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

Loading required package: carData

library(Epi) # for twoby2  
library(GGally) # for ggpairs

Loading required package: ggplot2

Registered S3 method overwritten by 'GGally':  
 method from   
 +.gg ggplot2

library(MKinfer) # for boot.t.test  
library(mosaic) # for favstats

Registered S3 method overwritten by 'mosaic':  
 method from   
 fortify.SpatialPolygonsDataFrame ggplot2

The 'mosaic' package masks several functions from core packages in order to add   
additional features. The original behavior of these functions should not be affected by this.

Attaching package: 'mosaic'

The following objects are masked from 'package:dplyr':  
  
 count, do, tally

The following object is masked from 'package:Matrix':  
  
 mean

The following object is masked from 'package:ggplot2':  
  
 stat

The following object is masked from 'package:Epi':  
  
 factorize

The following objects are masked from 'package:car':  
  
 deltaMethod, logit

The following objects are masked from 'package:stats':  
  
 binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
 quantile, sd, t.test, var

The following objects are masked from 'package:base':  
  
 max, mean, min, prod, range, sample, sum

library(naniar) # deal with missing values  
library(nhanesA) # source of data  
library(vcd) # for mosaic (plot) and assoc (plot)

Loading required package: grid

Attaching package: 'vcd'

The following object is masked from 'package:mosaic':  
  
 mplot

library(janitor) # for tabyl and other things

Attaching package: 'janitor'

The following objects are masked from 'package:stats':  
  
 chisq.test, fisher.test

library(tidyverse) # for all kinds of things

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ lubridate 1.9.3 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.0  
✔ readr 2.1.5

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ mosaic::count() masks dplyr::count()  
✖ purrr::cross() masks mosaic::cross()  
✖ mosaic::do() masks dplyr::do()  
✖ tidyr::expand() masks Matrix::expand()  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
✖ tidyr::pack() masks Matrix::pack()  
✖ dplyr::recode() masks car::recode()  
✖ purrr::some() masks car::some()  
✖ mosaic::stat() masks ggplot2::stat()  
✖ mosaic::tally() masks dplyr::tally()  
✖ tidyr::unpack() masks Matrix::unpack()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

theme\_set(theme\_bw())

# 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)

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

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 grade with no diploma)",  
 "High School Grad" = "High school graduate/GED or equivalent",  
 "Some College/AA" = "Some college or AA degree",  
 "College Grad" = "College graduate or above"),  
 sroh = fct\_recode(sroh,  
 "Excellent" = "Excellent,",  
 "Very Good" = "Very good,",  
 "Good" = "Good,",  
 "Fair" = "Fair, or",  
 "Poor" = "Poor?")) |>  
 mutate(mean\_sbp = (sbp1 + sbp2 + sbp3)/3,  
 SEQN = as.character(SEQN)) |>  
 drop\_na() |>   
 droplevels()  
  
write\_rds(nh1982, "c02/data/nh1982.Rds")

## nh1982

glimpse(nh1982)

Rows: 1,982  
Columns: 10  
$ SEQN <chr> "109266", "109273", "109291", "109297", "109315", "109317", "…  
$ age <dbl> 29, 36, 42, 30, 30, 28, 33, 41, 35, 30, 41, 36, 29, 30, 32, 2…  
$ educ <fct> College Grad, Some College/AA, College Grad, Some College/AA,…  
$ sbp1 <dbl> 99, 116, 107, 105, 118, 110, 110, 106, 162, 111, 126, 131, 11…  
$ sbp2 <dbl> 99, 110, 111, 105, 123, 110, 105, 107, 148, 111, 133, 133, 11…  
$ sbp3 <dbl> 99, 115, 107, 102, 125, 110, 108, 113, 163, 113, 123, 133, 11…  
$ sroh <fct> Good, Good, Fair, Very Good, Good, Very Good, Excellent, Exce…  
$ hospital <fct> No, No, Yes, No, No, No, No, No, No, No, No, No, No, No, No, …  
$ mentalh <fct> No, No, No, No, No, No, Yes, No, No, No, No, No, Yes, Yes, No…  
$ mean\_sbp <dbl> 99.00000, 113.66667, 108.33333, 104.00000, 122.00000, 110.000…

