432 Class 12

https://thomaselove.github.io/432-2025/

2025-02-20

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(janitor)  
library(naniar)  
library(broom)  
library(gt)  
library(gtsummary)  
library(mosaic)  
library(here) # introduced today  
library(conflicted) # introduced today  
library(tableone) # building Table One  
library(survey) # survey-specific tools for weighting  
library(nhanesA) # data source  
library(easystats)  
library(tidyverse)  
  
theme\_set(theme\_bw())   
conflicts\_prefer(dplyr::filter(), base::max(), base::sum(), base::mean())

## Today’s Agenda

1. The **conflicted** package
2. The **here** package
3. Building Table One
4. Working with Survey Weights

## What does the here package do?

The **here** package creates paths relative to the top-level directory. The package displays the top-level of the current project on load or any time you call here().

here()

[1] "D:/Teaching/432/2025/432-slides-2025"

here("c12", "data", "bradley.csv")

[1] "D:/Teaching/432/2025/432-slides-2025/c12/data/bradley.csv"

or

here("c12/data/bradley.csv")

[1] "D:/Teaching/432/2025/432-slides-2025/c12/data/bradley.csv"

## The here package!



## More on the here package and related issues

* [Jenny Bryan on the here package](https://github.com/jennybc/here_here)
* here package details are available at <https://here.r-lib.org/index.html>
* [Workflow: Scripts and Projects](https://r4ds.hadley.nz/workflow-scripts.html) from R for Data Science

## The conflicted package

<https://conflicted.r-lib.org/>

* The goal of conflicted is to provide an alternative conflict resolution strategy. R’s default conflict resolution system gives precedence to the most recently loaded package. This can make it hard to detect conflicts, particularly when introduced by an update to an existing package. conflicted takes a different approach, making every conflict an error and forcing you to choose which function to use.

## Conflict Resolution here

In the package setup today, I used the following code to resolve errors that came up in these slides:

conflicts\_prefer(dplyr::filter(), base::max(), base::sum(), base::mean())

* That was the complete set of conflicts I had to resolve to get the slides to run.
* As it turns out, though, there are lots of other conflicts that didn’t crop up in building these slides, because the relevant conflict never emerged…

## What else does conflicted do?

conflict\_scout()

