432 Class 03

https://thomaselove.github.io/432-2024/

2024-01-23

Today's Agenda

Incorporating Survey Weights ...

1. In estimating means and confidence intervals
2. In building linear regression models
3. Into a more detailed t-test approach using NHANES

Primary Source:
https://bookdown.org/rwnahhas/RMPH/survey-design.html

Today's R Setup

```
knitr::opts_chunk$set(comment = NA)

library(broom)
library(janitor)
library(gtsummary)
library(gt)
library(mosaic)

library(nhanesA) # data source
library(survey) # survey-specific tools

library(tidyverse)

theme_set(theme_bw())
```

Section 1

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.

$$\text{Sample Subject's Weight} = \frac{1}{Prob(selection)}$$

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.

```r
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

```r
dim(demo_raw)
```

```
[1] 10175    47
```

```r
dim(tchol_raw)
```

```
[1] 8291    3
```

```r
joined_df <- inner_join(demo_raw, tchol_raw, by = c("SEQN"))

dim(joined_df)
```

```
[1] 8291   49
```

Create a small analytic tibble

```r
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)
```

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()

```r
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")
```

Approach A

Data from nh1314 sample, unadjusted

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

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")
```

Approach B

Data from nh1314 sample, unadjusted

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

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	
8430.461	24694.05	34642.05	59560.74	125680.3	44528.66	26027

Using gtsummary() to describe nh1314 (unweighted)

```
table1 <- nh1314 |>
  tbl_summary(include = -SEQN)

table1
```

Characteristic	**N = 1,801**
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)

See https://www.danieldsjoberg.com/gtsummary/ for more options.

Create nh_design survey design

```r
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
```

Weighted Counts

```r
svytotal( ~ one, nh_design )
```

```
       total      SE
one 80196108 1104558
```

By Groups

```r
svyby( ~ one, ~ FEMALE, nh_design, svytotal)
```

```
  FEMALE      one      se
0      0 39694756 1255122
1      1 40501352 1196260
```

```r
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
```

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
```

```
    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 5.254835
```

Unweighted Summaries of TOTCHOL

```r
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	
69	156	178	203	417	181.012	37.408	1801	0	0.8

```r
nh1314 |> group_by(FEMALE, US_MIL) |>
  summarise(n = n(), mean = mean(TOTCHOL), se = sd(TOTCHOL)/sq
```

```
# A tibble: 4 x 5
# Groups:   FEMALE [2]
  FEMALE US_MIL     n  mean    se
   <dbl>  <dbl> <int> <dbl> <dbl>
1      0      0   829  182.  1.33
```

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

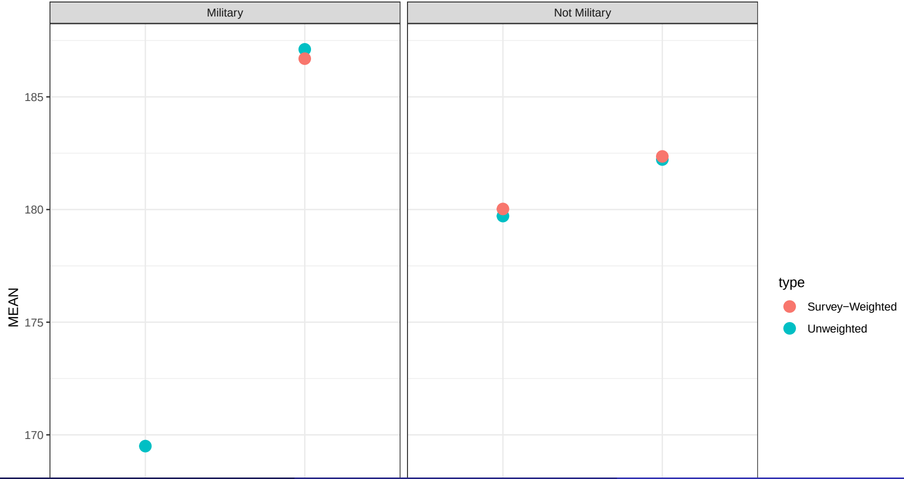
```r
res <- tibble(
  type = rep(c("Unweighted", "Survey-Weighted"),4),
  female = c(rep("Female", 4), rep("Male", 4)),
  us_mil = rep(c("Military", "Military", "Not Military", "Not
  MEAN = c(169.5, 164.1984, 179.71, 180.0248, 187.11, 186.696

res |> gt()
```

type	female	us_mil	MEAN
Unweighted	Female	Military	169.5000
Survey-Weighted	Female	Military	164.1984
Unweighted	Female	Not Military	179.7100
Survey-Weighted	Female	Not Military	180.0248
Unweighted	Male	Military	187.1100
Survey-Weighted	Male	Military	186.6966
Unweighted	Male	Not Military	182.2200
Survey-Weighted	Male	Not Military	182.3569

