# 432 Class 02

https://thomaselove.github.io/432-2023/

2023-01-19

# Today's Agenda

1. Comparing Means
2. Comparing Rates
3. Fitting Linear Models
4. Setting Up Lab 1, due Monday 2023-01-23 at 9 PM.

The most relevant sections of the Course Notes are Chapters 1-5.

# Today's R Setup

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

library(broom) # for tidy, glance and augment
library(car) # for boxCox and vif
library(Epi) # for twoby2
library(GGally) # for ggpairs
library(knitr) # for kable
library(MKinfer) # for boot.t.test
library(mosaic) # for favstats
library(naniar) # deal with missing values
library(nhanesA) # source of data
library(vcd) # for mosaic (plot) and assoc (plot)
library(janitor) # for tabyl and other things
library(tidyverse) # for all kinds of things

theme_set(theme_bw())
```

Section 1

# Building an NHANES Data Set (see Course Notes Chapters 1-2)

# How I Built Our Data (2017 - March 2020 NHANES)

Variables from P-DEMO:

- SEQN
- RIDAGEYR (age) *restricted to ages 26-42 here*
- DMDEDUC2 (educ)

Variables from BPXO (linked by SEQN):

- BPXOSY1 (sbp1)
- BPXOSY2 (sbp2)
- BPXOSY3 (sbp3)

Variables from HUQ (linked by SEQN)

- HUQ010 (sroh)
- HUQ071 (hospital)
- HUQ090 (mentalh)

Total: 1982 observations on 9 variables: includes all available NHANES participants ages 26-42 with complete data on these nine variables.

# Building the Data (using nhanesA package)

```r
p_demo <- nhanes('P_DEMO') |>
  select(SEQN, RIDAGEYR, DMDEDUC2)

p_bpxo <- nhanes('P_BPXO') |>
  select(SEQN, BPXOSY1, BPXOSY2, BPXOSY3)

p_huq <- nhanes('P_HUQ') |>
  select(SEQN, HUQ010, HUQ071, HUQ090)

df_list <- list(p_demo, p_bpxo, p_huq)

nh_raw <- df_list |>
  reduce(left_join, by = 'SEQN') |>
  drop_na() |>
  filter(RIDAGEYR >= 26 & RIDAGEYR <= 42) |>
  as_tibble()
```

# Renaming and Cleaning Variables (1)

```
nh1982 <- nh_raw |>
  rename(age = RIDAGEYR, educ = DMDEDUC2,
         sbp1 = BPXOSY1, sbp2 = BPXOSY2,
         sbp3 = BPXOSY3, sroh = HUQ010,
         hospital = HUQ071, mentalh = HUQ090) |>
  replace_with_na_at(
    .vars = c("educ", "sroh", "hospital", "mentalh"),
    condition = ~ .x %in% c(7,9)) |>
  mutate(across(c(hospital, mentalh), ~ 2 - .x)) |>
  mutate(mean_sbp = (sbp1 + sbp2 + sbp3)/3,
         SEQN = as.character(SEQN))
```

# Renaming and Cleaning Variables (2)

```
nh1982 <- nh1982 |>
  mutate(educ = fct_recode(factor(educ),
              "Less than 9th Grade" = "1",
              "9th - 11th Grade" = "2",
              "High School Grad" = "3",
              "Some College / AA" = "4",
              "College Grad" = "5")) |>
  mutate(sroh = fct_recode(factor(sroh),
              "Excellent" = "1",
              "Very Good" = "2",
              "Good" = "3",
              "Fair" = "4",
              "Poor" = "5")) |>
  drop_na()