30 conflicts

• `chisq.test()`: janitor and stats

• `clean\_names()`: insight and janitor

• `cor()`: mosaic and stats

• `cor\_test()`: correlation and mosaic

• `count()`: mosaic and dplyr

• `cov()`: mosaic and stats

• `cross()`: purrr and mosaic

• `do()`: mosaic and dplyr

• `dotchart()`: survey and graphics

• `expand()`: tidyr and Matrix

• `filter()`: dplyr

• `fisher.test()`: janitor and stats

• `IQR()`: mosaic and stats

• `lag()`: dplyr and stats

• `max()`: base

• `mean()`: base

• `min()`: mosaic and base

• `pack()`: tidyr and Matrix

• `prod()`: mosaic and base

• `prop.test()`: mosaic and stats

• `range()`: mosaic and base

• `remove\_empty()`: datawizard and janitor

• `remove\_empty\_rows()`: datawizard and janitor

• `rescale()`: datawizard and mosaic

• `sample()`: mosaic and base

• `stat()`: mosaic and ggplot2

• `sum()`: base

• `tally()`: mosaic and dplyr

• `unpack()`: tidyr and Matrix

• `var()`: mosaic and stats

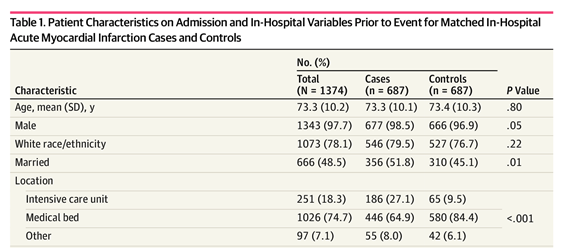
# Building a Table One

## An Original Clinical Investigation

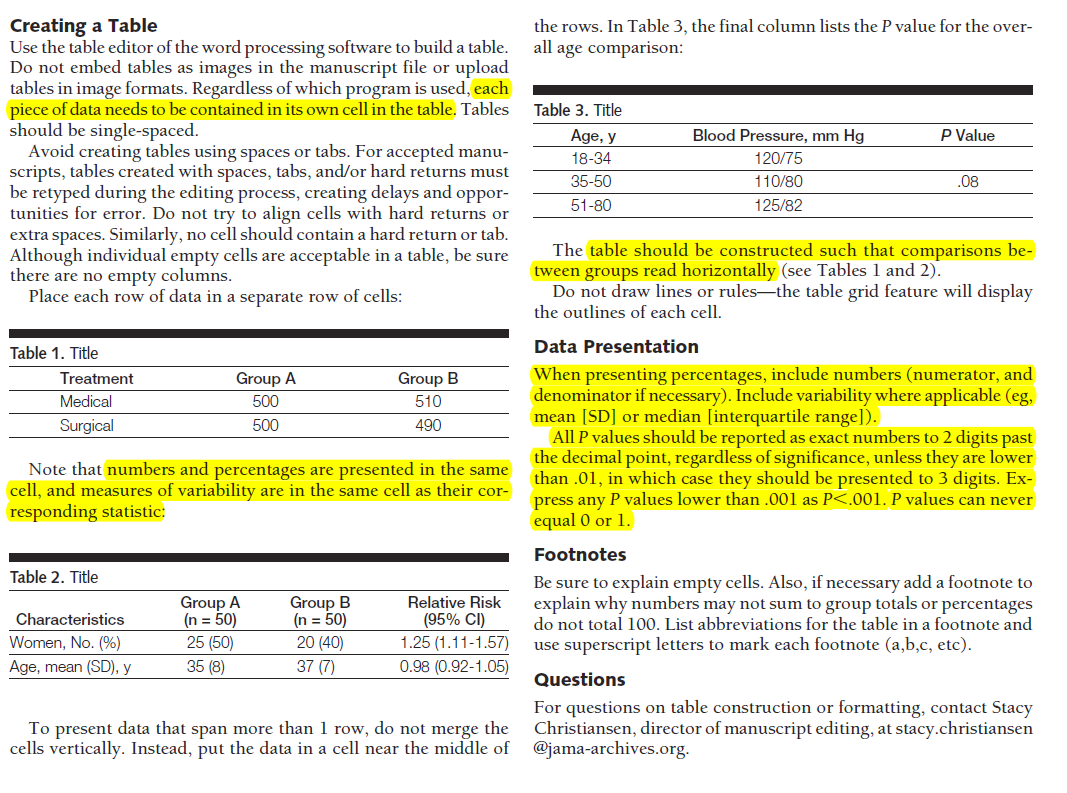


[Link to Source](https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2720923)

## Part of Bradley et al.’s Table 1



## Table Creation Instructions, JAMA: [linked here](https://jama.jamanetwork.com/data/ifora-forms/jama/tablecreationinst.pdf)



## A Data Set

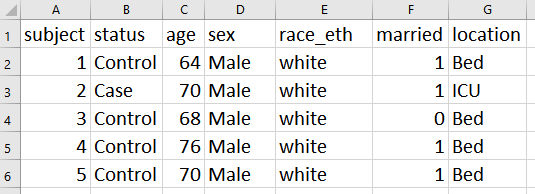
The bradley.csv data set is simulated, but consists of 1,374 observations (687 Cases and 687 Controls) containing:

* a subject identification code, in subject
* status (case or control)
* age (in years)
* sex (Male or Female)
* race/ethnicity (white or non-white)
* married (1 = yes or 0 = no)
* location (ICU, bed, other)

The bradley.csv data closely match the summary statistics provided in Table 1 of the Bradley et al. article. Our job is to recreate that part of Table 1, as best as we can.

## The bradley.csv data (first 5 rows)

* The bradley\_sim.md file on our web site shows you how I simulated the data.



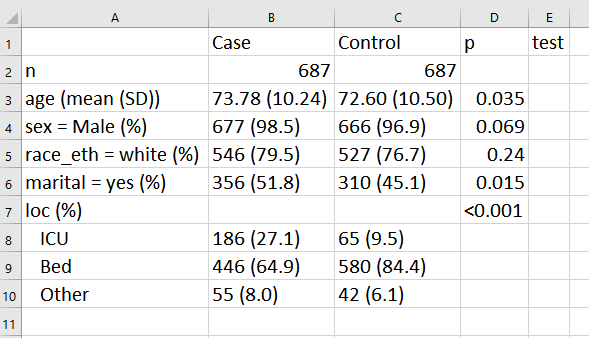
## To “Live” Coding

On our web site (Data and Code + Class 12 materials)

* In the data folder:
  + bradley.csv data file
* bradley\_table1.qmd Quarto script
* bradley\_table1.md Results of running Quarto
* bradley\_table1\_result.csv is the table generated by that Quarto script

# To The “Live Code” at <https://rpubs.com/TELOVE/bradley-table1-432>

## Opening bradley\_table1\_result.csv in Excel



## Learning More About Table 1

Chapter 18 of the Course Notes covers two larger examples, and more details, like…

* specifying factors, and re-ordering them when necessary
* using non-normal summaries or exact categorical tests
* dealing with warning messages and with missing data
* producing Table 1 in R so you can cut and paste it into Excel or Word

and Lab 5 asks you to do this with a familiar data set.

# Incorporating survey weights (an introduction)

## What are survey weights?

In many surveys, each sampled subject is assigned a weight that is equivalent to the reciprocal of his/her probability of selection into the sample.

but more sophisticated sampling designs require more complex weighting schemes. Usually these are published as part of the survey data.

I’ll demonstrate part of the survey package today.

## An NHANES Example

Let’s use the NHANES 2013-14 data and pull in both the demographics (DEMO\_H) and total cholesterol (TCHOL\_H) databases.

demo\_raw <- nhanes('DEMO\_H', translated = FALSE)  
tchol\_raw <- nhanes('TCHOL\_H', translated = FALSE)

Detailed descriptions available at

* <https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DEMO_H.htm>
* <https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/TCHOL_H.htm>

## Weighting in NHANES

Weights for each sampled person in NHANES account for the complex survey design. The weight describes the number of people in the population represented by the sampled person, and is created in three steps:

1. the base weight is computed, which accounts for the unequal probabilities of selection given that some demographic groups were over-sampled;
2. adjustments are made for non-response; and
3. post-stratification adjustments are made to match estimates of the U.S. civilian non-institutionalized population available from the Census Bureau.

Source: <https://wwwn.cdc.gov/nchs/nhanes/tutorials/Module3.aspx>

## Weights in our NHANES data

The DEMO file contains two kinds of sampling weights:

* the interview weight (WTINT2yr), and
* the MEC exam weight (WTMEC2yr)

NHANES also provides several weights for subsamples. In NHANES, we identify the variable of interest that was collected on the smallest number of respondents. The sample weight that applies to that variable is the appropriate one to use. In our first case, we will study total cholesterol and use the weights from the MEC exam.

