431 Class 16

https://thomaselove.github.io/431-2024/

2024-10-17

## Today’s Agenda

* Discussion of Quiz 1
* Dealing with Missing Data
* Simple (Single) Imputation using the mice package
* Multiple Imputation using the mice package

(MICE = Multiple Imputation through Chained Equations)

* Breakout Activity 3 Results

## Today’s Packages

library(janitor)  
library(googlesheets4)  
library(knitr)  
library(kableExtra)  
library(mice)  
library(naniar)  
library(xfun)  
library(easystats)  
library(tidyverse)  
  
theme\_set(theme\_light())  
  
source("c16/data/Love-431.R")

## 700 Veterans with HBP

700 male veterans with a hypertension diagnosis.

* Pulse Pressure = Systolic BP - Diastolic BP.

hbp700 <- read\_csv("c16/data/hbp700.csv", show\_col\_types = FALSE) |>  
 janitor::clean\_names() |>  
 mutate(across(where(is.character), as\_factor)) |>  
 mutate(pp = sbp - dbp) |>  
 mutate(subject = as.character(subject))   
  
dim(hbp700)

[1] 700 10

hbp700 |> head(4)

# A tibble: 4 × 10  
 subject age race nincome bmi tobacco ldl sbp dbp pp  
 <chr> <dbl> <fct> <dbl> <dbl> <fct> <dbl> <dbl> <dbl> <dbl>  
1 hbp001 65 NHBlack 27900 23.4 Never 104 150 87 63  
2 hbp002 64 NHWhite NA 35.5 Former 79 119 77 42  
3 hbp003 62 NHBlack 71600 NA Former 77 118 80 38  
4 hbp004 82 <NA> 37300 24.5 Former 63 113 52 61

## Missingness in hbp700

miss\_var\_summary(hbp700)

# A tibble: 10 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
 1 race 112 16   
 2 tobacco 110 15.7   
 3 bmi 19 2.71   
 4 nincome 11 1.57   
 5 ldl 7 1   
 6 sbp 1 0.143  
 7 dbp 1 0.143  
 8 pp 1 0.143  
 9 subject 0 0   
10 age 0 0

miss\_case\_table(hbp700)

# A tibble: 4 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 461 65.9   
2 1 219 31.3   
3 2 17 2.43   
4 3 3 0.429

## Categorical Variables in hbp700

hbp700 |> tabyl(race, tobacco) |>   
 adorn\_totals(where = c("row", "col"))

race Never Former Current NA\_ Total  
 NHBlack 81 260 4 80 425  
 NHWhite 31 109 2 21 163  
 <NA> 17 86 0 9 112  
 Total 129 455 6 110 700

hbp700 |> tabyl(race) |> adorn\_pct\_formatting()

race n percent valid\_percent  
 NHBlack 425 60.7% 72.3%  
 NHWhite 163 23.3% 27.7%  
 <NA> 112 16.0% -

hbp700 |> tabyl(tobacco) |> adorn\_pct\_formatting()

tobacco n percent valid\_percent  
 Never 129 18.4% 21.9%  
 Former 455 65.0% 77.1%  
 Current 6 0.9% 1.0%  
 <NA> 110 15.7% -

## Quantitative Variables in hbp700

s\_1 <- hbp700 |> reframe(lovedist(age)) |> mutate(var = "age")  
s\_2 <- hbp700 |> reframe(lovedist(nincome)) |> mutate(var = "nincome")  
s\_3 <- hbp700 |> reframe(lovedist(bmi)) |> mutate(var = "bmi")  
s\_4 <- hbp700 |> reframe(lovedist(ldl)) |> mutate(var = "ldl")  
s\_5 <- hbp700 |> reframe(lovedist(sbp)) |> mutate(var = "sbp")  
s\_6 <- hbp700 |> reframe(lovedist(dbp)) |> mutate(var = "dbp")  
s\_7 <- hbp700 |> reframe(lovedist(pp)) |> mutate(var = "pp")  
  
rbind(s\_1, s\_2, s\_3, s\_4, s\_5, s\_6, s\_7) |>   
 relocate(var)

# A tibble: 7 × 11  
 var n miss mean sd med mad min q25 q75 max  
 <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 age 700 0 65.6 9.43 6.5 e1 8.90e0 3.8 e1 5.9 e1 7.1 e1 9 e1  
2 nincome 700 11 43821. 17621. 4.18e4 1.65e4 1.68e4 2.86e4 5.17e4 1.05e5  
3 bmi 700 19 31.1 6.45 3.04e1 5.93e0 1.68e1 2.66e1 3.48e1 6.26e1  
4 ldl 700 7 82.7 27.8 8 e1 2.37e1 4 e1 6.3 e1 9.6 e1 2.25e2  
5 sbp 700 1 133. 17.8 1.32e2 1.78e1 8.8 e1 1.21e2 1.44e2 2.12e2  
6 dbp 700 1 76.4 10.1 7.6 e1 8.90e0 4.9 e1 7 e1 8.2 e1 1.22e2  
7 pp 700 1 57.0 14.1 5.6 e1 1.33e1 2.2 e1 4.7 e1 6.6 e1 1.11e2

## Today’s Outcome: Pulse Pressure

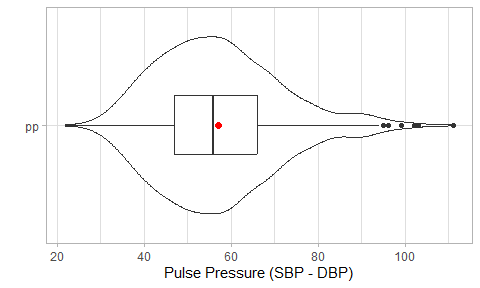
* Pulse Pressure = Systolic BP - Diastolic BP.

hbp700 <- hbp700 |> mutate(pp = sbp - dbp)  
  
ggplot(hbp700, aes(x = pp, y = "pp")) + geom\_violin() +   
 geom\_boxplot(width = 0.3) + labs(x = "Pulse Pressure (SBP - DBP)", y = "") +  
 stat\_summary(fun = "mean", geom = "point", col = "red", size = 2)

Warning: Removed 1 row containing non-finite outside the scale range  
(`stat\_ydensity()`).

Warning: Removed 1 row containing non-finite outside the scale range  
(`stat\_boxplot()`).

Warning: Removed 1 row containing non-finite outside the scale range  
(`stat\_summary()`).



