432 Class 02

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

2024-01-18

Today's Agenda

1. Comparing Means
2. Comparing Rates
3. Fitting Linear Models
4. Setting Up Lab 1, due Tuesday 2024-01-23 at Noon.

Course Notes: most relevant material in 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(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)

2017 - March 2020 NHANES Data

1982 NHANES subjects ages 26-42 with complete data on these 9 variables:

Variable	Source	Description
SEQN	P-DEMO	Subject ID: Link (also in BPXO and HUQ)
age	P_DEMO	RIDAGEYR (restricted to ages 26-42 here)
educ	P_DEMO	DMDEDUC2 (five-category factor)
sbp1	BPXO	BPXOSY1 = 1st Systolic BP reading, in mm Hg
sbp2	BPXO	BPXOSY2 = 2nd Systolic BP reading
sbp3	BPXO	BPXOSY3 = 3rd Systolic BP reading
sroh	HUQ	HUQ010 = five-categories E, VG, G, F, P
hospital	HUQ	HUQ071 = Yes or No
mentalh	HUQ	HUQ090 = Yes or No

Building the Data (using nhanesA)

```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

```r
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("Refused", "Don't know")) |>
  mutate(educ = fct_recode(educ,
           "Less than 9th Grade" = "Less than 9th grade",
           "9th - 11th Grade" = "9-11th grade (Includes 12th
           "High School Grad" = "High school graduate/GED or
           "Some College/AA" = "Some college or AA degree",
           "College Grad" = "College graduate or above"),
         sroh = fct_recode(sroh,
               "Excellent" = "Excellent,",
               "Very Good" = "Very good,",
```

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 <fct> No, No, Yes, No, No, No, No, No, No, No, No,
$ mentalh  <fct> No, No, No, No, No, No, Yes, No, No, No, No,
$ 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 SBP in mm Hg (range: 76 to 209, mean 116)
hospital	Yes if hospitalized in last 12m, else No (8% Yes)
mentalh	Yes if saw a mental health professional in last 12m, else No (12% Yes)
sroh	Self-reported Overall Health (see next slide)
educ	Educational Attainment (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")
```

Paired SBP Differences

Normal Q–Q plot
SBP1 – SBP2 Differences
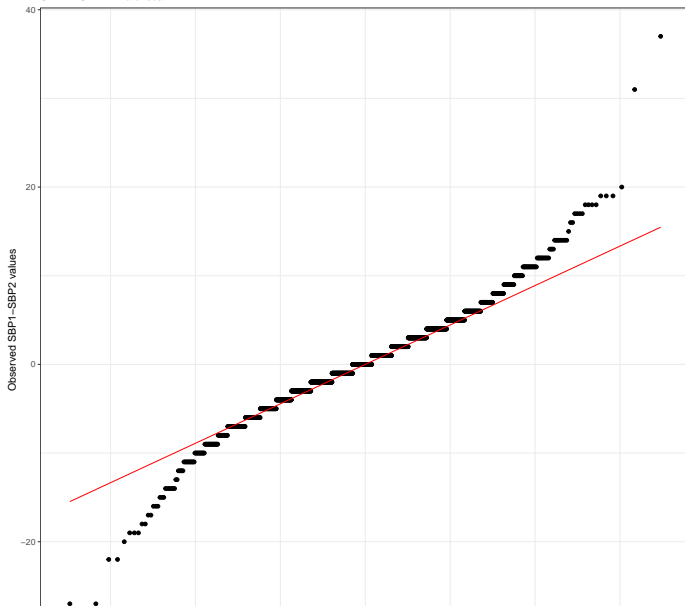
Comparing Paired Samples

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

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

```
set.seed(20240118)
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 = 0.04204
bootstrap mean of x (SE) = 0.2517891 (0.1185932)
90 percent bootstrap percentile confidence interval:
 0.05630676 0.45716448

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
```
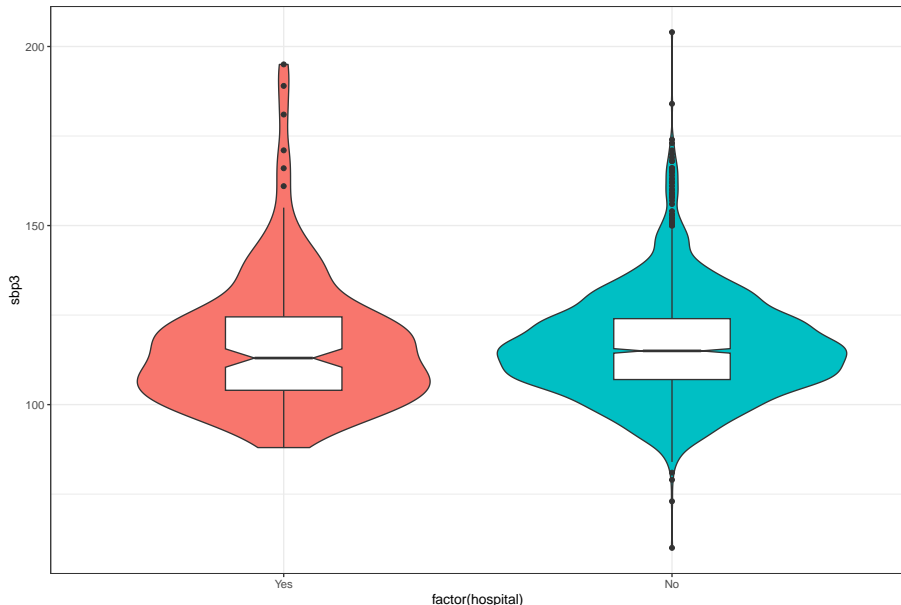
Comparing sbp3 by hospital: Independent Samples