## What Variables Do We Need?

* SEQN = subject identifying code
* RIAGENDR = sex (1 = M, 2 = F)
* RIDAGEYR = age (in years at screening, topcode at 80)
* DMQMILIZ = served active duty in US Armed Forces (yes/no)
* RIDSTATR = 2 if subject took both interview and MEC exam
* WTMEC2YR - Full sample 2 year MEC exam weight
* LBXTC = Total Cholesterol (mg/dl) - this is our outcome

The first 5 are in DEMO\_H, and the first and last are in TCHOL\_H.

## Merge the DEMO and TCHOL files

dim(demo\_raw)

[1] 10175 47

dim(tchol\_raw)

[1] 8291 3

joined\_df <- inner\_join(demo\_raw, tchol\_raw, by = c("SEQN"))  
  
dim(joined\_df)

[1] 8291 49

## Create and save a small analytic tibble

nh1314 <- joined\_df |> # has n = 8291  
 tibble() |>  
 filter(complete.cases(LBXTC)) |> # now n = 7624  
 filter(RIDSTATR == 2) |> # still 7624  
 filter(RIDAGEYR > 19 & RIDAGEYR < 40) |> # now n = 1802  
 filter(DMQMILIZ < 3) |> # drop 7 = refused, n = 1801  
 mutate(FEMALE = RIAGENDR - 1,  
 AGE = RIDAGEYR,  
 US\_MIL = ifelse(DMQMILIZ == 1, 1, 0),  
 WT\_EX = WTMEC2YR,  
 TOTCHOL = LBXTC) |>  
 select(SEQN, FEMALE, AGE, TOTCHOL, US\_MIL, WT\_EX)  
  
write\_rds(nh1314, here("c12/data/nh1314.Rds"))

* The nh1314.Rds file is on our 432-data page if you need it.

## nh1314 analytic sample

nh1314 |> select(AGE, TOTCHOL, WT\_EX) |> summary()

AGE TOTCHOL WT\_EX   
 Min. :20.00 Min. : 69 Min. : 8430   
 1st Qu.:24.00 1st Qu.:156 1st Qu.: 24694   
 Median :30.00 Median :178 Median : 34642   
 Mean :29.47 Mean :181 Mean : 44529   
 3rd Qu.:34.00 3rd Qu.:203 3rd Qu.: 59561   
 Max. :39.00 Max. :417 Max. :125680

nh1314 |> tabyl(FEMALE, US\_MIL) |>   
 adorn\_totals(where = c("row", "col")) |> adorn\_title()

US\_MIL   
 FEMALE 0 1 Total  
 0 829 45 874  
 1 921 6 927  
 Total 1750 51 1801

## Formatting df\_stats with gt()

df\_stats(~ AGE + TOTCHOL, data = nh1314) |>  
 mutate(across(mean:sd, ~ round\_half\_up(.x, 2))) |>   
 gt() |> tab\_options(table.font.size = 20) |>  
 tab\_header(title = "Approach A",   
 subtitle = "Data from nh1314 sample, unadjusted")

Table 1: Approach A

Data from nh1314 sample, unadjusted

| response | min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AGE | 20 | 24 | 30 | 34 | 39 | 29.47 | 5.80 | 1801 | 0 |
| TOTCHOL | 69 | 156 | 178 | 203 | 417 | 181.01 | 37.41 | 1801 | 0 |

## Formatting df\_stats with gt()

df\_stats(~ AGE + TOTCHOL, data = nh1314) |>  
 gt() |> fmt\_number(columns = mean:sd, decimals = 2) |>  
 tab\_options(table.font.size = 20) |>  
 tab\_header(title = "Approach B",   
 subtitle = "Data from nh1314 sample, unadjusted")

Table 1: Approach B

Data from nh1314 sample, unadjusted

| response | min | Q1 | median | Q3 | max | mean | sd | n | missing |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AGE | 20 | 24 | 30 | 34 | 39 | 29.47 | 5.80 | 1801 | 0 |
| TOTCHOL | 69 | 156 | 178 | 203 | 417 | 181.01 | 37.41 | 1801 | 0 |

## Our nh1314 analytic sample: Weights

Each weight represents the number of people exemplified by that subject.

favstats(~ WT\_EX, data = nh1314) |>  
 rename(na = missing) |> gt() |>  
 tab\_options(table.font.size = 20)

| min | Q1 | median | Q3 | max | mean | sd | n | na |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8430.461 | 24694.05 | 34642.05 | 59560.74 | 125680.3 | 44528.66 | 26027.44 | 1801 | 0 |

## Describing nh1314 (unweighted)

* using tbl\_summary() from the **gtsummary** package. See <https://www.danieldsjoberg.com/gtsummary/> for more options.

table1 <- nh1314 |>  
 tbl\_summary(include = -SEQN)  
  
table1

| **Characteristic** | **N = 1,801***1* |
| --- | --- |
| FEMALE | 927 (51%) |
| AGE | 30.0 (24.0, 34.0) |
| TOTCHOL | 178 (156, 203) |
| US\_MIL | 51 (2.8%) |
| WT\_EX | 34,642 (24,694, 59,561) |
| *1*n (%); Median (Q1, Q3) | |

## Create nh\_design survey design

nh\_design <-   
 svydesign(  
 id = ~ SEQN,  
 weights = ~ WT\_EX,  
 data = nh1314)   
  
nh\_design <- update( nh\_design, one = 1)   
  