Estimated Means, plotted

```
ggplot(res, aes(x = female, y = MEAN, col = type)) +
  geom_point(size = 4) +
  facet_wrap(~ us_mil)
```



Section 2

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.

```r
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 =
```

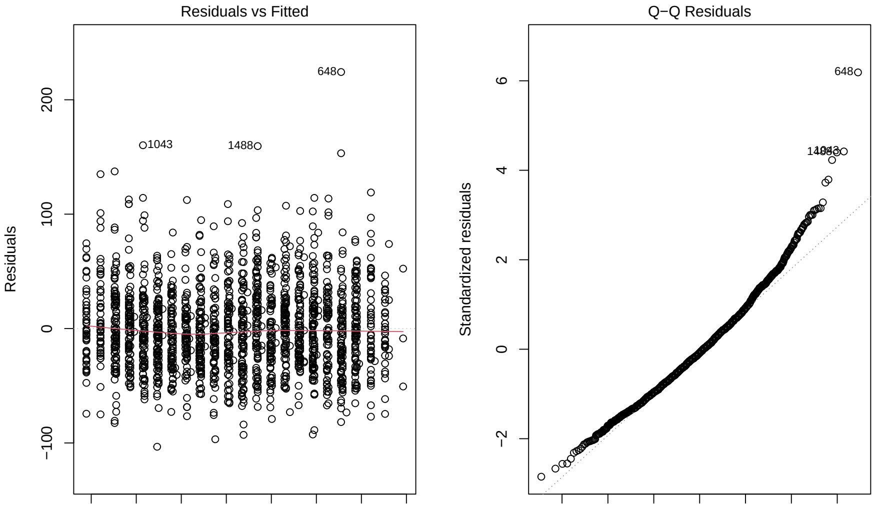
term	estimate	std.error	p.value	conf.low	c
(Intercept)	136.345657	4.4915861	9.534649e-164	128.953845	143.
AGE	1.571367	0.1474222	9.149426e-26	1.328754	1.
FEMALE	-3.312433	1.7274350	5.532719e-02	-6.155276	-0.
US_MIL	2.003854	5.2026231	7.001628e-01	-6.558113	10.

```
glance(mod1) |> select(r2 = r.squared, adjr2 = adj.r.squared,
    sigma, nobs, df) |> gt() |> tab_options(table.font.size =
```

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

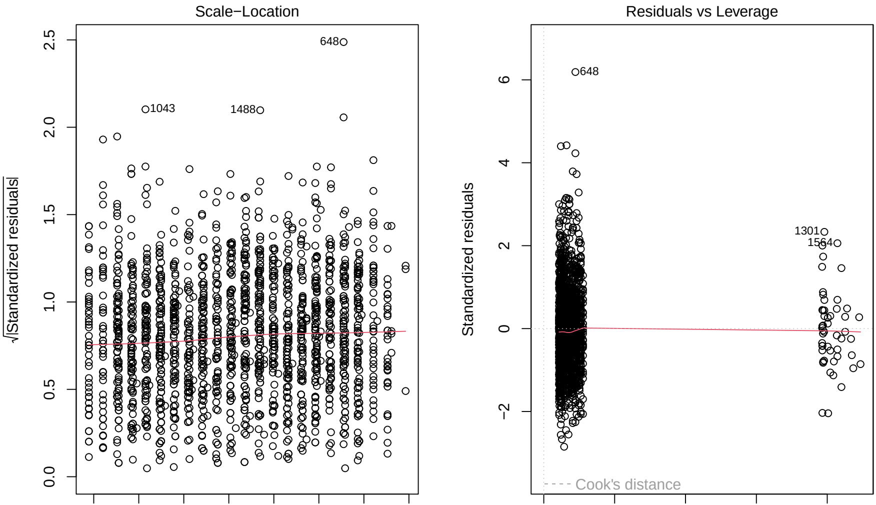
mod1 Residuals (plots 1, 2)

