431 Class 12

thomaselove.github.io/431

2021-09-30

Today's Agenda

1. Using data from NHANES
2. A complex data management challenge
3. Using `dbp` to predict `sbp` again
4. Considering a transformation of our outcome

Today's Packages

```
library(NHANES) # for access to NHANES data
library(ggpubr) # to add equation to scatterplot easily
library(equatiomatic)
library(glue)
library(janitor)
library(knitr)
library(broom)
library(magrittr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

Ingesting and Managing Today's Data

Today's Data

The `NHANES` data file located in the `NHANES` package is our source.

> *NHANES stands for National Health and Nutrition Examination surveys. The NHANES target population is "the non-institutionalized civilian resident population of the United States". Since 1999, approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component, in a mobile examination centre.*
>
> *`NHANES` and `NHANESraw` each include 75 variables for the 2009-2010 and 2011-2012 sample years, with complex sampling weights included in `NHANESraw`. `NHANES` contains 10,000 rows of data resampled from `NHANESraw` to undo oversampling effects. `NHANES` can be treated, for educational purposes, as if it were a simple random sample from the American population.*

- For today, we'll do the one thing you should never do with NHANES data, which is to ignore the sampling weights.

Today's Data Ingest and Management

- We'll walk through these steps in the next few slides.

```r
set.seed(20210930)
nh12 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, BPSysAve, BPDiaAve, Age, Smoke100,
           Race1, HealthGen, SurveyYr) %>%
    rename(Subject = ID, SBP = BPSysAve, DBP = BPDiaAve,
           SROH = HealthGen) %>%
    clean_names() %>%
    mutate(across(where(is.character), as_factor)) %>%
    mutate(subject = as.character(subject)) %>%
    filter(age > 20 & age < 80) %>%
    filter(dbp > 39) %>%
    distinct() %>%
    slice_sample(n = 700) %>%
    droplevels()
```

Today's Data Management: Step 1

- Select the eight variables of interest from NHANES.

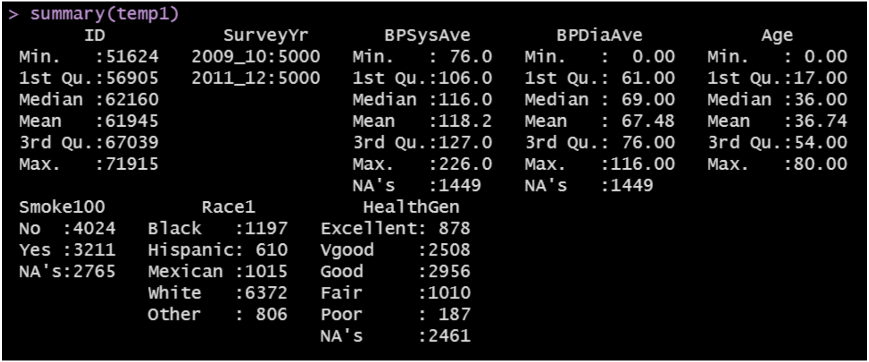
```
temp1 <- NHANES %>%
    select(ID, SurveyYr, BPSysAve, BPDiaAve, Age, Smoke100,
           Race1, HealthGen)
 #   filter(SurveyYr == "2011_12") %>%
 #   rename(Subject = ID, SBP = BPSysAve, DBP = BPDiaAve,
 #          SROH = HealthGen) %>%
 #    clean_names() %>%
 #    mutate(across(where(is.character), as_factor)) %>%
 #    mutate(subject = as.character(subject)) %>%
 #    filter(age > 20 & age < 80) %>%
 #    filter(dbp > 39) %>%
 #    distinct() %>%
 #    slice_sample(n = 700) %>%
 #    droplevels()
```

temp1 is a tibble.