## this one = 1 business will help us count  
  
nh\_design

Independent Sampling design (with replacement)  
update(nh\_design, one = 1)

## Unweighted Counts

### Overall

sum(weights(nh\_design, "sampling") != 0)

[1] 1801

### By Groups

svyby( ~ one, ~ FEMALE, nh\_design, unwtd.count)

FEMALE counts se  
0 0 874 0  
1 1 927 0

svyby( ~ one, ~ FEMALE + US\_MIL, nh\_design, unwtd.count)

FEMALE US\_MIL counts se  
0.0 0 0 829 0  
1.0 1 0 921 0  
0.1 0 1 45 0  
1.1 1 1 6 0

## Weighted Counts

svytotal( ~ one, nh\_design )

total SE  
one 80196108 1104558

### By Groups

svyby( ~ one, ~ FEMALE, nh\_design, svytotal)

FEMALE one se  
0 0 39694756 1255122  
1 1 40501352 1196260

svyby( ~ one, ~ FEMALE \* US\_MIL, nh\_design, svytotal)

FEMALE US\_MIL one se  
0.0 0 0 37185326.4 1225990.7  
1.0 1 0 40151728.1 1192408.4  
0.1 0 1 2509429.8 419477.5  
1.1 1 1 349624.1 157476.1

## Use survey design for weighted means

What is the mean of total cholesterol, overall and in groups?

svymean( ~ TOTCHOL, nh\_design, na.rm = TRUE)

mean SE  
TOTCHOL 181.25 1.0172

svyby(~ TOTCHOL, ~ FEMALE, nh\_design, svymean, na.rm = TRUE)

FEMALE TOTCHOL se  
0 0 182.6313 1.515072  
1 1 179.8881 1.359801

svyby(~ TOTCHOL, ~ FEMALE + US\_MIL, nh\_design, svymean, na.rm = TRUE)

FEMALE US\_MIL TOTCHOL se  
0.0 0 0 182.3569 1.575994  
1.0 1 0 180.0248 1.368408  
0.1 0 1 186.6966 5.354835  
1.1 1 1 164.1984 6.535223

## Unweighted Summaries of TOTCHOL

favstats(~ TOTCHOL, data = nh1314) |>   
 mutate(se = sd / sqrt(n)) |>  
 gt() |> fmt\_number(columns = c(mean, sd, se), decimals = 3) |>  
 tab\_options(table.font.size = 20)

| min | Q1 | median | Q3 | max | mean | sd | n | missing | se |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 69 | 156 | 178 | 203 | 417 | 181.012 | 37.408 | 1801 | 0 | 0.881 |

nh1314 |> group\_by(FEMALE, US\_MIL) |>  
 summarise(n = n(), mean = mean(TOTCHOL), se = sd(TOTCHOL)/sqrt(n)) |> gt()

`summarise()` has grouped output by 'FEMALE'. You can override using the  
`.groups` argument.

| US\_MIL | n | mean | se |
| --- | --- | --- | --- |
| 0 | | | |
| 0 | 829 | 182.2159 | 1.332123 |
| 1 | 45 | 187.1111 | 5.466170 |
| 1 | | | |
| 0 | 921 | 179.7058 | 1.207026 |
| 1 | 6 | 169.5000 | 7.868714 |

## Survey-Weighted Measures of uncertainty

Mean of total cholesterol within groups with 90% CI?

grouped\_result <- svyby(~ TOTCHOL, ~ FEMALE + US\_MIL,   
 nh\_design, svymean, na.rm = TRUE)  
coef(grouped\_result)

0.0 1.0 0.1 1.1   
182.3569 180.0248 186.6966 164.1984

confint(grouped\_result, level = 0.90)

5 % 95 %  
0.0 179.7646 184.9492  
1.0 177.7739 182.2756  
0.1 177.8887 195.5045  
1.1 153.4489 174.9478

* Get standard errors with se(grouped\_result), too.

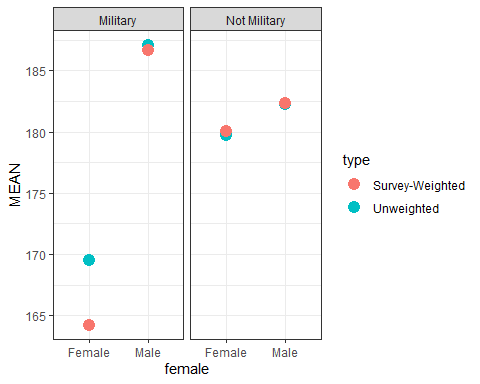
## Store estimated means in res

res <- tibble(  
 type = rep(c("Unweighted", "Survey-Weighted"),4),  
 female = c(rep("Female", 4), rep("Male", 4)),  
 us\_mil = rep(c("Military", "Military", "Not Military", "Not Military"), 2),  
 MEAN = c(169.5, 164.198, 179.706, 180.025,   
 187.111, 186.697, 182.220, 182.357))  
res |> gt()

| type | female | us\_mil | MEAN |
| --- | --- | --- | --- |
| Unweighted | Female | Military | 169.500 |
| Survey-Weighted | Female | Military | 164.198 |
| Unweighted | Female | Not Military | 179.706 |
| Survey-Weighted | Female | Not Military | 180.025 |
| Unweighted | Male | Military | 187.111 |
| Survey-Weighted | Male | Military | 186.697 |
| Unweighted | Male | Not Military | 182.220 |
| Survey-Weighted | Male | Not Military | 182.357 |

## Estimated Means, plotted

ggplot(res, aes(x = female, y = MEAN, col = type)) +  
 geom\_point(size = 4) +  
 facet\_wrap(~ us\_mil)



# Building Models and Survey Weights

## Modeling TOTCHOL in nh1314

First, we’ll ignore weighting, and fit a model with main effects of all three predictors (mod1), then a model (mod2) which incorporates an interaction of FEMALE and US\_MIL.

mod1 <- lm(TOTCHOL ~ AGE + FEMALE + US\_MIL, data = nh1314)  
  
mod2 <- lm(TOTCHOL ~ AGE + FEMALE \* US\_MIL, data = nh1314)

The interaction term means that the effect of FEMALE on TOTCHOL depends on the US\_MIL status.