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



mod1 Residuals (plots 3, 5)

```
par(mfrow = c(1,2)); plot(mod1, which = c(3,5))
```



mod2, unweighted

```
tidy(mod2, conf.int = TRUE, conf.level = 0.90) |>
  select(-statistic) |> gt() |> tab_options(table.font.size =
```

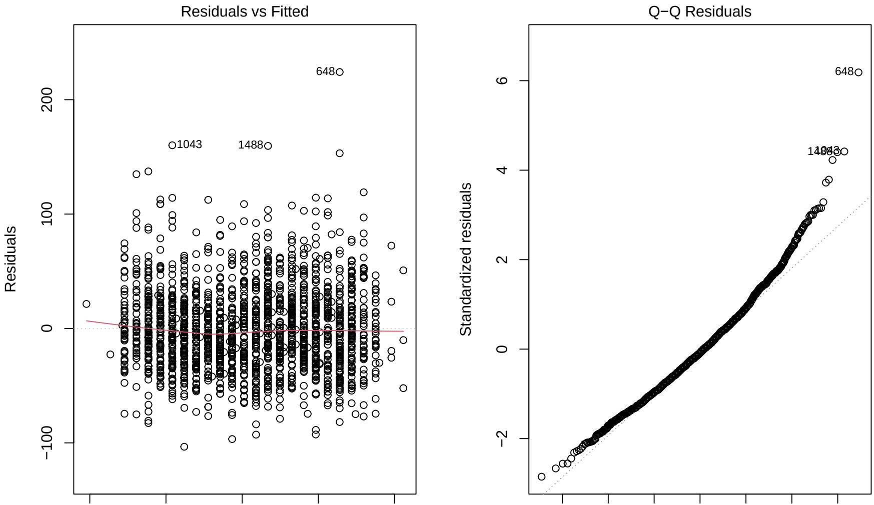
term	estimate	std.error	p.value	conf.lo
(Intercept)	136.299221	4.4922913	1.340759e-163	128.90624
AGE	1.570077	0.1474422	1.015235e-25	1.32743
FEMALE	-3.151959	1.7380814	6.992615e-02	-6.01232
US_MIL	3.639800	5.5547809	5.123873e-01	-5.50171
FEMALE:US_MIL	-13.342653	15.8650968	4.004561e-01	-39.45188

```
glance(mod2) |>
  select(r2 = r.squared, adjr2 = adj.r.squared, AIC, BIC, sigm
         nobs, df) |> gt() |> tab_options(table.font.size = 20
```

r2	adjr2	AIC	BIC	sigma	nobs	df
0.06134611	0.05025557	18054	18086.97	36.28248	1801	4

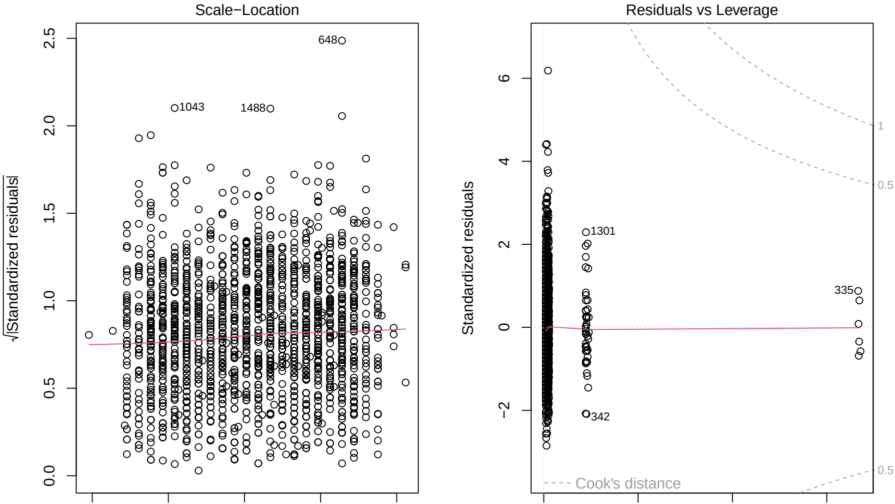
mod2 Residuals (plots 1, 2)

```r
par(mfrow = c(1,2)); plot(mod2, which = c(1,2))
```



mod2 Residuals (plots 3, 5)

```
par(mfrow = c(1,2)); plot(mod2, which = c(3,5))
```



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 =
```

term	estimate	std.error	p.value	conf.low	
(Intercept)	137.1292664	5.0039123	1.906826e-138	128.894318	145
AGE	1.5646634	0.1696597	7.889576e-20	1.285454	1
FEMALE	-3.2123089	2.0091506	1.100321e-01	-6.518772	0
US_MIL	0.5935502	5.0392343	9.062506e-01	-7.699528	8