## Using race to predict sbp - dbp

* 700 subjects: 1 missing sbp & dbp plus 112 missing race.

fit1 <- lm((sbp-dbp) ~ race, data = hbp700)  
summary(fit1)

Call:  
lm(formula = (sbp - dbp) ~ race, data = hbp700)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-28.191 -10.611 -1.611 8.809 53.809   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 57.1906 0.6882 83.104 <2e-16 \*\*\*  
raceNHWhite -1.5795 1.3100 -1.206 0.228   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.19 on 585 degrees of freedom  
 (113 observations deleted due to missingness)  
Multiple R-squared: 0.002479, Adjusted R-squared: 0.0007737   
F-statistic: 1.454 on 1 and 585 DF, p-value: 0.2284

## Assessing fit1

n\_obs(fit1)

[1] 587

model\_parameters(fit1)

Parameter | Coefficient | SE | 95% CI | t(585) | p  
----------------------------------------------------------------------  
(Intercept) | 57.19 | 0.69 | [55.84, 58.54] | 83.10 | < .001  
race [NHWhite] | -1.58 | 1.31 | [-4.15, 0.99] | -1.21 | 0.228

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit1)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
4783.684 | 4783.725 | 4796.809 | 0.002 | 7.737e-04 | 14.163 | 14.187

## Add in age, as a covariate

fit2 <- lm((sbp-dbp) ~ race + age, data = hbp700)  
summary(fit2)

Call:  
lm(formula = (sbp - dbp) ~ race + age, data = hbp700)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.297 -10.062 -1.709 7.953 53.321   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 36.32490 4.09642 8.867 < 2e-16 \*\*\*  
raceNHWhite -2.03827 1.28523 -1.586 0.113   
age 0.32355 0.06266 5.164 3.32e-07 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 13.89 on 584 degrees of freedom  
 (113 observations deleted due to missingness)  
Multiple R-squared: 0.04604, Adjusted R-squared: 0.04277   
F-statistic: 14.09 on 2 and 584 DF, p-value: 1.055e-06

## Assessing fit2

n\_obs(fit2) # any more missingness?

[1] 587

model\_parameters(fit2)

Parameter | Coefficient | SE | 95% CI | t(584) | p  
----------------------------------------------------------------------  
(Intercept) | 36.32 | 4.10 | [28.28, 44.37] | 8.87 | < .001  
race [NHWhite] | -2.04 | 1.29 | [-4.56, 0.49] | -1.59 | 0.113   
age | 0.32 | 0.06 | [ 0.20, 0.45] | 5.16 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit2)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
4759.475 | 4759.543 | 4776.975 | 0.046 | 0.043 | 13.850 | 13.886

# Multiple Imputation: Potential and Pitfalls

## Sterne et al. 2009 *BMJ*

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

In this article, we review the reasons why missing data may lead to bias and loss of information in epidemiological and clinical research. We discuss the circumstances in which multiple imputation may help by reducing bias or increasing precision, as well as describing potential pitfalls in its application. Finally, we describe the recent use and reporting of analyses using multiple imputation in general medical journals, and suggest guidelines for the conduct and reporting of such analyses.

* https://www.bmj.com/content/338/bmj.b2393

**Note**: The next 7 slides are derived from Sterne et al.

## An Example from Sterne et al.

Consider, for example, a study investigating the association of systolic blood pressure with the risk of subsequent coronary heart disease, in which data on systolic blood pressure are missing for some people.

The probability that systolic blood pressure is missing is likely to:

* decrease with age (doctors are more likely to measure it in older people),
* decrease with increasing body mass index, and
* decrease with history of smoking (doctors are more likely to measure it in people with heart disease risk factors or comorbidities).

If we assume that data are missing at random and that we have systolic blood pressure data on a representative sample of individuals within strata of age, smoking, body mass index, and coronary heart disease, then we can use multiple imputation to estimate the overall association between systolic blood pressure and coronary heart disease.

## Missing Data Mechanisms

* **Missing completely at random** There are no systematic differences between the missing values and the observed values.
  + For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer.
* **Missing at random** Any systematic difference between the missing and observed values can be explained by other observed data.
  + For example, missing BP measurements may be lower than measured BPs but only because younger people more often have a missing BP.
* **Missing not at random** Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values.
  + For example, people with high BP may be more likely to have headaches that cause them to miss clinic appointments.

“Missing at random” is an **assumption** that justifies the analysis, and is not a property of the data.

## Trouble: Data missing not at random

Sometimes, it is impossible to account for systematic differences between missing and observed values using the available data.

* In such (MNAR) cases, multiple imputation may give misleading results.
  + Those results can be either more or less misleading than a complete case analysis.
* For example, consider a study investigating predictors of depression.
  + If individuals are more likely to miss appointments because they are depressed on the day of the appointment, then it may be impossible to make the MAR assumption plausible, even if a large number of variables is included in the imputation model.

Where complete cases and multiple imputation analyses give different results, the analyst should attempt to understand why, and this should be reported in publications.

## What if the data are MCAR?

If we assume data are MAR, then unbiased and statistically more powerful analyses (compared with analyses based on complete cases) can generally be done by including individuals with incomplete data.

There are circumstances in which analyses of **complete cases** will not lead to bias.

* Missing data in predictor variables do not cause bias in analyses of complete cases if the reasons for the missing data are unrelated to the outcome.
  + In such cases, imputing missing data may lessen the loss of precision and power resulting from exclusion of individuals with incomplete predictor variables but are not required in order to avoid bias.

## Stages of Multiple Imputation (1 of 2)

Multiple imputation … aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them.

The first stage is to create multiple copies of the dataset, with the missing values replaced by imputed values.

* The imputation procedure must fully account for all uncertainty in predicting the missing values by injecting appropriate variability into the multiple imputed values; we can never know the true values of the missing data.

Note that single Imputation of missing values usually causes standard errors to be too small, since it fails to account for the fact that we are uncertain about the missing values.

## Stages of Multiple Imputation (2 of 2)

The second stage is to use standard statistical methods to fit the model of interest to each of the imputed datasets.

* Estimated associations in each of the imputed datasets will differ because of the variation introduced in the imputation of the missing values, and they are only useful when averaged together to give overall estimated associations.
* Standard errors are calculated using Rubin’s rules, which take account of the variability in results between the imputed datasets, reflecting the uncertainty associated with the missing values.
* Valid inferences are obtained because we are averaging over the distribution of the missing data given the observed data.

## Back to our little example

fit1 <- lm(pp ~ race, data = hbp700)  
fit2 <- lm(pp ~ race + age, data = hbp700)

How many subjects have complete / missing data affecting our models?

hbp\_sub <- hbp700 |> select(subject, sbp, dbp, pp, race, age)  
  
hbp\_sub |> pct\_complete\_case()

[1] 83.85714

hbp\_sub |> pct\_miss\_case()

[1] 16.14286

So how many imputations should we create?

