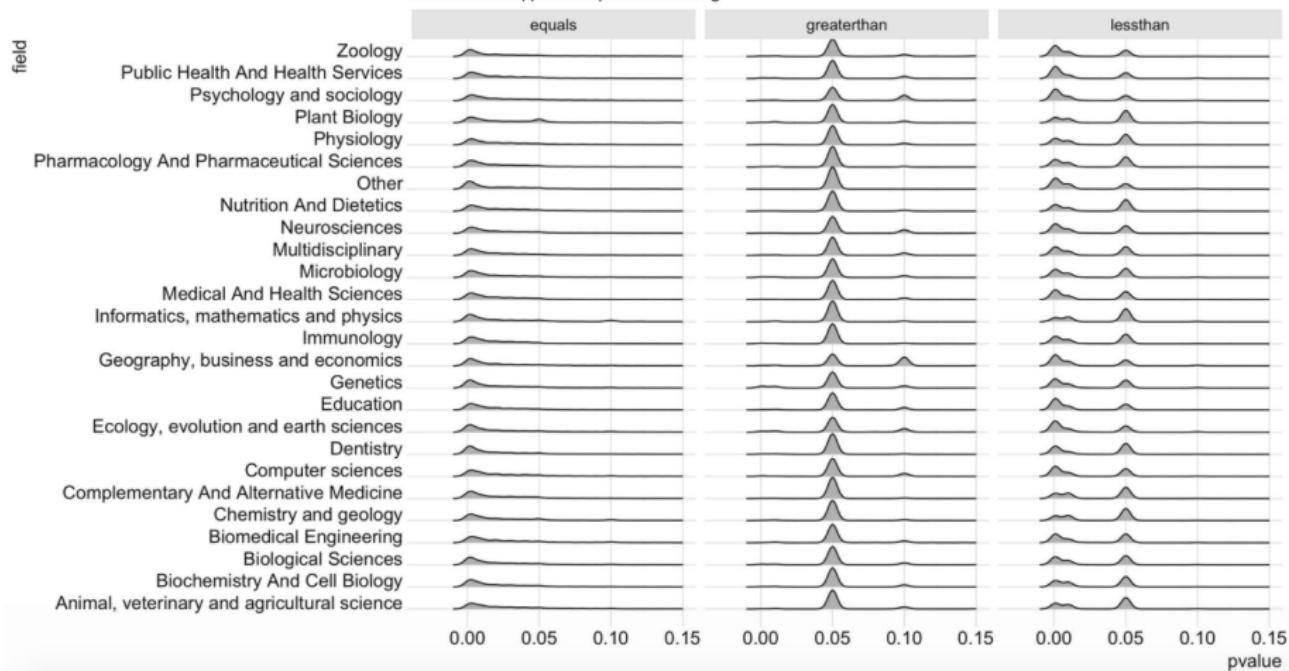
2.5 million p values in 25 scientific fields: Jeff Leek



from Michael Lopez

Distribution of pvalues by operator (=, >, <)

Economics dropped: all operators missing



Simple way for editors to improve science: If your journal still uses “statistical significance” in 2017, retire your statistical consultant

Practices that reduce scientific inference to mechanical “bright-line” rules (such as “ $p < 0.05$ ”) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making.

American Statistical Association, 2016

But many journals do present findings as “statistically significant” or “not statistically significant”.

- How can an editor work with statistical consultants who ignore the ASA without publicly justifying their views?
- Would the editor work with a cardiology consultant who ignores the American Heart Association without providing any justification?

Unfortunately...

There are a lot of candidates for the most outrageous misuse of “statistical significance” out there.



Alvaro Alonso
@alonso_epi

...