write_rds(nh1982, "c02/data/nh1982.Rds")
```

# nh1982

```
glimpse(nh1982)
```

```
Rows: 1,982
Columns: 10
$ SEQN     <chr> "109266", "109273", "109291", "109297", "1093
$ age      <dbl> 29, 36, 42, 30, 30, 28, 33, 41, 35, 30, 41, 3
$ educ     <fct> College Grad, Some College / AA, College Grad
$ sbp1     <dbl> 99, 116, 107, 105, 118, 110, 110, 106, 162, 1
$ sbp2     <dbl> 99, 110, 111, 105, 123, 110, 105, 107, 148, 1
$ sbp3     <dbl> 99, 115, 107, 102, 125, 110, 108, 113, 163, 1
$ sroh     <fct> Good, Good, Fair, Very Good, Good, Very Good,
$ hospital <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ mentalh  <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0,
$ mean_sbp <dbl> 99.00000, 113.66667, 108.33333, 104.00000, 12
```

# Codebook (excerpt, without SEQN)

| Variable | Description (n = 1982) |
|---|---|
| age | Age in years (range 26-42, mean = 34) |
| meansbp | Mean of sbp1, sbp2, sbp3 in mm Hg (range: 76 to 209, mean 116): we'll also use sbp1, sbp2 and sbp3. |
| hospital | 1 if hospitalized in last 12m, else 0 (8% are 1) |
| mentalh | 1 if saw a mental health professional in last 12m, else 0 (12% are 1) |
| sroh | Self-reported Overall Health (5 levels: see next slide) |
| educ | Educational Attainment (5 levels: see next slide) |

# SROH and Educational Attainment

```
nh1982 |> tabyl(sroh) |> adorn_pct_formatting()
```

```
      sroh   n percent
 Excellent 294   14.8%
 Very Good 598   30.2%
      Good 728   36.7%
      Fair 321   16.2%
      Poor  41    2.1%
```

```
nh1982 |> tabyl(educ) |> adorn_pct_formatting()
```

```
                 educ   n percent
 Less than 9th Grade  90    4.5%
    9th - 11th Grade 209   10.5%
    High School Grad 418   21.1%
   Some College / AA 677   34.2%
        College Grad 588   29.7%
```

# Ingesting the Data (from .Rds)

If you don't want to work through the nhanesA import and tidying, you can simply work with the nh1982.Rds file provided on our 432-data page.

```
nh1982 <- read_rds("c02/data/nh1982.Rds")

## not run here...
```

# Section 2

## Comparing Means (see Course Notes Chapter 3)

# Paired or Independent Samples?

In Analysis 1, we will compare the means of SBP1 and SBP2 for our 1982 participants.

In Analysis 2, we will compare the mean of SBP3 between our 159 participants who were hospitalized and the 1823 who were not?

Which of these analyses uses paired samples, and why?

# Paired Samples Analysis

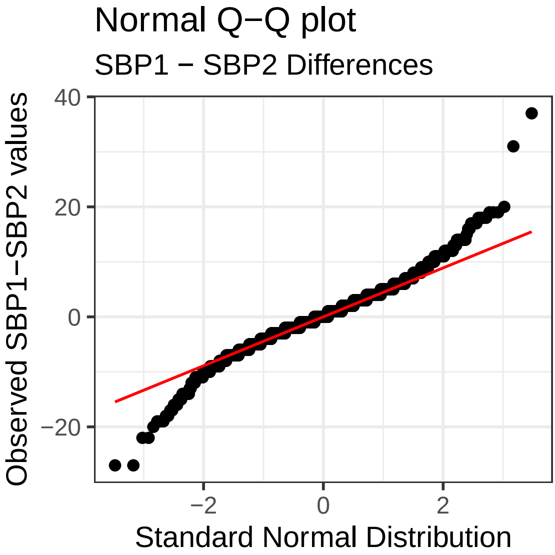
```
nh1982 <- nh1982 |> mutate(SBP_diff = sbp1 - sbp2)

favstats(~ SBP_diff, data = nh1982)
```

```
 min Q1 median Q3 max      mean       sd    n missing
 -27 -3      0  3  37 0.2482341 5.279749 1982       0
```