## Building 20 imputations with mice

set.seed(20241017)  
  
hbp\_mice20 <- mice(hbp\_sub, m = 20, printFlag = FALSE)

Warning: Number of logged events: 101

Summary on next slide…

## Summarize Imputation Process

summary(hbp\_mice20)

Class: mids  
Number of multiple imputations: 20   
Imputation methods:  
 subject sbp dbp pp race age   
 "" "pmm" "pmm" "pmm" "logreg" ""   
PredictorMatrix:  
 subject sbp dbp pp race age  
subject 0 1 1 1 1 1  
sbp 0 0 1 1 1 1  
dbp 0 1 0 1 1 1  
pp 0 1 1 0 1 1  
race 0 1 1 1 0 1  
age 0 1 1 1 1 0  
Number of logged events: 101   
 it im dep meth out  
1 0 0 constant subject  
2 1 1 race logreg sbp  
3 1 2 race logreg sbp  
4 1 3 race logreg sbp  
5 1 4 race logreg sbp  
6 1 5 race logreg sbp

See Heymans and Eekhout sections 4.6 - 4.14 for more information.

## Imputation Options within mice

Default methods include:

* pmm predictive mean matching (default choice for quantitative variables)
* logreg logistic regression (default for binary categorical variables)
* polyreg polytomous logistic regression (for nominal multi-categorical variables)
* polr proportional odds logistic regression (for ordinal categories)

but there are cart methods and many others available, too.

## What should we include in an imputation model?

1. If things you are imputing are not Normally distributed, this can pose special challenges, and either a transformation or choosing an imputation method which is robust to these concerns is helpful.
2. Include the outcome when imputing predictors. It causes you to conclude the relationship is weaker than it actually is, if you don’t.
3. The MAR assumption may only be reasonable when a certain variable is included in the model.
   * As a result, it’s usually a good idea to include as wide a range of variables in imputation models as possible. The concerns we’d have about parsimony in outcome models don’t apply here.

## Store one (or more) of the imputed data sets

This will store the fifth imputed data set in imp\_5.

imp\_5 <- complete(hbp\_mice20, 5) |> tibble()  
  
dim(imp\_5)

[1] 700 6

n\_miss(imp\_5)

[1] 0

## Fit fit2 on 5th imputation

fit2\_i5 <- lm((sbp-dbp) ~ race + age, data = imp\_5)  
summary(fit2\_i5)

Call:  
lm(formula = (sbp - dbp) ~ race + age, data = imp\_5)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-33.406 -9.828 -1.772 7.924 53.293   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 36.08264 3.66817 9.837 < 2e-16 \*\*\*  
raceNHWhite -1.97400 1.14716 -1.721 0.0857 .   
age 0.32765 0.05555 5.898 5.73e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 13.79 on 697 degrees of freedom  
Multiple R-squared: 0.04936, Adjusted R-squared: 0.04663   
F-statistic: 18.1 on 2 and 697 DF, p-value: 2.181e-08

## Assessing fit2\_i5

n\_obs(fit2\_i5) # any missingness?

[1] 700

model\_parameters(fit2\_i5)

Parameter | Coefficient | SE | 95% CI | t(697) | p  
----------------------------------------------------------------------  
(Intercept) | 36.08 | 3.67 | [28.88, 43.28] | 9.84 | < .001  
race [NHWhite] | -1.97 | 1.15 | [-4.23, 0.28] | -1.72 | 0.086   
age | 0.33 | 0.06 | [ 0.22, 0.44] | 5.90 | < .001

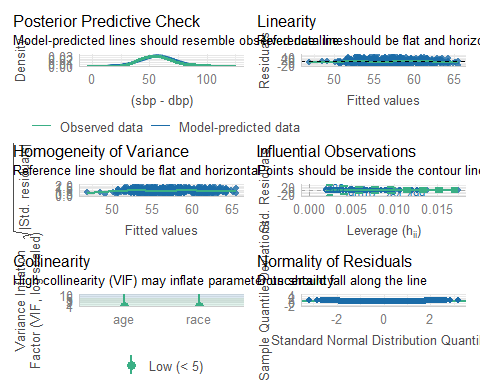
Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit2\_i5)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
5665.287 | 5665.345 | 5683.491 | 0.049 | 0.047 | 13.763 | 13.793

## Check model after single imputation

check\_model(fit2\_i5)



## fit2 on each imputed data frame

3 estimates (coefficients) times 20 imputed data sets = 60 rows.

m2\_mods <- with(hbp\_mice20, lm(pp ~ race + age))  
  
summary(m2\_mods)

# A tibble: 60 × 6  
 term estimate std.error statistic p.value nobs  
 <chr> <dbl> <dbl> <dbl> <dbl> <int>  
 1 (Intercept) 35.8 3.66 9.77 3.35e-21 700  
 2 raceNHWhite -1.91 1.15 -1.66 9.74e- 2 700  
 3 age 0.331 0.0555 5.97 3.71e- 9 700  
 4 (Intercept) 36.3 3.68 9.87 1.39e-21 700  
 5 raceNHWhite -1.77 1.18 -1.51 1.32e- 1 700  
 6 age 0.323 0.0555 5.82 9.05e- 9 700  
 7 (Intercept) 36.0 3.66 9.84 1.76e-21 700  
 8 raceNHWhite -1.84 1.15 -1.60 1.09e- 1 700  
 9 age 0.327 0.0554 5.91 5.29e- 9 700  
10 (Intercept) 36.0 3.66 9.82 2.01e-21 700  
# ℹ 50 more rows

## Pool across the 20 imputations

m2\_pool <- pool(m2\_mods)  
  
n\_obs(m2\_pool)

[1] 700 700 700 700 700 700 700 700 700 700 700 700 700 700 700 700 700 700 700  
[20] 700

model\_parameters(m2\_pool)

# Fixed Effects  
  
Parameter | Coefficient | SE | 95% CI | t | df | p  
------------------------------------------------------------------------------  
(Intercept) | 36.04 | 3.67 | [28.84, 43.24] | 9.83 | 693.73 | < .001  
race [NHWhite] | -2.06 | 1.20 | [-4.42, 0.30] | -1.71 | 570.26 | 0.088   
age | 0.33 | 0.06 | [ 0.22, 0.44] | 5.92 | 693.81 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

## Compare to our original fit2

n\_obs(fit2)

[1] 587

model\_parameters(fit2)

Parameter | Coefficient | SE | 95% CI | t(584) | p  
----------------------------------------------------------------------  
(Intercept) | 36.32 | 4.10 | [28.28, 44.37] | 8.87 | < .001  
race [NHWhite] | -2.04 | 1.29 | [-4.56, 0.49] | -1.59 | 0.113   
age | 0.32 | 0.06 | [ 0.20, 0.45] | 5.16 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

## Estimate and Adjusted

pool.r.squared(m2\_mods)

est lo 95 hi 95 fmi  
R^2 0.05052803 0.0233874 0.08659785 0.008321154

pool.r.squared(m2\_mods, adjusted = TRUE)

est lo 95 hi 95 fmi  
adj R^2 0.04780299 0.02148981 0.08316104 0.008788525

model\_performance(fit2) # original model with NAs

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
4759.475 | 4759.543 | 4776.975 | 0.046 | 0.043 | 13.850 | 13.886

Much, much more to come. See [Chapter 17 of our Course Book](https://thomaselove.github.io/431-book/17_adjtrans.html) for more discussion and an additional example.