```
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	21.6033	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 =
```

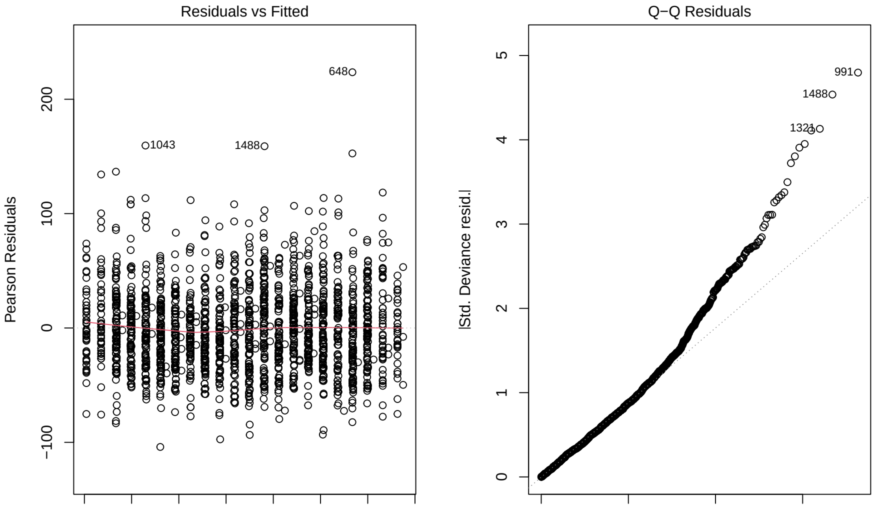
term	estimate	std.error	p.value	conf.low
(Intercept)	136.863878	5.0060799	6.872607e-138	128.62535
AGE	1.567633	0.1695865	6.517529e-20	1.28854
FEMALE	-2.868135	2.0285450	1.575681e-01	-6.20651
US_MIL	3.426681	5.4709976	5.311744e-01	-5.57695
FEMALE:US_MIL	-22.065349	8.5522325	9.956850e-03	-36.13977

```
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	22.19935	2341671	2498023	1800	2341633	1796

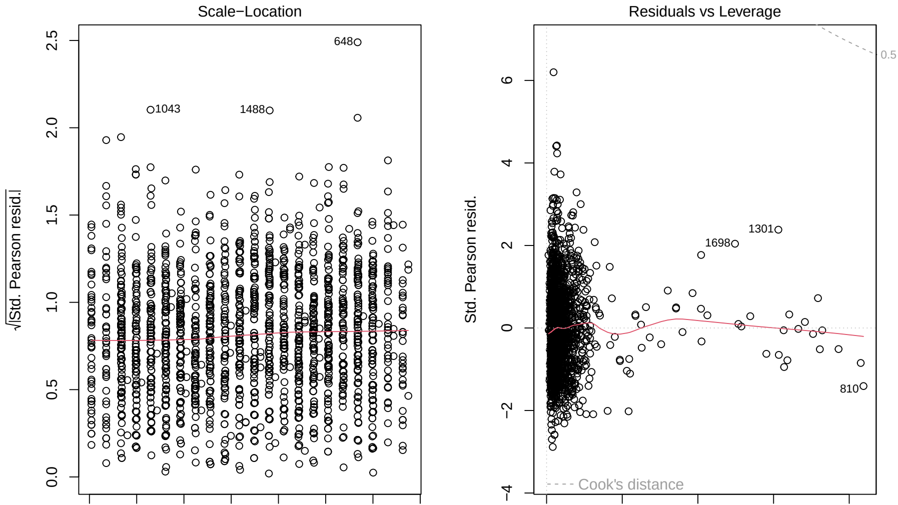
Residuals for Model glm1_results

```r
par(mfrow = c(1,2)); plot(glm1_results, which = c(1,2))
```