## mod1, unweighted

tidy(mod1, conf.int = TRUE, conf.level = 0.90) |>  
 select(-statistic) |> gt() |> tab\_options(table.font.size = 20)

| term | estimate | std.error | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 136.345657 | 4.4915861 | 9.534649e-164 | 128.953845 | 143.7374695 |
| AGE | 1.571367 | 0.1474222 | 9.149426e-26 | 1.328754 | 1.8139805 |
| FEMALE | -3.312433 | 1.7274350 | 5.532719e-02 | -6.155276 | -0.4695898 |
| US\_MIL | 2.003854 | 5.2026231 | 7.001628e-01 | -6.558113 | 10.5658215 |

glance(mod1) |> select(r2 = r.squared, adjr2 = adj.r.squared, AIC, BIC,  
 sigma, nobs, df) |> gt() |> tab\_options(table.font.size = 20)

| r2 | adjr2 | AIC | BIC | sigma | nobs | df |
| --- | --- | --- | --- | --- | --- | --- |
| 0.06097646 | 0.0594088 | 18052.71 | 18080.19 | 36.27952 | 1801 | 3 |

## mod1, unweighted

* Now using functions from the easystats framework…

model\_parameters(mod1, pretty\_names = FALSE, ci = 0.90)

Parameter | Coefficient | SE | 90% CI | t(1797) | p  
----------------------------------------------------------------------  
(Intercept) | 136.35 | 4.49 | [128.95, 143.74] | 30.36 | < .001  
AGE | 1.57 | 0.15 | [ 1.33, 1.81] | 10.66 | < .001  
FEMALE | -3.31 | 1.73 | [ -6.16, -0.47] | -1.92 | 0.055   
US\_MIL | 2.00 | 5.20 | [ -6.56, 10.57] | 0.39 | 0.700

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

model\_performance(mod1)

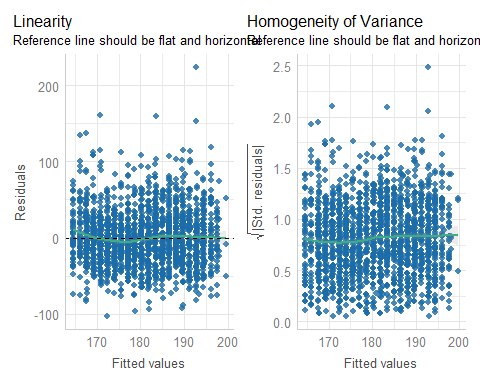
# Indices of model performance  
  
AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma  
-----------------------------------------------------------------------  
18052.707 | 18052.740 | 18080.187 | 0.061 | 0.059 | 36.239 | 36.280

n\_obs(mod1)

[1] 1801

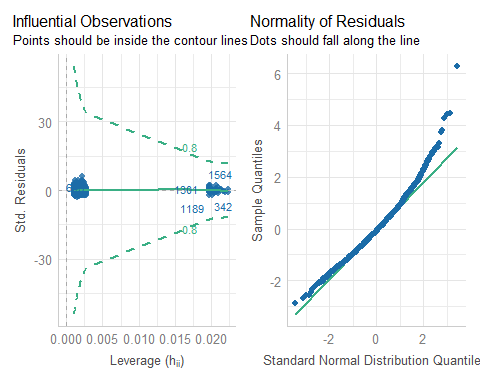
## mod1 Residuals (first 2 plots)

check\_model(mod1, check = c("linearity", "homogeneity"))



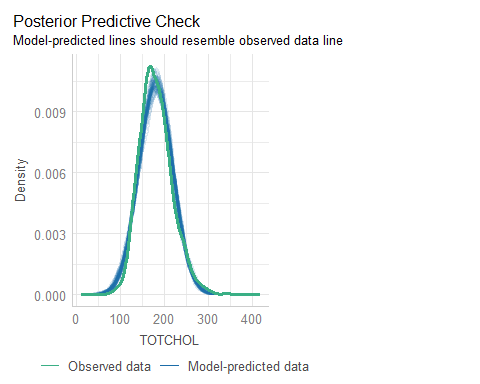
## mod1 Residuals (plots 3-4)

check\_model(mod1, detrend = FALSE, check = c("outliers", "qq"))



## mod1 prediction check

check\_model(mod1, check = "pp\_check")



## mod2, unweighted

tidy(mod2, conf.int = TRUE, conf.level = 0.90) |>  
 select(-statistic) |> gt() |> tab\_options(table.font.size = 20)