More p-value silliness. HR 0.90, 95%CI 0.81-0.99-->
'effect'; HR 0.89, 95%CI 0.78-1.0009-->no 'effect'
jaha.ahajournals.org/content/6/5/e0... @ken_rothman

Normalization of Testosterone Levels After Testosterone Replacement Therapy Is Associated With Decreased Incidence of Atrial Fibrillation

Rishi Sharma, MD, MHSa; Olurinde A. Oni, MBBS, MPH; Kamal Gupta, MD; Mukut Sharma, PhD; Ram Sharma, PhD; Vikas Singh, MD, MHSa; Deepak Parashara, MD; Surineni Kamalakar, MBBS, MPH; Buddhadeb Dawn, MD; Guoqing Chen, MD, PhD, MPH; John A. Ambrose, MD; Rajat S. Barua, MD, PhD

Background—Atrial fibrillation (AF) is the most common cardiac dysrhythmia associated with significant morbidity and mortality. Several small studies have reported that low serum total testosterone (TT) levels were associated with a higher incidence of AF. In contrast, it is also reported that anabolic steroid use is associated with an increase in the risk of AF. To date, no study has explored the effect of testosterone normalization on new incidence of AF after testosterone replacement therapy (TRT) in patients with low testosterone.

Methods and Results—Using data from the Veterans Administrations Corporate Data Warehouse, we identified a national cohort of 76 639 veterans with low TT levels and divided them into 3 groups. Group 1 had TRT resulting in normalization of TT levels (normalized TRT), group 2 had TRT without normalization of TT levels (nonnormalized TRT), and group 3 did not receive TRT (no TRT). Propensity score-weighted stabilized inverse probability of treatment weighting Cox proportional hazard methods were used for analysis of the data from these groups to determine the association between post-TRT levels of TT and the incidence of AF. **Group 1 (40 856 patients, median age 66 years) had significantly lower risk of AF than group 2 (23 939 patients, median age 65 years; hazard ratio 0.80, 95% CI 0.71-0.89, $P=0.0005$)** and group 3 (11 844 patients, median age 67 years; hazard ratio 0.79, 95% CI 0.69-0.89, $P<0.0001$).

Normalization of Testosterone Levels After Testosterone Replacement Therapy Is Associated With Decreased Incidence of Atrial Fibrillation

Rishi Sharma, MD, MHSA; Olurinde A. Oni, MBBS, MPH; Kamal Gupta, MD; Mukut Sharma, PhD; Ram Sharma, PhD; Vikas Singh, MD, MHSA; Deepak Parashara, MD; Surineni Kamalakar, MBBS, MPH; Buddhadeb Dawn, MD; Guoqing Chen, MD, PhD, MPH; John A. Ambrose, MD; Rajat S. Barua, MD, PhD

Background—Atrial fibrillation (AF) is the most common cardiac dysrhythmia associated with significant morbidity and mortality. Several small studies have reported that low serum total testosterone (TT) levels were associated with a higher incidence of AF. In contrast, it is also reported that anabolic steroid use is associated with an increase in the risk of AF. To date, no study has explored the effect of testosterone normalization on new incidence of AF after testosterone replacement therapy (TRT) in patients with low testosterone.

Methods and Results—Using data from the Veterans Administrations Corporate Data Warehouse, we identified a national cohort of 76 639 veterans with low TT levels and divided them into 3 groups. Group 1 had TRT resulting in normalization of TT levels (normalized TRT), group 2 had TRT without normalization of TT levels (nonnormalized TRT), and group 3 did not receive TRT (no TRT). Propensity score–weighted stabilized inverse probability of treatment weighting Cox proportional hazard methods were used for analysis of the data from these groups to determine the association between post-TRT levels of TT and the incidence of AF. Group 1 (40 856 patients, median age 66 years) had significantly lower risk of AF than group 2 (23 939 patients, median age 65 years; hazard ratio 0.90, 95% CI 0.81–0.99, $P=0.0255$) and group 3 (11 853 patients, median age 67 years; hazard ratio 0.79, 95% CI 0.70–0.89, $P=0.0001$). There was no statistical difference between groups 2 and 3 (hazard ratio 0.89, 95% CI 0.78–1.0009, $P=0.0675$) in incidence of AF.

Conclusions—These novel results suggest that normalization of TT levels after TRT is associated with a significant decrease in the incidence of AF. (*J Am Heart Assoc.* 2017;6:e004880. DOI: 10.1161/JAHA.116.004880.)