```
temp1
```

```
# A tibble: 10,000 x 8
      ID SurveyYr BPSysAve BPDiaAve   Age Smoke100
   <int> <fct>       <int>    <int> <int> <fct>
 1 51624 2009_10       113       85    34 Yes
 2 51624 2009_10       113       85    34 Yes
 3 51624 2009_10       113       85    34 Yes
 4 51625 2009_10        NA       NA     4 <NA>
 5 51630 2009_10       112       75    49 Yes
 6 51638 2009_10        86       47     9 <NA>
 7 51646 2009_10       107       37     8 <NA>
 8 51647 2009_10       118       64    45 No
 9 51647 2009_10       118       64    45 No
10 51647 2009_10       118       64    45 No
# ... with 9,990 more rows, and 2 more variables:
#   Race1 <fct>, HealthGen <fct>
```

Summarizing `temp1`



```
> summary(temp1)
       ID              SurveyYr        BPSysAve         BPDiaAve           Age       
 Min.   :51624   2009_10:5000   Min.   : 76.0   Min.   :  0.00   Min.   : 0.00  
 1st Qu.:56905   2011_12:5000   1st Qu.:106.0   1st Qu.: 61.00   1st Qu.:17.00  
 Median :62160                  Median :116.0   Median : 69.00   Median :36.00  
 Mean   :61945                  Mean   :118.2   Mean   : 67.48   Mean   :36.74  
 3rd Qu.:67039                  3rd Qu.:127.0   3rd Qu.: 76.00   3rd Qu.:54.00  
 Max.   :71915                  Max.   :226.0   Max.   :116.00   Max.   :80.00  
                                NA's   :1449    NA's   :1449                    
 Smoke100          Race1           HealthGen   
 No  :4024   Black   :1197   Excellent: 878  
 Yes :3211   Hispanic: 610   Vgood    :2508  
 NA's:2765   Mexican :1015   Good     :2956  
             White   :6372   Fair     :1010  
             Other   : 806   Poor     : 187  
                             NA's     :2461  
```

Today's Data: Step 2

- Restrict to 2011-12 data, and rename some variables.

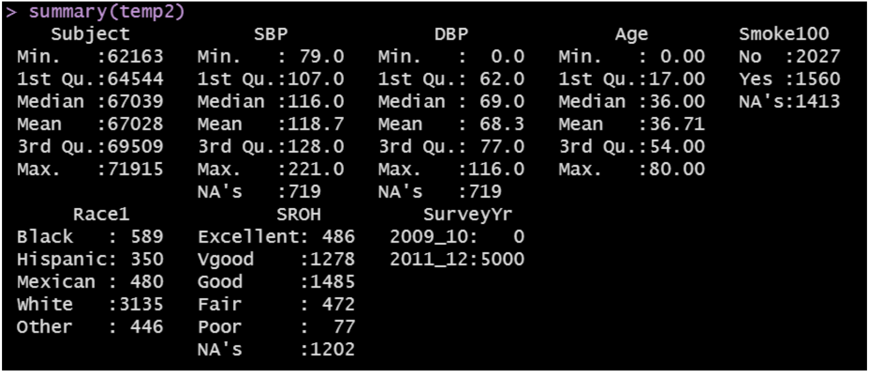
```
temp2 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, BPSysAve, BPDiaAve, Age,
           Smoke100, Race1, HealthGen, SurveyYr) %>%
    rename(Subject = ID, SBP = BPSysAve,
           DBP = BPDiaAve, SROH = HealthGen)
```

The temp2 tibble

```
temp2
```

```
# A tibble: 5,000 x 8
   Subject   SBP   DBP   Age Smoke100 Race1   SROH
     <int> <int> <int> <int> <fct>    <fct>   <fct>
 1   62163   107    37    14 <NA>     Other   Good
 2   62172   103    72    43 Yes      Black   Good
 3   62174    97    39    80 No       White   Fair
 4   62174    97    39    80 No       White   Fair
 5   62175    NA    NA     5 <NA>     White   <NA>
 6   62176   107    69    34 No       White   Vgood
 7   62178   121    72    80 No       White   Fair
 8   62180   107    66    35 No       White   Good
 9   62186   108    64    17 <NA>     Black   Vgood
10   62190   113    27    15 <NA>     Mexican Excellent
# ... with 4,990 more rows, and 1 more variable:
#   SurveyYr <fct>
```

Summary of temp2