# Breakout Activity 3

## R Packages

library(janitor)  
library(googlesheets4)  
library(naniar)  
library(xfun)  
library(easystats)  
library(tidyverse)  
  
theme\_set(theme\_light())  
  
source("c16/data/Love-431.R")

## Ingest the movies\_2024-10-15 data

gs4\_deauth()  
  
url <- "https://docs.google.com/spreadsheets/d/16fm1693sFjau9sIM-ORamxLL3ZMrBeRSxXqWpP8IBDs/edit?gid=0#gid=0"  
  
mov\_raw <- read\_sheet(url, na = c("", "NA"))

✔ Reading from "movies\_2024-10-15".

✔ Range '2024-10-15 Data'.

mov\_clean <- mov\_raw |>   
 janitor::clean\_names() |>  
 mutate(across(where(is.character), as\_factor)) |>  
 mutate(across(c(mov\_id, movie, director, star\_1, star\_2, star\_3, origin,  
 fc\_link, rt\_link, imdb\_cats, synopsis, imdb\_id, imdb\_link),  
 as.character))

# Tukey (from Ajay)

## Data Clean for Tukey (from Ajay)

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| potentially triggering events | primary language English | length | streaming services |

mov\_t1 <- mov\_clean |>  
 select(mov\_id, movie, triggers, lang\_1, length, stream\_n) |>  
 mutate(lang\_english = ifelse(lang\_1 == "English", 1, 0),  
 length\_s = scale(length, center = TRUE, scale = TRUE),  
 stream\_n = as\_factor(stream\_n),  
 stream\_f = fct\_collapse(stream\_n,   
 "0-2" = c(0, 1, 2), "4-5" = c(4, 5)),  
 stream\_f = fct\_relevel(stream\_f, "0-2", "3", "4-5"))

## Missing Status

Tukey (from Ajay)

miss\_case\_table(mov\_t1)

# A tibble: 2 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 225 98.7   
2 1 3 1.32

miss\_var\_summary(mov\_t1) |> filter(n\_miss > 0)

# A tibble: 1 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 triggers 3 1.32

## Outcome and Covariate

Tukey (from Ajay)

a1 <- mov\_t1 |> reframe(lovedist(triggers)) |> mutate(var = "triggers")  
a2 <- mov\_t1 |> reframe(lovedist(length)) |> mutate(var = "length")  
a3 <- mov\_t1 |> reframe(lovedist(length\_s)) |> mutate(var = "length\_s")  
  
rbind(a1, a2, a3) |> relocate(var)

# A tibble: 3 × 11  
 var n miss mean sd med mad min q25 q75 max  
 <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 trigg… 228 3 2.60e+ 1 18.9 24 22.2 0 10 41 77   
2 length 228 0 1.23e+ 2 25.6 118. 24.5 70 103 138 207   
3 lengt… 228 0 -2.46e-16 1 -0.223 0.956 -2.08 -0.790 0.578 3.27

## Correlations

Tukey (from Ajay)

correlation(mov\_t1 |> select(triggers, length\_s, length))

# Correlation Matrix (pearson-method)  
  
Parameter1 | Parameter2 | r | 95% CI | t | df | p  
----------------------------------------------------------------------  
triggers | length\_s | 0.15 | [0.02, 0.28] | 2.33 | 223 | 0.042\*   
triggers | length | 0.15 | [0.02, 0.28] | 2.33 | 223 | 0.042\*   
length\_s | length | 1.00 | [1.00, 1.00] | Inf | 226 | < .001\*\*\*  
  
p-value adjustment method: Holm (1979)  
Observations: 225-228

## Binary and Multi-Category Variables

Tukey (from Ajay)

mov\_t1 |> tabyl(lang\_english, stream\_n) |>  
 adorn\_totals(where = c("row", "col"))

lang\_english 0 1 2 3 4 5 Total  
 0 4 3 5 12 4 0 28  
 1 5 4 5 68 106 12 200  
 Total 9 7 10 80 110 12 228

mov\_t1 |> tabyl(lang\_english, stream\_f) |>  
 adorn\_totals(where = c("row", "col"))

lang\_english 0-2 3 4-5 Total  
 0 12 12 4 28  
 1 14 68 118 200  
 Total 26 80 122 228

## Fit 1

Tukey (from Ajay)

fit1\_t1 <- lm(triggers ~ lang\_english, data = mov\_t1)  
  
n\_obs(fit1\_t1)

[1] 225

model\_parameters(fit1\_t1)

Parameter | Coefficient | SE | 95% CI | t(223) | p  
-------------------------------------------------------------------  
(Intercept) | 12.96 | 3.59 | [5.88, 20.04] | 3.61 | < .001  
lang english | 14.69 | 3.82 | [7.16, 22.22] | 3.85 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit1\_t1)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1951.153 | 1951.262 | 1961.402 | 0.062 | 0.058 | 18.240 | 18.322

## Fit 2

Tukey (from Ajay)

fit2\_t1 <- lm(triggers ~ lang\_english + length\_s, data = mov\_t1)  
  
n\_obs(fit2\_t1)

[1] 225

model\_parameters(fit2\_t1)

Parameter | Coefficient | SE | 95% CI | t(222) | p  
-------------------------------------------------------------------  
(Intercept) | 11.25 | 3.57 | [4.22, 18.28] | 3.15 | 0.002   
lang english | 16.65 | 3.80 | [9.16, 24.14] | 4.38 | < .001  
length s | 3.77 | 1.21 | [1.38, 6.16] | 3.11 | 0.002

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit2\_t1)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1943.547 | 1943.728 | 1957.211 | 0.101 | 0.093 | 17.855 | 17.975

## Fit 3

Tukey (from Ajay)

fit3\_t1 <- lm(triggers ~ stream\_f + length\_s, data = mov\_t1)  
  
n\_obs(fit3\_t1)