Key Words: atrial fibrillation • testosterone • testosterone replacement therapy



Mike Babyak
@mababyak

...

Replying to @_MiguelHernan

I've often tried to make a similar point to colleagues. They would never dream of ignoring medical consensus on the approach to an assay or dx procedure, but often cast statisticians as being "fussy" for trying to have them adhere to best statistical practice.

10:06 AM · Dec 31, 2017 · Twitter Web Client



Ken Rothman
@ken_rothman

...

Replying to @oncology_bg @pash22 and 8 others

.We shouldn't be "deciding" to reject or accept. We should be measuring effects. See, e.g.,

p-Hacking

Hack Your Way To Scientific Glory (fivethirtyeight)

Hack Your Way To Scientific Glory



You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

- Employment
- Inflation
- GDP
- Stock prices

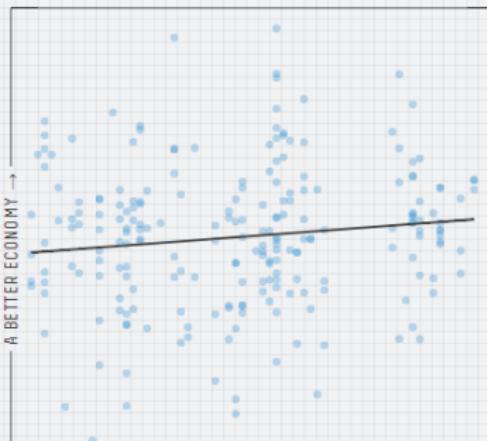
Other options

- Factor in power

Weight more powerful
positions more heavily

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



Result: Almost

Your 0.06 p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

“Researcher Degrees of Freedom”, 1

[I]t is unacceptably easy to publish statistically significant evidence consistent with any hypothesis.

*The culprit is a construct we refer to as **researcher degrees of freedom**. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?*

Simmons et al. [link](#)

“Researcher Degrees of Freedom”, 2

... It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields statistical significance, and to then report only what worked. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

For more, see

- Gelman's blog [2012 – 11 – 01](#) “Researcher Degrees of Freedom”,
- Paper by [*Simmons*](#) and others, defining the term.

And this is really hard to deal with...

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or p-hacking and the research hypothesis was posited ahead of time

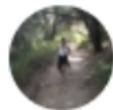
Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

- [Link](#) to the paper from Gelman and Loken

Grim Reality

- Editorial, Educational and Other Institutional Practices Will Have to Change
- It Is Going to Take Work, and It Is Going to Take Time
- Why Will Change Finally Happen Now?

Confidence Intervals - do they solve our problem?



Chelsea Parlett Pelleriti

@ChelseaParlett

Follow



Hey Stats folk, what's your 280 character definition of a confidence interval? 🤔

4:30 PM - 13 Mar 2018

Confidence Intervals - do they solve our problem?



Thomas Leeper
@thosleeper

[Follow](#)



Replies to @ChelseaParlett

An interval drawn such that, were repeated, equal-sized samples of units drawn from the population of units using an identical sampling procedure and the same estimator was applied to each sample, $100*(1-\alpha)\%$ of those intervals would contain the population parameter of interest.

4:58 PM - 13 Mar 2018

Confidence Intervals - do they solve our problem?



Joran Elias
@joranelias

Follow



A confidence interval is a measure of uncertainty such that all definitions of it elicit corrections from Bayesians.

(Didn't need all 280.)

Confidence Intervals - do they solve our problem?



Jenny Bryan
@JennyBryan

Following



Pedantry about the definition of a confidence interval ... why is this the hill statisticians choose to die on? Every time you feel the urge, go convert a table to a figure. It is likely to do more good.

Confidence Intervals - do they solve our problem?



Frank Harrell @f2harrell · 28 Dec 2017

Tables and figures are important but so is this. We need to get this right. Too many faulty conclusions being drawn with frequentist statistical analysis. If one is going to be a frequentist one should make exactly correct interpretations.