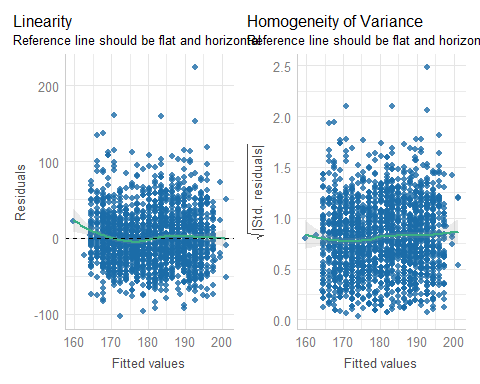
| term | estimate | std.error | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 136.299221 | 4.4922913 | 1.340759e-163 | 128.906246 | 143.6921962 |
| AGE | 1.570077 | 0.1474422 | 1.015235e-25 | 1.327431 | 1.8127229 |
| FEMALE | -3.151959 | 1.7380814 | 6.992615e-02 | -6.012324 | -0.2915943 |
| US\_MIL | 3.639800 | 5.5547809 | 5.123873e-01 | -5.501717 | 12.7813170 |
| FEMALE:US\_MIL | -13.342653 | 15.8650968 | 4.004561e-01 | -39.451882 | 12.7665766 |

glance(mod2) |>   
 select(r2 = r.squared, adjr2 = adj.r.squared, AIC, BIC, sigma,   
 nobs, df) |> gt() |> tab\_options(table.font.size = 20)

| r2 | adjr2 | AIC | BIC | sigma | nobs | df |
| --- | --- | --- | --- | --- | --- | --- |
| 0.06134611 | 0.05925557 | 18054 | 18086.97 | 36.28248 | 1801 | 4 |

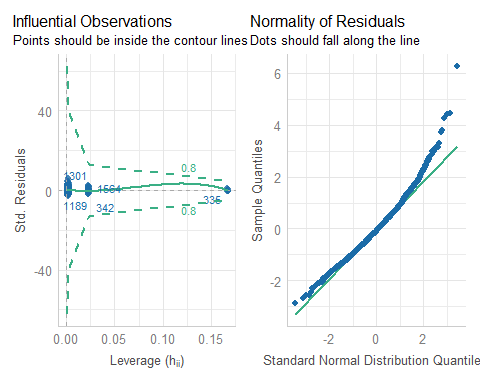
## mod2 Residuals (first 2 plots)

check\_model(mod2, check = c("linearity", "homogeneity"))



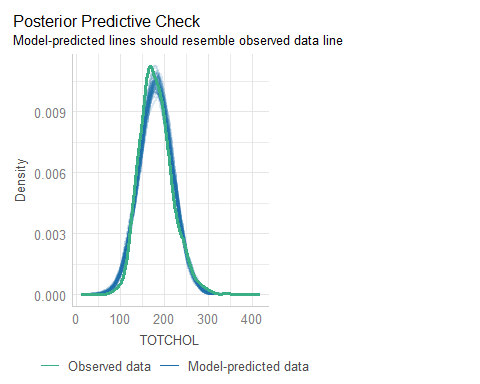
## mod2 Residuals (plots 3-4)

check\_model(mod2, detrend = FALSE, check = c("outliers", "qq"))



## mod2 prediction check

check\_model(mod2, check = "pp\_check")



## Survey-weighted models via svyglm

Again, we’ll run two models, first without and second with an interaction term between FEMALE and US\_MIL.

glm1\_results <- svyglm(TOTCHOL ~ AGE + FEMALE + US\_MIL,   
 nh\_design, family = gaussian())

glm2\_results <- svyglm(TOTCHOL ~ AGE + FEMALE \* US\_MIL,   
 nh\_design, family = gaussian())

Gaussian family used to generate linear regressions here.

## Weighted Model 1

tidy(glm1\_results, conf.int = TRUE, conf.level = 0.90) |>  
 select(-statistic) |> gt() |> tab\_options(table.font.size = 20)

| term | estimate | std.error | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 137.1292664 | 5.0039123 | 1.906826e-138 | 128.894318 | 145.36421489 |
| AGE | 1.5646634 | 0.1696597 | 7.889576e-20 | 1.285454 | 1.84387266 |
| FEMALE | -3.2123089 | 2.0091506 | 1.100321e-01 | -6.518772 | 0.09415432 |
| US\_MIL | 0.5935502 | 5.0392343 | 9.062506e-01 | -7.699528 | 8.88662816 |

glance(glm1\_results) |> select(nobs, AIC, BIC, everything()) |>   
 gt() |> tab\_options(table.font.size = 20)

| nobs | AIC | BIC | null.deviance | df.null | deviance | df.residual |
| --- | --- | --- | --- | --- | --- | --- |
| 1801 | 18036.35 | 2344965 | 2498023 | 1800 | 2344935 | 1797 |

## Weighted Model 2

tidy(glm2\_results, conf.int = TRUE, conf.level = 0.90) |>  
 select(-statistic) |> gt() |> tab\_options(table.font.size = 20)