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

```
  hospital min  Q1 median   Q3 max     mean       sd    n
1      Yes  88 104    113 124.5 195 116.7107 18.50260  159
2       No  60 107    115 124.0 204 116.1130 14.50532 1823
```

```
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

SBP (3rd reading) by Hospitalization

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 via linear model

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

tidy(lm2, conf.int = TRUE, conf.level = 0.90)

# A tibble: 2 x 7
  term         estimate std.error statistic p.value conf.low co
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>
1 (Intercept)   117.        1.18     99.0    0        115.
2 hospitalNo     -0.598     1.23     -0.486  0.627     -2.62

glance(lm2) |> select(r.squared, sigma)

# A tibble: 1 x 2
  r.squared sigma
      <dbl> <dbl>
1  0.000119  14.9
```

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    Yes   No Total
     Yes     37  210   247
      No    122 1613  1735
   Total    159 1823  1982
```

Standard Epidemiological Format

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

```
mosaic(~ educ + sroh, data = nh1982, highlighting = "sroh")
```
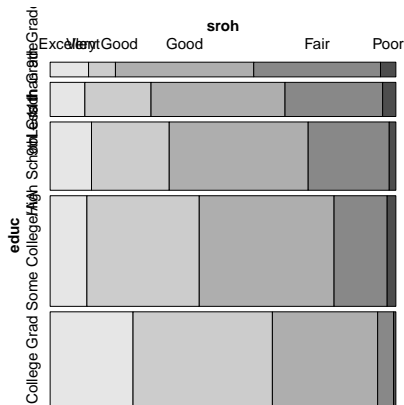
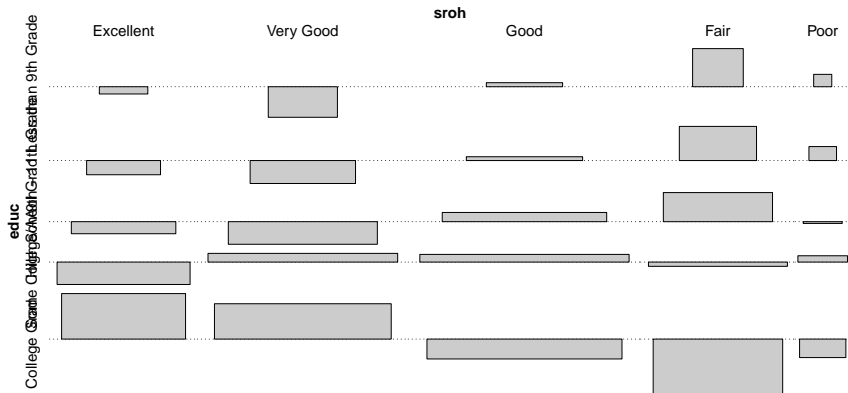# Pearson χ^2 test for our 5x5 Table

```
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

```
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.

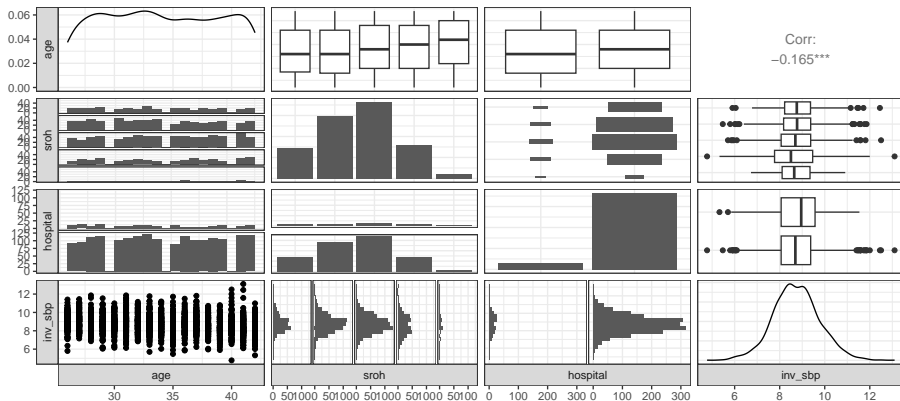# Box-Cox Plot to suggest potential outcome transformations