Q 2

⤵

Heart 10

✉

▼



Jenny Bryan

@JennyBryan

Following

Replying to @f2harrell

I just feel like the people we're often trying to reach aren't making informed comparisons of frequentist vs Bayesian methods, they're still struggling with decision making under uncertainty

Using Bayesian Ideas: Confidence Intervals

My current favorite (hypothetical) example is an epidemiology study of some small effect where the point estimate of the odds ratio is 3.0 with a 95% conf interval of [1.1, 8.2].

As a 95% conf interval, this is fine (assuming the underlying assumptions regarding sampling, causal identification, etc. are valid).

(but on some level you need to deal with the fact that...)

... real-world odds ratios are much more likely to be near 1.1 than to be near 8.2.

See [Gelman](#) 2014-12-11.

Uncertainty intervals?

I've (Gelman) become increasingly uncomfortable with the term "confidence interval" for several reasons:

- The well-known difficulties in interpretation (officially the confidence statement can be interpreted only on average, but people typically implicitly give the Bayesian interpretation to each case.)
- The ambiguity between confidence intervals and predictive intervals.
- The awkwardness of explaining that confidence intervals are big in noisy situations where you have less confidence, and confidence intervals are small when you have more confidence.

So here's my proposal. Let's use the term "uncertainty interval" instead. The uncertainty interval tells you how much uncertainty you have.

See [Gelman](#) 2010-12-21.

Some Noisy Recent Suggestions

Benjamin et al 2017 Redefine Statistical Significance

We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

Motivations:

- links to Bayes Factor interpretation
- 0.005 is stringent enough to “break” the current system - makes it very difficult for researchers to reach threshold with noisy, useless studies.

Visit the main [*article*](#). Visit an explanatory piece in [*Science*](#).

Lakens et al. Justify Your Alpha

"In response to recommendations to redefine statistical significance to $p \leq .005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level." Visit [link](#).

Abandon Statistical Significance

Gelman blog [2017 – 09 – 26](#) on “Abandon Statistical Significance”

“Measurement error and variation are concerns even if your estimate is more than 2 standard errors from zero. Indeed, if variation or measurement error are high, then you learn almost nothing from an estimate even if it happens to be ‘statistically significant.’ ”

Read the whole paper [here](#)

VIEWPOINT

John P. A. Ioannidis,
MD, DSc
Stanford Prevention
Research Center,
Meta-Research
Innovation Center at
Stanford, Departments
of Medicine, Health
Research and Policy,
Biomedical Data
Science, and Statistics,
Stanford University,
Stanford, California.

The Proposal to Lower *P* Value Thresholds to .005

Pvalues and accompanying methods of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report *P* values in the abstract, full text, or both include some values of .05 or less.¹ However, many of the claims that these reports highlight are likely false.² Recognizing the major importance of the statistical significance conundrum, the American Statistical Association (ASA) published³ a statement on *P* values in 2016. The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious. The contributors to the ASA statement also wrote 20 independent, accompanying commentaries focusing on different aspects and prioritizing different solutions. Another large coalition of 72 methodologists recently proposed⁴ a specific, simple move: lowering the routine *P* value threshold for claiming statistical significance from .05 to .005 for new discoveries. The proposal met with strong endorsement in some circles and concerns in others.

P values are misinterpreted, overtrusted, and misused. The language of the ASA statement enables the dis-

fully considered how low a *P* value should be for a research finding to have a sufficiently high chance of being true. For example, adoption of genome-wide significance thresholds ($P < 5 \times 10^{-8}$) in population genomics has made discovered associations highly replicable and these associations also appear consistently when tested in new populations. The human genome is very complex, but the extent of multiplicity of significance testing involved is known, the analyses are systematic and transparent, and a requirement for $P < 5 \times 10^{-8}$ can be cogently arrived at.