[1] 225

model\_parameters(fit3\_t1)

Parameter | Coefficient | SE | 95% CI | t(221) | p  
----------------------------------------------------------------------  
(Intercept) | 18.42 | 3.67 | [11.19, 25.65] | 5.02 | < .001  
stream f [3] | 4.58 | 4.21 | [-3.72, 12.88] | 1.09 | 0.278   
stream f [4-5] | 11.06 | 4.03 | [ 3.13, 19.00] | 2.75 | 0.007   
length s | 3.00 | 1.22 | [ 0.59, 5.40] | 2.45 | 0.015

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit3\_t1)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1953.522 | 1953.796 | 1970.602 | 0.069 | 0.056 | 18.174 | 18.338

# Hard R Cafe

## Data Clean for Hard R Cafe

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| potentially triggering events | country of origin includes US | year (or age) | oscars won |

mov\_hrc <- mov\_clean |>  
 select(mov\_id, movie, triggers, origin, year, oscars) |>  
 mutate(origin\_us = as.numeric(str\_detect(origin, pattern = "USA")),  
 age = 2024 - year,  
 age\_s = scale(age, center = TRUE, scale = TRUE),  
 oscars = as\_factor(oscars),  
 oscar\_f = fct\_lump(oscars, n = 2, other\_level = "2+"))

## Missing Status

Hard R Cafe

miss\_case\_table(mov\_hrc)

# A tibble: 2 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 225 98.7   
2 1 3 1.32

miss\_var\_summary(mov\_hrc) |> filter(n\_miss > 0)

# A tibble: 1 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 triggers 3 1.32

## Outcome and Covariate

Hard R Cafe

a1 <- mov\_hrc |> reframe(lovedist(triggers)) |> mutate(var = "triggers")  
a2 <- mov\_hrc |> reframe(lovedist(age)) |> mutate(var = "age")  
a3 <- mov\_hrc |> reframe(lovedist(age\_s)) |> mutate(var = "age\_s")  
  
rbind(a1, a2, a3) |> relocate(var)

# A tibble: 3 × 11  
 var n miss mean sd med mad min q25 q75 max  
 <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 triggers 228 3 2.60e+ 1 18.9 24 22.2 0 10 41 77   
2 age 228 0 2.11e+ 1 14.3 18 11.9 0 11 28 82   
3 age\_s 228 0 -1.50e-17 1 -0.220 0.831 -1.48 -0.710 0.481 4.26

## Correlations

Hard R Cafe

correlation(mov\_hrc |> select(triggers, age\_s, age))

# Correlation Matrix (pearson-method)  
  
Parameter1 | Parameter2 | r | 95% CI | t | df | p  
-------------------------------------------------------------------------  
triggers | age\_s | -0.06 | [-0.19, 0.08] | -0.83 | 223 | 0.811   
triggers | age | -0.06 | [-0.19, 0.08] | -0.83 | 223 | 0.811   
age\_s | age | 1.00 | [ 1.00, 1.00] | Inf | 226 | < .001\*\*\*  
  
p-value adjustment method: Holm (1979)  
Observations: 225-228

## Binary and Multi-Category Variables

Hard R Cafe

mov\_hrc |> tabyl(origin\_us, oscars) |>  
 adorn\_totals(where = c("row", "col"))

origin\_us 0 1 2 3 4 5 6 7 8 11 Total  
 0 29 1 1 2 0 0 0 0 0 0 33  
 1 123 29 13 11 8 3 2 2 2 2 195  
 Total 152 30 14 13 8 3 2 2 2 2 228

mov\_hrc |> tabyl(origin\_us, oscar\_f) |>  
 adorn\_totals(where = c("row", "col"))

origin\_us 0 1 2+ Total  
 0 29 1 3 33  
 1 123 29 43 195  
 Total 152 30 46 228

## Fit 1

Hard R Cafe

fit1\_hrc <- lm(triggers ~ origin\_us, data = mov\_hrc)  
  
n\_obs(fit1\_hrc)

[1] 225

model\_parameters(fit1\_hrc)

Parameter | Coefficient | SE | 95% CI | t(223) | p  
------------------------------------------------------------------  
(Intercept) | 14.10 | 3.29 | [7.62, 20.58] | 4.29 | < .001  
origin us | 13.75 | 3.54 | [6.77, 20.73] | 3.88 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit1\_t1)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1951.153 | 1951.262 | 1961.402 | 0.062 | 0.058 | 18.240 | 18.322

## Fit 2

Hard R Cafe

fit2\_hrc <- lm(triggers ~ origin\_us + age\_s, data = mov\_hrc)  
  
n\_obs(fit2\_hrc)

[1] 225

model\_parameters(fit2\_hrc)

Parameter | Coefficient | SE | 95% CI | t(222) | p  
-------------------------------------------------------------------  
(Intercept) | 14.20 | 3.30 | [ 7.70, 20.69] | 4.31 | < .001  
origin us | 13.64 | 3.55 | [ 6.64, 20.63] | 3.84 | < .001  
age s | -0.84 | 1.25 | [-3.30, 1.63] | -0.67 | 0.504

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit2\_hrc)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1952.421 | 1952.603 | 1966.085 | 0.065 | 0.057 | 18.210 | 18.333

## Fit 3

Hard R Cafe

fit3\_hrc <- lm(triggers ~ oscar\_f + age\_s, data = mov\_hrc)  
  
n\_obs(fit3\_hrc)

[1] 225

model\_parameters(fit3\_hrc)

Parameter | Coefficient | SE | 95% CI | t(221) | p  
--------------------------------------------------------------------  
(Intercept) | 22.42 | 1.50 | [19.47, 25.37] | 14.98 | < .001  
oscar f [1] | 10.50 | 3.67 | [ 3.28, 17.72] | 2.86 | 0.005   
oscar f [2+] | 10.66 | 3.12 | [ 4.50, 16.81] | 3.41 | < .001  
age s | -1.56 | 1.25 | [-4.02, 0.91] | -1.24 | 0.215

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit3\_hrc)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1952.596 | 1952.870 | 1969.677 | 0.073 | 0.060 | 18.137 | 18.300

# Tukey (from Atticus)

## Data Clean: Tukey (from Atticus)

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| potentially triggering events | star 1 gender | length | oscars won |

mov\_t2 <- mov\_clean |>  
 select(mov\_id, movie, triggers, gen\_1, length, oscars) |>  
 mutate(length\_s = scale(length, center = TRUE, scale = TRUE),  
 oscars = as\_factor(oscars),  
 oscar\_f = fct\_lump(oscars, n = 2, other\_level = "2+"))