```
boxCox(temp_mod2)
```

Profile Log–likelihood
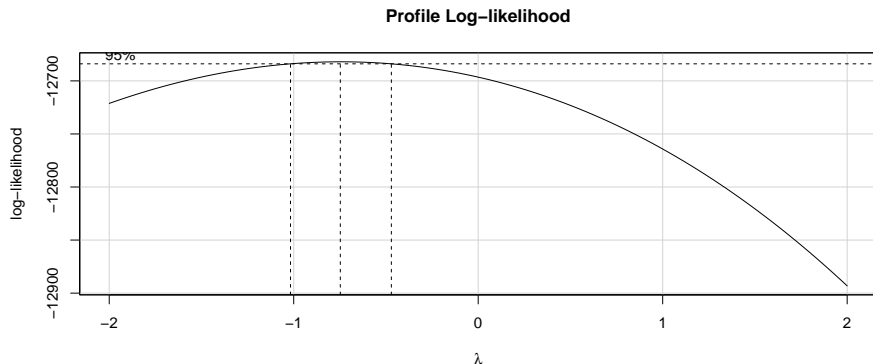

```
nh1982 <- nh1982 |> mutate(inv_sbp = 1000/mean_sbp)
```

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)

# A tibble: 2 x 7
  term         estimate std.error statistic  p.value conf.low c
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
1 (Intercept)    9.93      0.161      61.5  0           9.66
2 age           -0.0349    0.00469    -7.44 1.51e-13   -0.0426
```

Tidied Coefficients for Model m2

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

tidy(m2, conf.int = TRUE, conf.level = 0.9)
```

```
# A tibble: 7 x 7
  term          estimate std.error statistic  p.value conf.low
  <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
1 (Intercept)   10.0       0.185      54.3   0          9.74
2 age           -0.0338    0.00470    -7.19  9.27e-13  -0.0415
3 srohVery Good -0.0552    0.0727     -0.759 4.48e- 1  -0.175
4 srohGood      -0.110     0.0705     -1.56  1.20e- 1  -0.226
5 srohFair      -0.265     0.0825     -3.21  1.33e- 3  -0.401
6 srohPoor      -0.176     0.171      -1.03  3.02e- 1  -0.457
7 hospitalNo    -0.0464    0.0849     -0.546 5.85e- 1  -0.186
```

# Fit Summaries for Models `m1` and `m2`
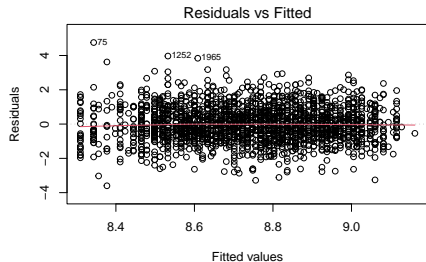
```
bind_rows(glance(m1), glance(m2)) |>
  mutate(model = c("m1", "m2")) |>
  select(model, r2 = r.squared, adjr2 = adj.r.squared,
         sigma, AIC, BIC, nobs, df, df.residual)
```

```
# A tibble: 2 x 9
  model    r2  adjr2 sigma   AIC   BIC  nobs    df df.residua
  <chr> <dbl>  <dbl> <dbl> <dbl> <dbl> <int> <dbl>        <int
1 m1   0.0272 0.0267  1.02 5714. 5731.  1982     1          198
2 m2   0.0334 0.0304  1.02 5711. 5756.  1982     6          197
```

Which model appears to fit the data better?

Residual Plots for Model `m2`
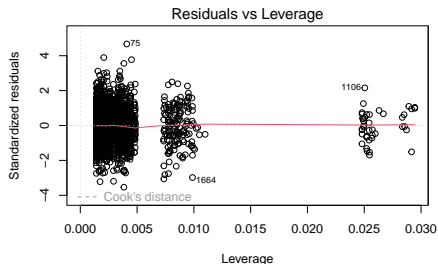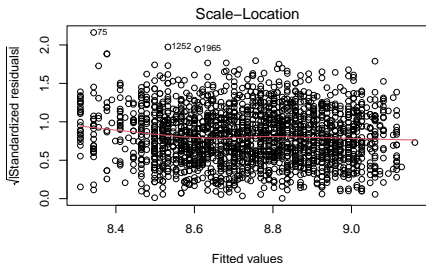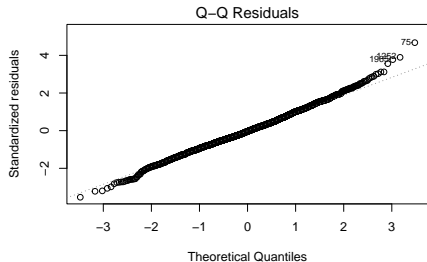
Making a Prediction in New Data

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.

Making a Prediction in New Data

```
new_person <- tibble(age = 29, sroh = "Good", hospital = "No")
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)
```

```
# A tibble: 2 x 6
  model fit_meansbp .fitted   age sroh  hospital
  <chr>       <dbl>   <dbl> <dbl> <chr> <chr>
1 m1           112.    8.92    29 Good  No
2 m2           112.    8.90    29 Good  No
```

Section 5

Setting Up Lab 1, due Tuesday 2024-01-23 at Noon

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 Tuesday Noon
  - Answer Sketch available Wednesday
- Developing Inferences Using Survey Weights
- Linear Regression and ANOVA/ANCOVA models