However, for most other types of biomedical research, the multiplicity involved is unclear and the analyses are nonsystematic and nontransparent. For most observational exploratory research that lacks preregistered protocols and analysis plans, it is unclear how many analyses were performed and what various analytic paths were explored. Hidden multiplicity, nonsystematic exploration, and selective reporting may affect even experimental research and randomized trials. Even though it is now more common to have a preexisting protocol and statistical analysis plan and preregistration of

RESEARCH ARTICLE

Second-generation *p*-values: Improved rigor, reproducibility, & transparency in statistical analyses

Jeffrey D. Blume^{1*}, Lucy D'Agostino McGowan², William D. Dupont³, Robert A. Greevy,
Jr.¹

Second-generation p values

Verifying that a statistically significant result is scientifically meaningful is not only good scientific practice, it is a natural way to control the Type I error rate. Here we introduce a novel extension of the p -value—a second-generation p -value (p_δ)—that formally accounts for scientific relevance and leverages this natural Type I Error control. The approach relies on a pre-specified interval null hypothesis that represents the collection of effect sizes that are scientifically uninteresting or are practically null. The second-generation p -value is the proportion of data-supported hypotheses that are also null hypotheses. As such, second-generation p -values indicate when the data are compatible with null hypotheses ($p_\delta = 1$), or with alternative hypotheses ($p_\delta = 0$), or when the data are inconclusive ($0 < p_\delta < 1$). Moreover, second-generation p -values provide a proper scientific adjustment for multiple comparisons and reduce false discovery rates. This is an advance for environments rich in data, where traditional p -value adjustments are needlessly punitive. Second-generation p -values promote transparency, rigor and reproducibility of scientific results by *a priori* specifying which candidate hypotheses are practically meaningful and by providing a more reliable statistical summary of when the data are compatible with alternative or null hypotheses.

COMMENT

P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

OK, so what **SHOULD** we do?

The American Statistician Volume 73, 2019, Supplement 1

Articles on:

- ① Getting to a Post “ $p < 0.05$ ” Era
 - ② Interpreting and Using p
 - ③ Supplementing or Replacing p
 - ④ Adopting more holistic approaches
 - ⑤ Reforming Institutions: Changing Publication Policies and Statistical Education
-
- Note that there is an enormous list of “things to do” in Section 7 of the main editorial, too.

Statistical Inference in the 21st Century



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: <https://www.tandfonline.com/loi/utas20>

Moving to a World Beyond “ $p < 0.05$ ”

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019)
Moving to a World Beyond “ $p < 0.05$ ”, *The American Statistician*, 73:sup1, 1-19, DOI:
[10.1080/00031305.2019.1583913](https://doi.org/10.1080/00031305.2019.1583913)

To link to this article: <https://doi.org/10.1080/00031305.2019.1583913>

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

- Statistical methods do not rid data of their uncertainty.

Statistical methods do not rid data of their uncertainty. “Statistics,” Gelman (2016) says, “is often sold as a sort of alchemy that transmutes randomness into certainty, an ‘uncertainty laundering’ that begins with data and concludes with success as measured by statistical significance.” To accept uncertainty requires that we “treat statistical results as being much more incomplete and uncertain than is currently the norm” (Amrhein, Trafimow, and Greenland 2019). We must “countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error” (Calin-Jageman and Cumming 2019).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

We can make acceptance of uncertainty more natural to our thinking by accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate. Reporting and interpreting point and interval estimates should be routine.

How will accepting uncertainty change anything? To begin, it will prompt us to seek better measures, more sensitive designs, and larger samples, all of which increase the rigor of research.