## Missing Status

Tukey (from Atticus)

miss\_case\_table(mov\_t2)

# A tibble: 2 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 225 98.7   
2 1 3 1.32

miss\_var\_summary(mov\_t2) |> filter(n\_miss > 0)

# A tibble: 1 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 triggers 3 1.32

## Outcome and Covariate

Tukey (from Atticus)

a1 <- mov\_t2 |> reframe(lovedist(triggers)) |> mutate(var = "triggers")  
a2 <- mov\_t2 |> reframe(lovedist(length)) |> mutate(var = "length")  
a3 <- mov\_t2 |> reframe(lovedist(length\_s)) |> mutate(var = "length\_s")  
  
rbind(a1, a2, a3) |> relocate(var)

# A tibble: 3 × 11  
 var n miss mean sd med mad min q25 q75 max  
 <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 trigg… 228 3 2.60e+ 1 18.9 24 22.2 0 10 41 77   
2 length 228 0 1.23e+ 2 25.6 118. 24.5 70 103 138 207   
3 lengt… 228 0 -2.46e-16 1 -0.223 0.956 -2.08 -0.790 0.578 3.27

## Correlations

Tukey (from Atticus)

correlation(mov\_t2 |> select(triggers, length\_s, length))

# Correlation Matrix (pearson-method)  
  
Parameter1 | Parameter2 | r | 95% CI | t | df | p  
----------------------------------------------------------------------  
triggers | length\_s | 0.15 | [0.02, 0.28] | 2.33 | 223 | 0.042\*   
triggers | length | 0.15 | [0.02, 0.28] | 2.33 | 223 | 0.042\*   
length\_s | length | 1.00 | [1.00, 1.00] | Inf | 226 | < .001\*\*\*  
  
p-value adjustment method: Holm (1979)  
Observations: 225-228

## Binary and Multi-Category Variables

Tukey (from Atticus)

mov\_t2 |> tabyl(gen\_1, oscars) |>  
 adorn\_totals(where = c("row", "col"))

gen\_1 0 1 2 3 4 5 6 7 8 11 Total  
 M 114 25 13 11 8 2 2 0 1 2 178  
 F 38 5 1 2 0 1 0 2 1 0 50  
 Total 152 30 14 13 8 3 2 2 2 2 228

mov\_t2 |> tabyl(gen\_1, oscar\_f) |>  
 adorn\_totals(where = c("row", "col"))

gen\_1 0 1 2+ Total  
 M 114 25 39 178  
 F 38 5 7 50  
 Total 152 30 46 228

## Fit 1

Tukey (from Atticus)

fit1\_t2 <- lm(triggers ~ gen\_1, data = mov\_t2)  
  
n\_obs(fit1\_t2)

[1] 225

model\_parameters(fit1\_t2)

Parameter | Coefficient | SE | 95% CI | t(223) | p  
-------------------------------------------------------------------  
(Intercept) | 26.33 | 1.43 | [23.52, 29.15] | 18.42 | < .001  
gen 1 [F] | -1.69 | 3.03 | [-7.67, 4.28] | -0.56 | 0.577

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit1\_t2)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1965.284 | 1965.393 | 1975.533 | 0.001 | -0.003 | 18.822 | 18.906

## Fit 2

Tukey (from Atticus)

fit2\_t2 <- lm(triggers ~ gen\_1 + length\_s, data = mov\_t2)  
  
n\_obs(fit2\_t2)

[1] 225

model\_parameters(fit2\_t2)

Parameter | Coefficient | SE | 95% CI | t(222) | p  
-------------------------------------------------------------------  
(Intercept) | 26.12 | 1.42 | [23.32, 28.91] | 18.40 | < .001  
gen 1 [F] | -0.63 | 3.04 | [-6.63, 5.36] | -0.21 | 0.835   
length s | 2.85 | 1.26 | [ 0.37, 5.34] | 2.26 | 0.025

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit2\_t2)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1962.161 | 1962.343 | 1975.826 | 0.024 | 0.015 | 18.609 | 18.734

## Fit 3

Tukey (from Atticus)

fit3\_t2 <- lm(triggers ~ oscar\_f + length\_s, data = mov\_t2)  
  
n\_obs(fit3\_t2)

[1] 225

model\_parameters(fit3\_t2)

Parameter | Coefficient | SE | 95% CI | t(221) | p  
--------------------------------------------------------------------  
(Intercept) | 22.84 | 1.50 | [19.87, 25.80] | 15.18 | < .001  
oscar f [1] | 9.74 | 3.67 | [ 2.51, 16.96] | 2.66 | 0.008   
oscar f [2+] | 9.17 | 3.19 | [ 2.89, 15.45] | 2.88 | 0.004   
length s | 1.97 | 1.25 | [-0.49, 4.43] | 1.58 | 0.116

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit3\_t2)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
1951.653 | 1951.927 | 1968.733 | 0.077 | 0.064 | 18.099 | 18.262

# Halloween Time Stats

## Data Clean for Halloween Time Stats

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| budget | origin in US | year (or age) | Bechdel-Wallace |

mov\_hts <- mov\_clean |>  
 select(mov\_id, movie, budget, origin, year, bw\_rating) |>  
 mutate(budgetM = budget/1000000,  
 origin\_us = as.numeric(str\_detect(origin, pattern = "USA")),  
 age = 2024 - year,  
 age\_s = scale(age, center = TRUE, scale = TRUE))

## Missing Status

Halloween Time Stats

miss\_case\_table(mov\_hts)

# A tibble: 4 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 203 89.0   
2 1 3 1.32  
3 2 15 6.58  
4 3 7 3.07

miss\_var\_summary(mov\_hts) |> filter(n\_miss > 0)

# A tibble: 3 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 budget 22 9.65  
2 budgetM 22 9.65  
3 bw\_rating 10 4.39

## Outcome and Covariate

Halloween Time Stats

a1 <- mov\_hts |> reframe(lovedist(budgetM)) |> mutate(var = "budgetM")  
a2 <- mov\_hts |> reframe(lovedist(age)) |> mutate(var = "age")  
a3 <- mov\_hts |> reframe(lovedist(age\_s)) |> mutate(var = "age\_s")  
  
rbind(a1, a2, a3) |> relocate(var)

# A tibble: 3 × 11  
 var n miss mean sd med mad min q25 q75 max  
 <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 budgetM 228 22 5.91e+ 1 66.5 30 35.6 0.2 13 90 356   
2 age 228 0 2.11e+ 1 14.3 18 11.9 0 11 28 82   
3 age\_s 228 0 -1.50e-17 1 -0.220 0.831 -1.48 -0.710 0.481 4.26