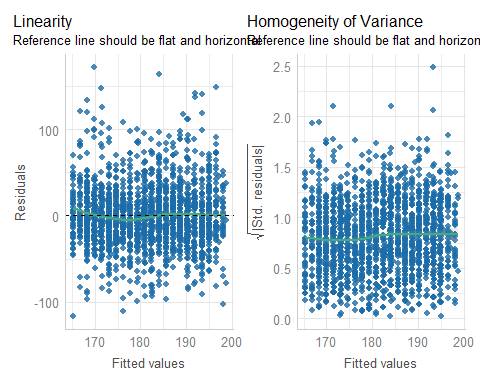
| term | estimate | std.error | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 136.863878 | 5.0060799 | 6.872607e-138 | 128.625359 | 145.1023957 |
| AGE | 1.567633 | 0.1695865 | 6.517529e-20 | 1.288544 | 1.8467218 |
| FEMALE | -2.868135 | 2.0285450 | 1.575681e-01 | -6.206517 | 0.4702464 |
| US\_MIL | 3.426681 | 5.4709976 | 5.311744e-01 | -5.576953 | 12.4303155 |
| FEMALE:US\_MIL | -22.065349 | 8.5522325 | 9.956850e-03 | -36.139779 | -7.9909185 |

glance(glm2\_results) |> select(nobs, AIC, BIC, everything()) |>  
 gt() |> tab\_options(table.font.size = 20)

| nobs | AIC | BIC | null.deviance | df.null | deviance | df.residual |
| --- | --- | --- | --- | --- | --- | --- |
| 1801 | 18034.39 | 2341671 | 2498023 | 1800 | 2341633 | 1796 |

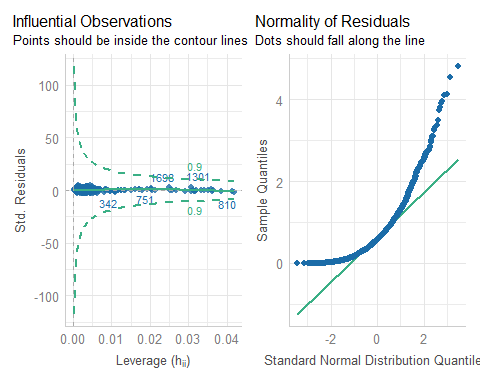
## glm1\_results Residuals (first 2 plots)

check\_model(glm1\_results, check = c("linearity", "homogeneity"))



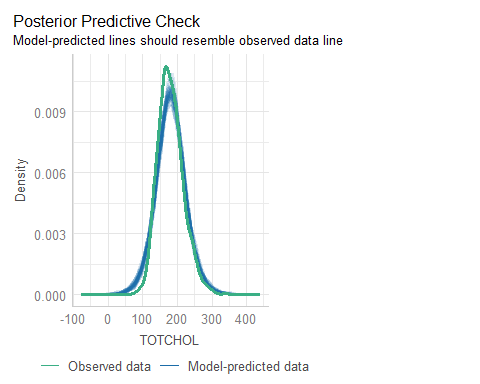
## glm1\_results Residuals (plots 3-4)

check\_model(glm1\_results, detrend = FALSE, check = c("outliers", "qq"))



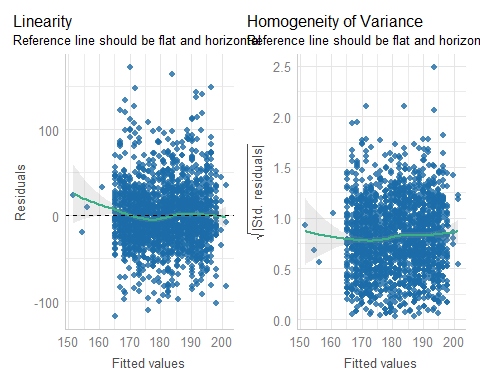
## glm1\_results prediction check

check\_model(glm1\_results, check = "pp\_check")



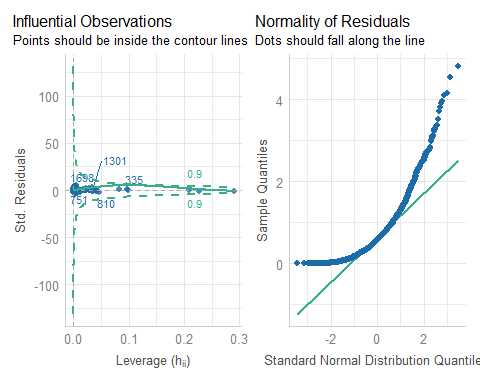
## glm2\_results Residuals (first 2 plots)

check\_model(glm2\_results, check = c("linearity", "homogeneity"))



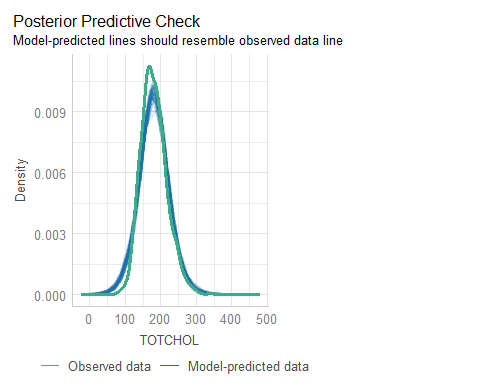
## glm2\_results Residuals (plots 3-4)

check\_model(glm2\_results, detrend = FALSE, check = c("outliers", "qq"))



## glm2\_results prediction check

check\_model(glm2\_results, check = "pp\_check")



## See Lab 5 Question 2

Using the nh\_3143 data you’ve used before, …

Estimate the percentage of the US non-institutionalized adult population within the ages of 30-59 who have smoked at least 100 cigarettes in their life that would describe their General Health as either “Excellent” or “Very Good”.

and you’ll do this without and then with sampling weights; due on Wednesday.