## 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...

# 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?

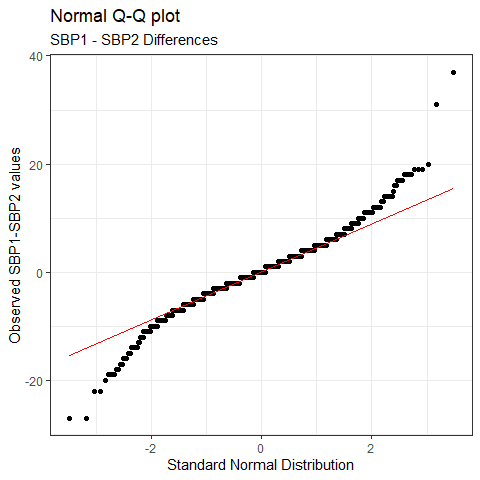
## Paired Samples Analysis

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



## 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

|  |
| --- |
|  |

## 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 × 7  
 term estimate std.error statistic p.value conf.low conf.high  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 117. 1.18 99.0 0 115. 119.   
2 hospitalNo -0.598 1.23 -0.486 0.627 -2.62 1.42

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

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

# 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  
  
 Exact P-value: 0.0001   
 Asymptotic P-value: 0.0000   
------------------------------------------------------

## 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

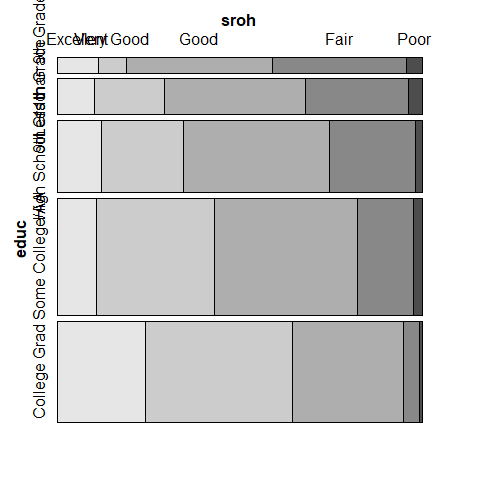
## Our 5x5 Table, showing SROH Proportions

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 test for our 5x5 Table

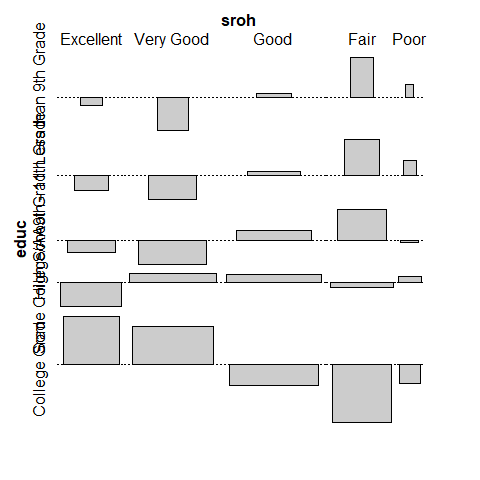
chisq.test(xtabs(~ educ + sroh, data = nh1982))

Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be  
incorrect

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)



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

## We’ll fit two models today

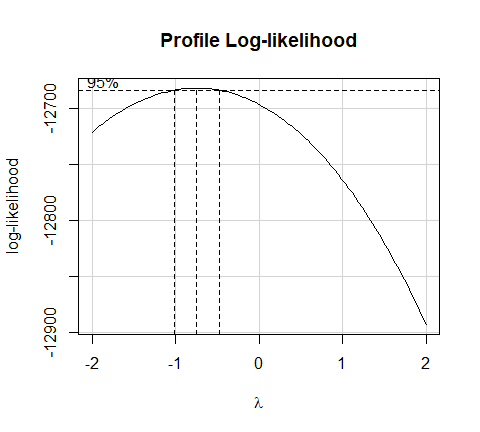
1. Predict mean SBP using Age alone.
2. Predict mean SBP (across three readings) using Age, Self-Reported Overall Health Status and Hospitalization Status.