## Correlations

Halloween Time Stats

correlation(mov\_hts |> select(budgetM, age\_s, age))

# Correlation Matrix (pearson-method)  
  
Parameter1 | Parameter2 | r | 95% CI | t | df | p  
--------------------------------------------------------------------------  
budgetM | age\_s | -0.38 | [-0.49, -0.26] | -5.91 | 204 | < .001\*\*\*  
budgetM | age | -0.38 | [-0.49, -0.26] | -5.91 | 204 | < .001\*\*\*  
age\_s | age | 1.00 | [ 1.00, 1.00] | Inf | 226 | < .001\*\*\*  
  
p-value adjustment method: Holm (1979)  
Observations: 206-228

## Binary and Multi-Category Variables

Halloween Time Stats

mov\_hts |> tabyl(origin\_us, bw\_rating) |>  
 adorn\_totals(where = c("row", "col"))

origin\_us 0 1 2 3 NA\_ Total  
 0 4 6 1 15 7 33  
 1 14 55 14 109 3 195  
 Total 18 61 15 124 10 228

## Fit 1

Halloween Time Stats

fit1\_hts <- lm(budgetM ~ origin\_us, data = mov\_hts)  
  
n\_obs(fit1\_hts)

[1] 206

model\_parameters(fit1\_hts)

Parameter | Coefficient | SE | 95% CI | t(204) | p  
---------------------------------------------------------------------  
(Intercept) | 11.00 | 14.49 | [-17.56, 39.56] | 0.76 | 0.449   
origin us | 53.28 | 15.25 | [ 23.22, 83.34] | 3.49 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit1\_hts)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
2307.089 | 2307.208 | 2317.073 | 0.056 | 0.052 | 64.472 | 64.787

## Fit 2

Halloween Time Stats

fit2\_hts <- lm(budgetM ~ origin\_us + age\_s, data = mov\_hts)  
  
n\_obs(fit2\_hts)

[1] 206

model\_parameters(fit2\_hts)

Parameter | Coefficient | SE | 95% CI | t(203) | p  
----------------------------------------------------------------------  
(Intercept) | 19.00 | 13.57 | [ -7.75, 45.74] | 1.40 | 0.163   
origin us | 44.72 | 14.28 | [ 16.56, 72.87] | 3.13 | 0.002   
age s | -24.75 | 4.36 | [-33.36, -16.14] | -5.67 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit2\_hts)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
2278.799 | 2278.998 | 2292.110 | 0.185 | 0.177 | 59.902 | 60.343

## Fit 3

Halloween Time Stats

fit3\_hts <- lm(budgetM ~ bw\_rating + age\_s, data = mov\_hts)  
  
n\_obs(fit3\_hts)

[1] 203

model\_parameters(fit3\_hts)

Parameter | Coefficient | SE | 95% CI | t(200) | p  
---------------------------------------------------------------------  
(Intercept) | 55.33 | 9.63 | [ 36.33, 74.32] | 5.74 | < .001  
bw rating | 2.25 | 4.08 | [ -5.80, 10.31] | 0.55 | 0.582   
age s | -25.65 | 4.53 | [-34.59, -16.72] | -5.66 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

model\_performance(fit3\_hts)

# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
--------------------------------------------------------------------  
2256.107 | 2256.309 | 2269.360 | 0.147 | 0.138 | 61.453 | 61.912

# Data Cleaning and Missing Status only

## Data Clean for Something Unique Tokyo Drift

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| budget | origin in US | length | oscars won |

mov\_sutd <- mov\_clean |>  
 select(mov\_id, movie, budget, origin, length, oscars) |>  
 mutate(budgetM = budget/1000000,  
 origin\_us = as.numeric(str\_detect(origin, pattern = "USA")),  
 length\_s = scale(length, center = TRUE, scale = TRUE),  
 oscars = as\_factor(oscars),  
 oscar\_f = fct\_lump(oscars, n = 2, other\_level = "2+"))

## Missing Status

Something Unique Tokyo Drift

miss\_case\_table(mov\_sutd)

# A tibble: 2 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 206 90.4   
2 2 22 9.65

miss\_var\_summary(mov\_sutd) |> filter(n\_miss > 0)

# A tibble: 2 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 budget 22 9.65  
2 budgetM 22 9.65

## Data Clean for PB&J

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| gross revenue | star 1 gender | budget | oscars won |

mov\_pbj <- mov\_clean |>  
 select(mov\_id, movie, gross\_world, gen\_1, budget, oscars) |>  
 mutate(grossM = gross\_world/1000000,  
 budget\_s = scale(budget, center = TRUE, scale = TRUE),  
 oscars = as\_factor(oscars),  
 oscar\_f = fct\_lump(oscars, n = 2, other\_level = "2+"))

## Missing Status

PB&J

miss\_case\_table(mov\_pbj)

# A tibble: 3 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 204 89.5   
2 2 18 7.89  
3 4 6 2.63

miss\_var\_summary(mov\_pbj) |> filter(n\_miss > 0)

# A tibble: 4 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 "budget" 22 9.65  
2 "" 22 9.65  
3 "gross\_world" 8 3.51  
4 "grossM" 8 3.51

## Data Clean for And Then There Were Three

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| gross revenue | drama | sex-nudity rating | ebert rating |

mov\_attwt <- mov\_clean |>  
 select(mov\_id, movie, gross\_world, drama, kim\_sn, ebert) |>  
 mutate(grossM = gross\_world/1000000,  
 ebert = as\_factor(ebert),  
 ebert\_f = fct\_lump(ebert, n = 4, other\_level = "1-2"),  
 ebert\_f = fct\_relevel(ebert\_f, "1-2"))

## Missing Status

And Then There Were Three

miss\_case\_table(mov\_attwt)

# A tibble: 5 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 146 64.0   
2 1 50 21.9   
3 2 11 4.82  
4 3 16 7.02  
5 5 5 2.19

miss\_var\_summary(mov\_attwt) |> filter(n\_miss > 0)

# A tibble: 5 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 kim\_sn 71 31.1   
2 ebert 29 12.7   
3 ebert\_f 29 12.7   
4 gross\_world 8 3.51  
5 grossM 8 3.51

## Data Clean for Vintage Macbooks

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| awards won | has Dr. Love seen | length | Bechdel-Wallace |

mov\_vm <- mov\_clean |>  
 select(mov\_id, movie, awards, dr\_love, length, bw\_rating) |>  
 mutate(length\_s = scale(length, center = TRUE, scale = TRUE))

## Missing Status

Vintage Macbooks

miss\_case\_table(mov\_vm)