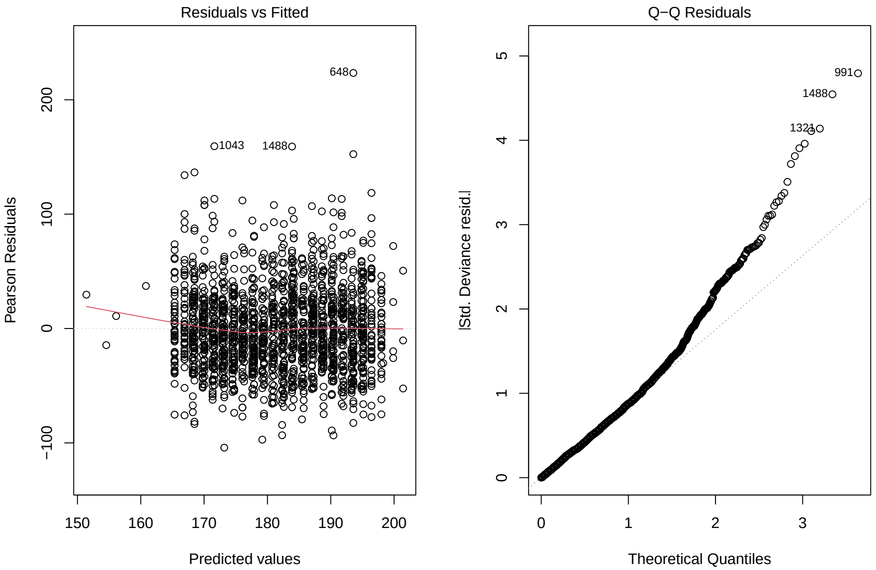


Residuals for Model `glm1_results`

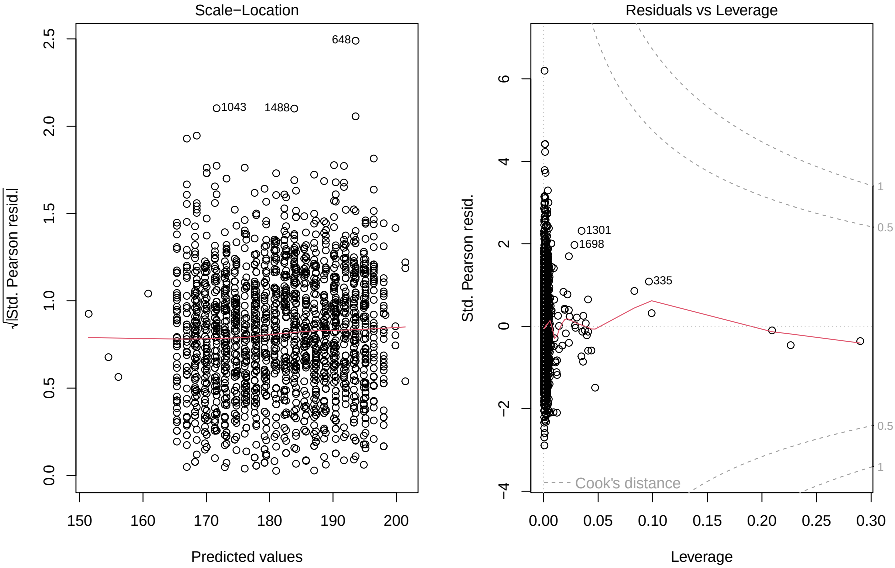
```
par(mfrow = c(1,2)); plot(glm1_results, which = c(3,5))
```



Residuals for Model `glm2_res`



Residuals for Model `glm2_res`



Section 3

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 and DPQ files and create derived variables

```
dat2 <- 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"))
```

Define Survey Design

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

```
NH_des_all <- svydesign(data = dat2, 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_dat2 <- NH_des_all |> subset(inAnalysis)

dim(NH_des_dat2)
```

```
[1] 9942   13
```

Define a function to call svymean and unweighted count

```r
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

```r
ourSummary(~ Depression, ~ one, NH_des_dat2)
```

```
  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_dat2)
```

```
  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_dat2)
```

```
  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

```r
## By sex and age
ourSummary(~ Depression, ~ Sex + Age_group, NH_des_dat2)
```

```
  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:

```r
svyttest(Depression ~ Sex, NH_des_dat2)
```

```

	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 t
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_dat2, Age_group == "4
```

```

	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 t
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_dat2,
                  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 t
95 percent confidence interval:
 -1.079836  2.450262
sample estimates:
difference in mean
         0.6852129
```

Next Time?

- Linear Regression and ANOVA / ANCOVA models
- Incorporating Polynomials into our models

Reminders

1. Please complete the **Minute Paper after Class 3** by noon tomorrow (Wednesday 2024-01-24)
2. Get started on **Lab 2**, due next Tuesday 2024-01-30 at Noon.
3. Continue reading **How To Be A Modern Scientist**