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

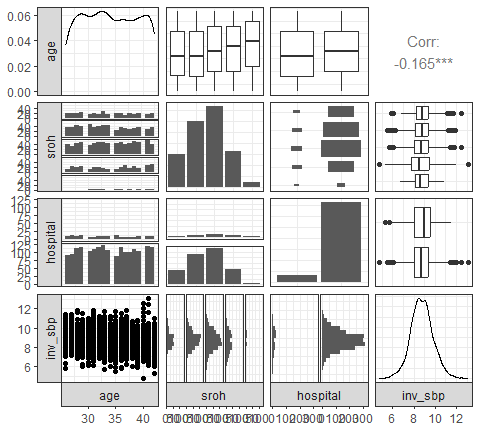
boxCox(temp\_mod2)



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)

# A tibble: 2 × 7  
 term estimate std.error statistic p.value conf.low conf.high  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 9.93 0.161 61.5 0 9.66 10.2   
2 age -0.0349 0.00469 -7.44 1.51e-13 -0.0426 -0.0272

## 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 × 7  
 term estimate std.error statistic p.value conf.low conf.high  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 10.0 0.185 54.3 0 9.74 10.3   
2 age -0.0338 0.00470 -7.19 9.27e-13 -0.0415 -0.0260   
3 srohVery Good -0.0552 0.0727 -0.759 4.48e- 1 -0.175 0.0644   
4 srohGood -0.110 0.0705 -1.56 1.20e- 1 -0.226 0.00627  
5 srohFair -0.265 0.0825 -3.21 1.33e- 3 -0.401 -0.129   
6 srohPoor -0.176 0.171 -1.03 3.02e- 1 -0.457 0.105   
7 hospitalNo -0.0464 0.0849 -0.546 5.85e- 1 -0.186 0.0933

## Fit Summaries for Models m1 and m2

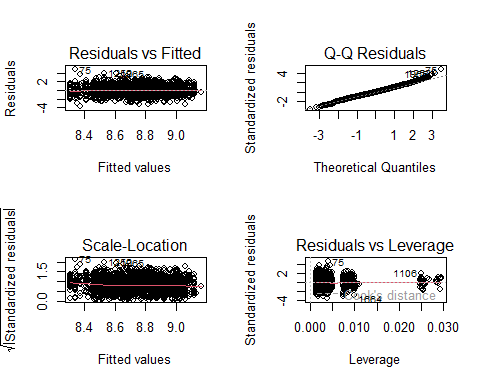
bind\_rows(glance(m1), glance(m2)) |>  
 mutate(model = c("m1", "m2")) |>   
 select(model, r2 = r.squared, adjr2 = adj.r.squared,   
 sigma, AIC, BIC, nobs, df, df.residual)

# A tibble: 2 × 9  
 model r2 adjr2 sigma AIC BIC nobs df df.residual  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <int>  
1 m1 0.0272 0.0267 1.02 5714. 5731. 1982 1 1980  
2 m2 0.0334 0.0304 1.02 5711. 5756. 1982 6 1975

Which model appears to fit the data better?

## Residual Plots for Model m2

par(mfrow = c(2,2)); plot(m2); par(mfrow = c(1,1))



## 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 × 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

# 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.)

1. Specify and fit the model, interpret food\_env coefficient.
2. Evaluate quality of model in terms of adherence to regression assumptions via four key residual plots.
3. 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