```
ggplot(nh1982, aes(sample = SBP_diff)) +
  geom_qq() + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q plot",
       subtitle = "SBP1 - SBP2 Differences",
       x = "Standard Normal Distribution",
       y = "Observed SBP1-SBP2 values")
```

# Normal Q-Q plot of Paired SBP Differences



# Comparing Paired Samples

Want a 90% confidence interval for the true mean of the paired SBP1 - SBP2 differences:

- t-based approach (equivalent to linear model) assumes Normality
- Wilcoxon signed rank approach doesn't assume Normality but makes inferences about the pseudo-median, not the mean
- bootstrap doesn't assume Normality, and describes the mean

```
set.seed(20230117)
boot.t.test(nh1982$SBP_diff, conf.level = 0.9,
            boot = TRUE, R = 999)
```

Results on the next slide...

# Bootstrap for Mean of SBP1-SBP2 Differences

```
    Bootstrap One Sample t-test

data:  nh1982$SBP_diff
bootstrap p-value = 0.05205
bootstrap mean of x (SE) = 0.2555684 (0.1184953)
90 percent bootstrap percentile confidence interval:
 0.06609485 0.44611504

Results without bootstrap:
t = 2.0931, df = 1981, p-value = 0.03646
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 0.05307362 0.44339459
sample estimates:
mean of x
0.2482341
```

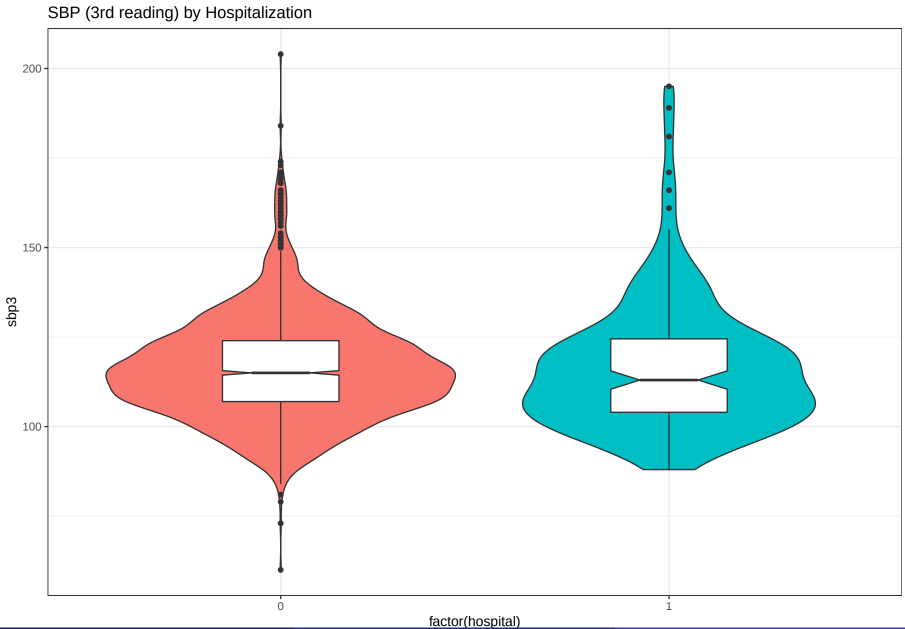
# Comparing sbp3 by hospital: Independent Samples

```
favstats(sbp3 ~ hospital, data = nh1982) |>
  select(-missing) |>
  kable(digits = 2)
```

| hospital | min | Q1 | median | Q3 | max | mean | sd | n |
|---|---|---|---|---|---|---|---|---|
| 0 | 60 | 107 | 115 | 124.0 | 204 | 116.11 | 14.51 | 1823 |
| 1 | 88 | 104 | 113 | 124.5 | 195 | 116.71 | 18.50 | 159 |

```
ggplot(nh1982, aes(x = factor(hospital), y = sbp3)) +
  geom_violin(aes(fill = factor(hospital))) +
  geom_boxplot(width = 0.3, notch = TRUE) +
  guides(fill = "none") +
  labs(title = "SBP (3rd reading) by Hospitalization")
```

# SBP (3rd reading) vs. Hospitalization Status



# Two Independent Samples, Comparing Means

Want a 90% confidence interval for the difference in means of SBP3 for people who were hospitalized - those who were not.

- Pooled t-based approach (equivalent to linear model) assumes Normality and equal population variances
- Welch t-based approach assumes Normality only
- bootstrap assumes neither
- Wilcoxon-Mann-Whitney rank sum assumes neither, but assesses a difference in locations, not the mean

# Pooled t test approach via linear model