# A tibble: 2 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 218 95.6   
2 1 10 4.39

miss\_var\_summary(mov\_vm) |> filter(n\_miss > 0)

# A tibble: 1 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 bw\_rating 10 4.39

## Data Clean for Ghostbusters

| Outcome | Binary | Covariate | Multi-Cat |
| --- | --- | --- | --- |
| awards won | star 1 gender | year (or age) | Bechdel-Wallace |

mov\_gb <- mov\_clean |>  
 select(mov\_id, movie, awards, gen\_1, year, bw\_rating) |>  
 mutate(age = 2024 - year,  
 age\_s = scale(age, center = TRUE, scale = TRUE))

## Missing Status

Ghostbusters

miss\_case\_table(mov\_gb)

# A tibble: 2 × 3  
 n\_miss\_in\_case n\_cases pct\_cases  
 <int> <int> <dbl>  
1 0 218 95.6   
2 1 10 4.39

miss\_var\_summary(mov\_gb) |> filter(n\_miss > 0)

# A tibble: 1 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <num>  
1 bw\_rating 10 4.39

## Session Information

session\_info()

R version 4.4.1 (2024-06-14 ucrt)  
Platform: x86\_64-w64-mingw32/x64  
Running under: Windows 11 x64 (build 22631)  
  
Locale:  
 LC\_COLLATE=English\_United States.utf8   
 LC\_CTYPE=English\_United States.utf8   
 LC\_MONETARY=English\_United States.utf8  
 LC\_NUMERIC=C   
 LC\_TIME=English\_United States.utf8   
  
Package version:  
 askpass\_1.2.1 backports\_1.5.0 base64enc\_0.1.3   
 bayestestR\_0.14.0 bit\_4.5.0 bit64\_4.5.2   
 blob\_1.2.4 boot\_1.3-31 broom\_1.0.7   
 bslib\_0.8.0 cachem\_1.1.0 callr\_3.7.6   
 cellranger\_1.1.0 cli\_3.6.3 clipr\_0.8.0   
 coda\_0.19-4.1 codetools\_0.2-20 colorspace\_2.1-1   
 compiler\_4.4.1 conflicted\_1.2.0 correlation\_0.8.5   
 cpp11\_0.5.0 crayon\_1.5.3 curl\_5.2.3   
 data.table\_1.16.2 datasets\_4.4.1 datawizard\_0.13.0   
 DBI\_1.2.3 dbplyr\_2.5.0 digest\_0.6.37   
 dplyr\_1.1.4 dtplyr\_1.3.1 easystats\_0.7.3   
 effectsize\_0.8.9 emmeans\_1.10.5 estimability\_1.5.1   
 evaluate\_1.0.1 fansi\_1.0.6 farver\_2.1.2   
 fastmap\_1.2.0 fontawesome\_0.5.2 forcats\_1.0.0   
 foreach\_1.5.2 fs\_1.6.4 gargle\_1.5.2   
 generics\_0.1.3 ggplot2\_3.5.1 ggrepel\_0.9.6   
 glmnet\_4.1-8 glue\_1.8.0 googledrive\_2.1.1   
 googlesheets4\_1.1.1 graphics\_4.4.1 grDevices\_4.4.1   
 grid\_4.4.1 gridExtra\_2.3 gtable\_0.3.5   
 haven\_2.5.4 highr\_0.11 hms\_1.1.3   
 htmltools\_0.5.8.1 httr\_1.4.7 ids\_1.0.1   
 insight\_0.20.5 isoband\_0.2.7 iterators\_1.0.14   
 janitor\_2.2.0 jomo\_2.7-6 jquerylib\_0.1.4   
 jsonlite\_1.8.9 kableExtra\_1.4.0 knitr\_1.48   
 labeling\_0.4.3 lattice\_0.22-6 lifecycle\_1.0.4   
 lme4\_1.1-35.5 lubridate\_1.9.3 magrittr\_2.0.3   
 MASS\_7.3-61 Matrix\_1.7-0 memoise\_2.0.1   
 methods\_4.4.1 mgcv\_1.9-1 mice\_3.16.0   
 mime\_0.12 minqa\_1.2.8 mitml\_0.4-5   
 modelbased\_0.8.8 modelr\_0.1.11 multcomp\_1.4-26   
 munsell\_0.5.1 mvtnorm\_1.3-1 naniar\_1.1.0   
 nlme\_3.1-164 nloptr\_2.1.1 nnet\_7.3-19   
 norm\_1.0.11.1 numDeriv\_2016.8.1.1 openssl\_2.2.2   
 ordinal\_2023.12.4.1 pan\_1.9 parallel\_4.4.1   
 parameters\_0.22.2 patchwork\_1.3.0 performance\_0.12.3   
 pillar\_1.9.0 pkgconfig\_2.0.3 plyr\_1.8.9   
 prettyunits\_1.2.0 processx\_3.8.4 progress\_1.2.3   
 ps\_1.8.0 purrr\_1.0.2 R6\_2.5.1   
 ragg\_1.3.3 rappdirs\_0.3.3 RColorBrewer\_1.1.3   
 Rcpp\_1.0.13 RcppEigen\_0.3.4.0.2 readr\_2.1.5   
 readxl\_1.4.3 rematch\_2.0.0 rematch2\_2.1.2   
 report\_0.5.9 reprex\_2.1.1 rlang\_1.1.4   
 rmarkdown\_2.28 rpart\_4.1.23 rstudioapi\_0.16.0   
 rvest\_1.0.4 sandwich\_3.1-1 sass\_0.4.9   
 scales\_1.3.0 see\_0.9.0 selectr\_0.4.2   
 shape\_1.4.6.1 snakecase\_0.11.1 splines\_4.4.1   
 stats\_4.4.1 stringi\_1.8.4 stringr\_1.5.1   
 survival\_3.7-0 svglite\_2.1.3 sys\_3.4.3   
 systemfonts\_1.1.0 textshaping\_0.4.0 TH.data\_1.1-2   
 tibble\_3.2.1 tidyr\_1.3.1 tidyselect\_1.2.1   
 tidyverse\_2.0.0 timechange\_0.3.0 tinytex\_0.53   
 tools\_4.4.1 tzdb\_0.4.0 ucminf\_1.2.2   
 UpSetR\_1.4.0 utf8\_1.2.4 utils\_4.4.1   
 uuid\_1.2.1 vctrs\_0.6.5 viridis\_0.6.5   
 viridisLite\_0.4.2 visdat\_0.6.0 vroom\_1.6.5   
 withr\_3.0.1 xfun\_0.48 xml2\_1.3.6   
 xtable\_1.8-4 yaml\_2.3.10 zoo\_1.8-12