It also helps us be modest . . . [and] leads us to be thoughtful.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

3.2. Be Thoughtful

What do we mean by this exhortation to “be thoughtful”? Researchers already clearly put much thought into their work. We are not accusing anyone of laziness. Rather, we are envisioning a sort of “statistical thoughtfulness.” In this perspective, statistically **thoughtful researchers** begin above all else with clearly expressed objectives. They recognize when they are doing exploratory studies and when they are doing more rigidly pre-planned studies. They invest in producing solid data. They consider not one but a multitude of data analysis techniques. And they think about so much more.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research looks ahead to prospective outcomes in the context of theory and previous research. Researchers would do well to ask, *What do we already know, and how certain are we in what we know?* And building on that and on the field's theory, *what magnitudes of differences, odds ratios, or other effect sizes are practically important?* These questions would naturally lead a researcher, for example, to use existing evidence from a literature review to identify specifically the findings that would be practically important for the key outcomes under study.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research includes careful consideration of the definition of a meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed. Afterwards is just too late; it is dangerously easy to justify observed results after the fact and to overinterpret trivial effect sizes as being meaningful. Many authors in this special issue argue that consideration of the effect size and its “scientific meaningfulness” is essential for reliable inference (e.g., Blume et al. 2019; Betensky 2019). This concern is also addressed in the literature on equivalence testing (Wellek 2017).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research considers “related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain...without giving priority to *p*-values or other purely statistical measures” (McShane et al. 2019).

Thoughtful researchers “use a toolbox of statistical techniques, employ good judgment, and keep an eye on developments in statistical and data science,” conclude Heck and Krueger (2019), who demonstrate how the *p*-value can be useful to researchers as a heuristic.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

In all instances, regardless of the value taken by p or any other statistic, consider what McShane et al. (2019) call the “currently subordinate factors”—the factors that should no longer be subordinate to “ $p < 0.05$.” These include relevant prior evidence, plausibility of mechanism, study design and data quality, and the real-world costs and benefits that determine what effects are scientifically important. The scientific context of your study matters, they say, and this should guide your interpretation.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

To **be open**, remember that one study is rarely enough. The words “a groundbreaking new study” might be loved by news writers but must be resisted by researchers. Breaking ground is only the first step in building a house. It will be suitable for habitation only after much more hard work.

Be open by providing sufficient information so that other researchers can execute meaningful alternative analyses. van Dongen et al. (2019) provide an illustrative example of such alternative analyses by different groups attacking the same problem.

Being open goes hand in hand with **being modest**.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Being modest requires a reality check (Amrhein, Trafimow, and Greenland 2019). “A core problem,” they observe, “is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct and of the reality being studied.”

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Be modest in recognizing there is not a “true statistical model” underlying every problem, which is why it is wise to **thoughtfully** consider many possible models (Lavine 2019).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Be modest about the role of statistical inference in scientific inference. “Scientific inference is a far broader concept than statistical inference,” says Hubbard, Haig, and Parsa (2019). “A major focus of scientific inference can be viewed as the pursuit of *significant sameness*, meaning replicable and empirically generalizable results among phenomena. Regrettably, the obsession with users of statistical inference to report *significant differences* in data sets actively thwarts cumulative knowledge development.”

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

The nexus of openness and modesty is to report everything while at the same time not concluding anything from a single study with unwarranted certainty. Because of the strong desire to inform and be informed, there is a relentless demand to state results with certainty. Again, accept uncertainty and embrace variation in associations and effects, because they are always there, like it or not. Understand that expressions of uncertainty are themselves uncertain. Accept that one study is rarely definitive, so encourage, sponsor, conduct, and publish replication studies.

Be modest by encouraging others to reproduce your work. Of course, for it to be reproduced readily, you will necessarily have been thoughtful in conducting the research and open in presenting it.

What I Think I Think Now

- Null hypothesis significance testing is much harder than I thought.
 - The null hypothesis is almost never a real thing.
 - Rather than rejiggering the cutoff, I would largely abandon the p value as a summary
 - Replication is far more useful than I thought it was.
- Some hills aren't worth dying on.
 - Think about uncertainty intervals more than confidence or credible intervals
 - Retrospective calculations about Type S (sign) and Type M (magnitude) errors can help me illustrate ideas.
- Which method to use is far less important than finding better data
 - The biggest mistake I make regularly is throwing away useful data
 - I'm not the only one with this problem.
- The best thing I do most days is communicate more clearly.
 - When stuck in a design, I think about how to get better data.
 - When stuck in an analysis, I try to turn a table into a graph.
- I have A LOT to learn.