```
lm2 <- lm(sbp3 ~ hospital, data = nh1982)

tidy(lm2, conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = 2)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 116.11 | 0.35 | 333.54 | 0.00 | 115.54 | 116.69 |
| hospital | 0.60 | 1.23 | 0.49 | 0.63 | -1.42 | 2.62 |

```
glance(lm2) |> select(r.squared, sigma) |>
  kable(digits = c(5,2))
```

| r.squared | sigma |
|---|---|
| 0.00012 | 14.86 |

# Section 3

## Comparing Rates (see Course Notes, Chapter 4)

# A Two-by-Two Contingency Table

```
nh1982 |> tabyl(mentalh, hospital) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_title()
```

```
         hospital
 mentalh        0   1 Total
       0     1613 122  1735
       1      210  37   247
   Total     1823 159  1982
```

# Standard Epidemiological Format

```
nh1982 <- nh1982 |>
  mutate(mentalh_f = fct_recode(factor(mentalh),
                  "Saw MHP" = "1", "No MHP" = "0"),
         mentalh_f = fct_relevel(mentalh_f,
                  "Saw MHP", "No MHP"),
         hospital_f = fct_recode(factor(hospital),
                  "Hosp." = "1", "No Hosp." = "0"),
         hospital_f = fct_relevel(hospital_f,
                  "Hosp.", "No Hosp."))

nh1982 |> tabyl(mentalh_f, hospital_f)
```

```
 mentalh_f Hosp. No Hosp.
   Saw MHP    37      210
    No MHP   122     1613
```

# Two by Two Table Analysis

```
twoby2(nh1982$mentalh_f, nh1982$hospital_f, conf.level = 0.90)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Hosp.
Comparing : Saw MHP vs. No MHP

        Hosp. No Hosp.    P(Hosp.) 90% conf. interval
Saw MHP    37      210      0.1498    0.1161   0.1911
No MHP    122     1613      0.0703    0.0609   0.0811

                                    90% conf. interval
             Relative Risk: 2.1303    1.5977   2.8405
         Sample Odds Ratio: 2.3295    1.6723   3.2449
Conditional MLE Odds Ratio: 2.3282    1.6287   3.2894
    Probability difference: 0.0795    0.0442   0.1217
```

# A Larger Two-Way Table

What is the association of Educational Attainment with Self-Reported Overall Health?

```
nh1982 |> tabyl(educ, sroh) |>
  adorn_totals(where =c("row","col"))|> adorn_title()
```

```
                          sroh
               educ Excellent Very Good Good Fair Poor Total
 Less than 9th Grade        10         7   36   33    4    90
    9th - 11th Grade        21        40   81   59    8   209
     High School Grad       50        94  168   98    8   418
   Some College / AA        72       220  264  104   17   677
        College Grad       141       237  179   27    4   588
               Total       294       598  728  321   41  1982
```

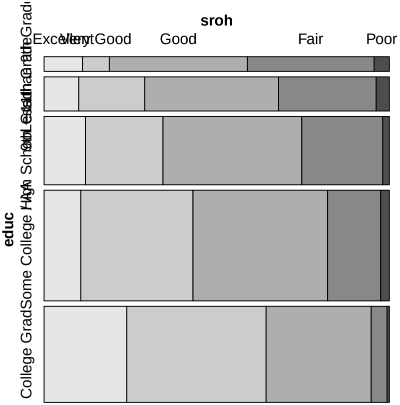
# Our 5x5 Table, showing SROH Proportions

```
nh1982 |> tabyl(educ, sroh) |>
  adorn_totals(where = c("row")) |>
  adorn_percentages(denominator = "row") |>
  adorn_pct_formatting() |> adorn_title()
```

```
                             sroh
                   educ Excellent Very Good  Good  Fair Poor
  Less than 9th Grade     11.1%      7.8% 40.0% 36.7% 4.4%
     9th - 11th Grade     10.0%     19.1% 38.8% 28.2% 3.8%
     High School Grad     12.0%     22.5% 40.2% 23.4% 1.9%
    Some College / AA     10.6%     32.5% 39.0% 15.4% 2.5%
         College Grad     24.0%     40.3% 30.4%  4.6% 0.7%
                Total     14.8%     30.2% 36.7% 16.2% 2.1%
```