# A More Complete Weighted NHANES Analysis

## New Question, New Data

Key Source: <https://wwwn.cdc.gov/nchs/data/tutorials/DB303_Fig1_R.R>

Now, we are looking at the percentage of persons aged 20 and over with depression, by age and sex, in the US in 2013-2016. Pull in data using nhanesA…

DEMO\_H <- nhanes('DEMO\_H', translated = FALSE) |>   
 select(SEQN, RIAGENDR, RIDAGEYR, SDMVSTRA, SDMVPSU, WTMEC2YR)  
DEMO\_I <- nhanes('DEMO\_I', translated = FALSE) |>  
 select(SEQN, RIAGENDR, RIDAGEYR, SDMVSTRA, SDMVPSU, WTMEC2YR)  
DEMO <- bind\_rows(DEMO\_H, DEMO\_I)   
DPQ\_H <- nhanes('DPQ\_H', translated = FALSE)   
DPQ\_I <- nhanes('DPQ\_I', translated = FALSE)   
DPQ <- bind\_rows(DPQ\_H, DPQ\_I)

## Merge DEMO & DPQ, create derived variables

nhanes2 <- left\_join(DEMO, DPQ, by = "SEQN") |> tibble() |>  
 # Set 7=Refused and 9=Don't Know To NA   
 mutate(across(.cols = DPQ010:DPQ090,   
 ~ ifelse(. >=7, NA, .))) %>%  
 mutate(one = 1,  
 PHQ9\_score = rowSums(select(. , DPQ010:DPQ090)),  
 Depression = ifelse(PHQ9\_score >= 10, 100, 0),  
 Sex = factor(RIAGENDR, labels = c("M", "F")),  
 Age\_group = cut(RIDAGEYR,   
 breaks = c(-Inf, 19, 39, 59, Inf),  
 labels = c("Under 20", "20-39", "40-59", "60+")),  
 WTMEC4YR = WTMEC2YR/2,  
 inAnalysis = (RIDAGEYR >= 20 & !is.na(PHQ9\_score))) |>  
 select(-starts\_with("DPQ"))  
  
write\_rds(nhanes2, here("c12/data/nhanes\_class12\_data2.Rds"))

* nhanes\_class12\_data2.Rds file is on our 432-data page.

## Define The Survey Design

Here’s the survey design for the overall data set:

NH\_des\_all <- svydesign(data = nhanes2, id = ~ SDMVPSU,   
 strata = ~ SDMVSTRA, weights = ~ WTMEC4YR, nest = TRUE)  
  
dim(NH\_des\_all)

[1] 20146 13

Here’s the survey design object for the subset of interest: adults aged 20 and over with a valid PHQ-9 depression score:

NH\_des\_2 <- NH\_des\_all |> subset(inAnalysis)  
  
dim(NH\_des\_2)

[1] 9942 13

## Define a function to call svymean and unweighted count

ourSummary <- function(varformula, byformula, design){  
 # Get mean, stderr, and unweighted sample size  
 c <- svyby(varformula, byformula, design, unwtd.count )   
 p <- svyby(varformula, byformula, design, svymean )   
 outSum <- left\_join(select(c,-se), p)   
 outSum  
}

### Estimate overall prevalence of depression

ourSummary(~ Depression, ~ one, NH\_des\_2)

Joining with `by = join\_by(one)`

one counts Depression se  
1 1 9942 8.056844 0.3599894

## Estimate prevalence of depression in various strata

## By sex  
ourSummary(~ Depression, ~ Sex, NH\_des\_2)

Joining with `by = join\_by(Sex)`

Sex counts Depression se  
1 M 4821 5.549344 0.4293217  
2 F 5121 10.427654 0.5658239

## By age  
ourSummary(~ Depression, ~ Age\_group, NH\_des\_2)

Joining with `by = join\_by(Age\_group)`

Age\_group counts Depression se  
1 20-39 3328 7.744613 0.5236944  
2 40-59 3307 8.429826 0.6164284  
3 60+ 3307 7.971216 0.7797954

## Estimate prevalence of depression by Age and Sex

## By sex and age  
ourSummary(~ Depression, ~ Sex + Age\_group, NH\_des\_2)

Joining with `by = join\_by(Sex, Age\_group)`

Sex Age\_group counts Depression se  
1 M 20-39 1654 5.513778 0.6461045  
2 F 20-39 1674 10.050321 0.8036891  
3 M 40-59 1556 5.222060 0.7699895  
4 F 40-59 1751 11.477238 1.2011361  
5 M 60+ 1611 6.052782 0.8295114  
6 F 60+ 1696 9.579923 1.0534115

## Compare Prevalence between Male and Female

Across all age groups:

svyttest(Depression ~ Sex, NH\_des\_2)

Design-based t-test  
  
data: Depression ~ Sex  
t = 6.8246, df = 29, p-value = 1.706e-07  
alternative hypothesis: true difference in mean is not equal to 0  
95 percent confidence interval:  
 3.416354 6.340267  
sample estimates:  
difference in mean   
 4.87831

## Compare Prevalence between Male and Female

In people ages 40-59:

svyttest(Depression ~ Sex, subset(NH\_des\_2, Age\_group == "40-59"))

Design-based t-test  
  
data: Depression ~ Sex  
t = 3.8688, df = 29, p-value = 0.0005706  
alternative hypothesis: true difference in mean is not equal to 0  
95 percent confidence interval:  
 2.948407 9.561949  
sample estimates:  
difference in mean   
 6.255178

## Differences by Age Group, among Adults

svyttest(Depression ~ Age\_group, subset(NH\_des\_2,   
 Age\_group=="20-39" | Age\_group=="40-59"))

Design-based t-test  
  
data: Depression ~ Age\_group  
t = 0.79398, df = 29, p-value = 0.4337  
alternative hypothesis: true difference in mean is not equal to 0  
95 percent confidence interval:  
 -1.079836 2.450262  
sample estimates:  
difference in mean   
 0.6852129