# Mosaic Plot for our 5x5 Table

```r
mosaic(~ educ + sroh, data = nh1982, highlighting = "sroh")
```



# Pearson $\chi^2$ test for our 5x5 Table

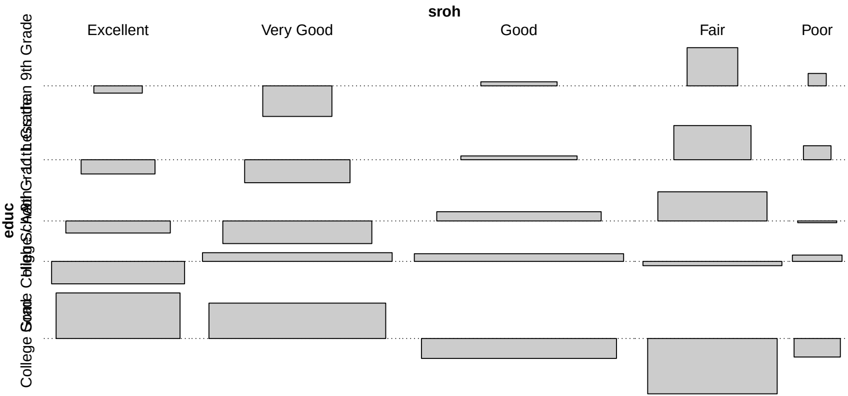
```r
chisq.test(xtabs(~ educ + sroh, data = nh1982))
```

```
    Pearson's Chi-squared test

data:  xtabs(~educ + sroh, data = nh1982)
X-squared = 225.99, df = 16, p-value < 2.2e-16
```

# Association Plot for our 5x5 Table

```r
assoc(~ educ + sroh, data = nh1982)
```



# Section 4

## Fitting Linear Models (see Course Notes, Chapter 5)

# We'll fit two models today

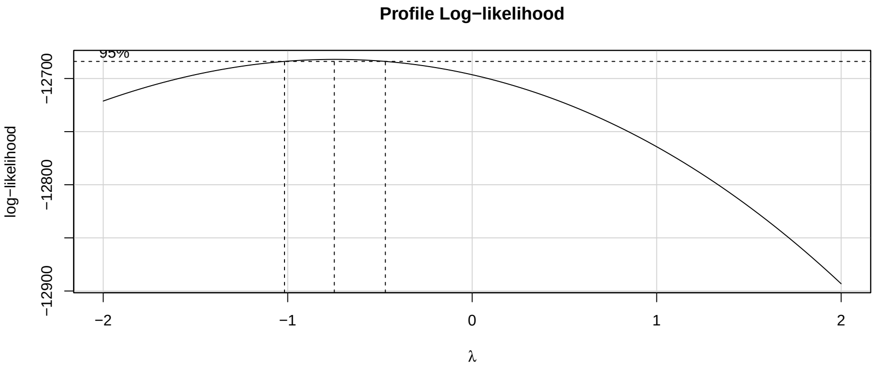
1. Predict mean SBP using Age alone.
2. Predict mean SBP (across three readings) using Age, Self-Reported Overall Health Status and Hospitalization Status.

```
temp_mod1 <- lm(mean_sbp ~ age, data = nh1982)
temp_mod2 <- lm(mean_sbp ~ age + sroh + hospital,
                data = nh1982)
```

Note that I'm not doing any predictive validation today (remember that I did that in Class 1), so I won't split the sample.

# Box-Cox Plot to suggest potential outcome transformations

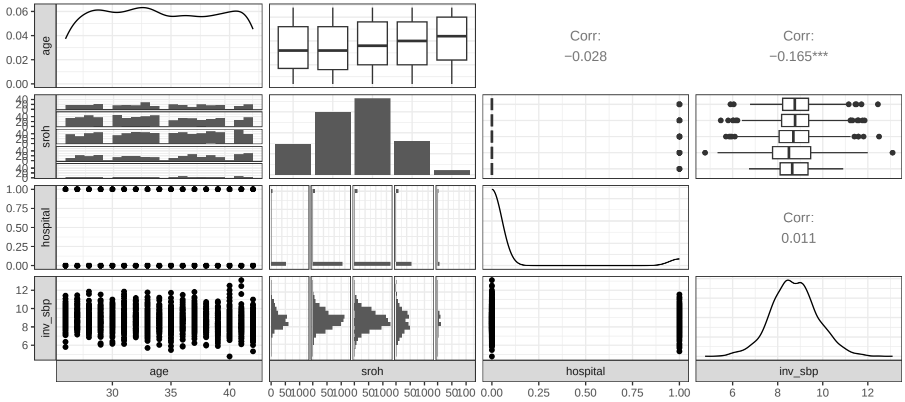
```
boxCox(temp_mod2)
```




```
nh1982 <- nh1982 |> mutate(inv_sbp = 1000/mean_sbp)
```

# Scatterplot Matrix (from ggpairs())

```
ggpairs(nh1982, columns = c(2, 7, 8, 14), switch = "both",
        lower=list(combo=wrap("facethist", bins=20)))
```



# Checking Collinearity: Variance Inflation Factors

```
vif(lm(inv_sbp ~ age + sroh + hospital, data = nh1982))
```

```
             GVIF Df GVIF^(1/(2*Df))
age      1.008723  1        1.004352
sroh     1.020544  4        1.002545
hospital 1.013797  1        1.006875
```

# Tidied Coefficients for Model m1

```
m1 <- lm(inv_sbp ~ age, data = nh1982)

tidy(m1, conf.int = TRUE, conf.level = 0.9) |>
  kable(digits = 2)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 9.93 | 0.16 | 61.52 | 0 | 9.66 | 10.20 |
| age | -0.03 | 0.00 | -7.44 | 0 | -0.04 | -0.03 |

# Tidied Coefficients for Model m2

```r
m2 <- lm(inv_sbp ~ age + sroh + hospital, data = nh1982)

tidy(m2, conf.int = TRUE, conf.level = 0.9) |>
  kable(digits = 2)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 9.99 | 0.17 | 58.76 | 0.00 | 9.71 | 10.27 |
| age | -0.03 | 0.00 | -7.19 | 0.00 | -0.04 | -0.03 |
| srohVery Good | -0.06 | 0.07 | -0.76 | 0.45 | -0.17 | 0.06 |
| srohGood | -0.11 | 0.07 | -1.56 | 0.12 | -0.23 | 0.01 |
| srohFair | -0.27 | 0.08 | -3.21 | 0.00 | -0.40 | -0.13 |
| srohPoor | -0.18 | 0.17 | -1.03 | 0.30 | -0.46 | 0.10 |
| hospital | 0.05 | 0.08 | 0.55 | 0.58 | -0.09 | 0.19 |

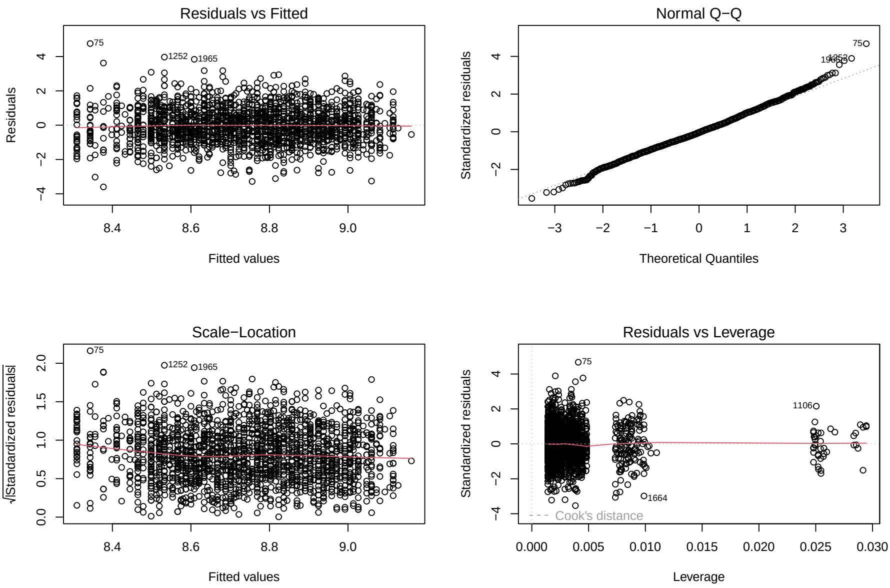
# Fit Summaries for Models m1 and m2

```
bind_rows(glance(m1), glance(m2)) |>
  mutate(model = c("m1", "m2")) |>
  select(model, r2 = r.squared, adjr2 = adj.r.squared,
         sigma, AIC, BIC, nobs, df, df.residual) |>
  kable(digits = c(0, 3, 3, 3, 1, 1, 0, 0))
```

| model | r2 | adjr2 | sigma | AIC | BIC | nobs | df | df.residual |
|---|---|---|---|---|---|---|---|---|
| m1 | 0.027 | 0.027 | 1.022 | 5713.8 | 5730.6 | 1982 | 1 | 1980 |
| m2 | 0.033 | 0.030 | 1.020 | 5711.2 | 5755.9 | 1982 | 6 | 1975 |

Which model appears to fit the data better?

# Residual Plots for Model `m2`



# Making a Prediction in New Data

Suppose a new person is age 29, was not hospitalized, and their SROH is "Good". What is their predicted mean systolic blood pressure?

- Our models predict 1000/mean_sbp and augment places that prediction into `.fitted`.
- To invert, divide `.fitted` by 1000, then take the reciprocal of that result. That's just 1000/`.fitted`.

```
new_person <- tibble(age = 29, sroh = "Good", hospital = 0)
bind_rows(augment(m1, newdata = new_person),
          augment(m2, newdata = new_person)) |>
  mutate(model = c("m1", "m2"), fit_meansbp = 1000/.fitted) |>
  select(model, fit_meansbp, .fitted, age, sroh, hospital) |>
```

| model | fit_meansbp | .fitted | age | sroh | hospital |
|---|---|---|---|---|---|
| m1 | 112.114 | 8.920 | 29 | Good | 0 |
| m2 | 112.309 | 8.904 | 29 | Good | 0 |

## Section 5

# Setting Up Lab 1, due Monday 2023-01-23 at 9 PM

# Lab 1 Question 1

I provide some County Health Rankings data for Ohio's 88 counties. You create a visualization involving information from at least three different variables using R and Quarto.

- Include proper labels and a meaningful title.
- Include a caption (75 words or fewer) that highlights the key result.
- What is the question you are trying to answer with this visualization?

There is a Quarto template for Lab 1, in addition to the data set.

# Lab 1 Question 2

Create a linear regression model to predict obese_pct as a function of food_env and median_income (all of these are quantitative variables.)

a. Specify and fit the model, interpret food_env coefficient.
b. Evaluate quality of model in terms of adherence to regression assumptions via four key residual plots.
c. Build a nice table comparing your model to a simple regression for obese_pct using only food_env, and then reflect on your findings.

# Next Week?

- Lab 1 due Monday 9 PM (Answer Sketch available Tuesday)
- Developing Inferences Using Survey Weights
- Linear Regression and ANOVA/ANCOVA models