```
> summary(temp2)
    Subject           SBP              DBP              Age          Smoke100  
 Min.   :62163   Min.   : 79.0   Min.   :  0.0   Min.   : 0.00   No  :2027  
 1st Qu.:64544   1st Qu.:107.0   1st Qu.: 62.0   1st Qu.:17.00   Yes :1560  
 Median :67039   Median :116.0   Median : 69.0   Median :36.00   NA's:1413  
 Mean   :67028   Mean   :118.7   Mean   : 68.3   Mean   :36.71              
 3rd Qu.:69509   3rd Qu.:128.0   3rd Qu.: 77.0   3rd Qu.:54.00              
 Max.   :71915   Max.   :221.0   Max.   :116.0   Max.   :80.00              
                 NA's   :719     NA's   :719                                
      Race1              SROH         SurveyYr   
 Black   : 589   Excellent: 486   2009_10:   0  
 Hispanic: 350   Vgood    :1278   2011_12:5000  
 Mexican : 480   Good     :1485                 
 White   :3135   Fair     : 472                 
 Other   : 446   Poor     :  77                 
                 NA's     :1202                 
```

Today's Data: Step 3

- Drop unused level (2009-10) from SurveyYr summary with `droplevels()`.
- Clean up the names to lower case with underscores using `clean_names()`.
- Use only distinct observations with `distinct()`.

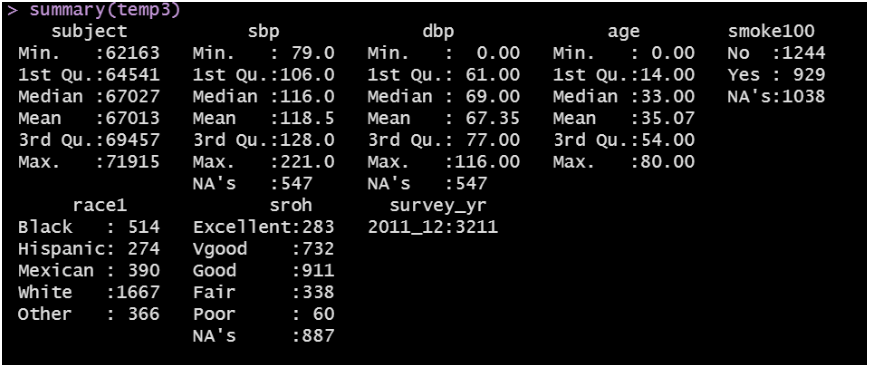
```
temp3 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, BPSysAve, BPDiaAve, Age,
           Smoke100, Race1, HealthGen, SurveyYr) %>%
    rename(Subject = ID, SBP = BPSysAve,
           DBP = BPDiaAve, SROH = HealthGen) %>%
    clean_names() %>%
    distinct() %>%
    droplevels()
```

The temp3 tibble

```
temp3
```

```
# A tibble: 3,211 x 8
   subject   sbp   dbp   age smoke100 race1    sroh
     <int> <int> <int> <int> <fct>    <fct>    <fct>
 1   62163   107    37    14 <NA>     Other    Good
 2   62172   103    72    43 Yes      Black    Good
 3   62174    97    39    80 No       White    Fair
 4   62175    NA    NA     5 <NA>     White    <NA>
 5   62176   107    69    34 No       White    Vgood
 6   62178   121    72    80 No       White    Fair
 7   62180   107    66    35 No       White    Good
 8   62186   108    64    17 <NA>     Black    Vgood
 9   62190   113    27    15 <NA>     Mexican Excellent
10   62199   110    65    57 Yes      White    Vgood
# ... with 3,201 more rows, and 1 more variable:
#   survey_yr <fct>
```

Summary of `temp3`



```
> summary(temp3)
    subject           sbp              dbp               age          smoke100
 Min.   :62163   Min.   : 79.0   Min.   :  0.00   Min.   : 0.00   No  :1244
 1st Qu.:64541   1st Qu.:106.0   1st Qu.: 61.00   1st Qu.:14.00   Yes : 929
 Median :67027   Median :116.0   Median : 69.00   Median :33.00   NA's:1038
 Mean   :67013   Mean   :118.5   Mean   : 67.35   Mean   :35.07
 3rd Qu.:69457   3rd Qu.:128.0   3rd Qu.: 77.00   3rd Qu.:54.00
 Max.   :71915   Max.   :221.0   Max.   :116.00   Max.   :80.00
                 NA's   :547     NA's   :547
      race1            sroh          survey_yr
 Black   : 514   Excellent:283   2011_12:3211
 Hispanic: 274   Vgood    :732
 Mexican : 390   Good     :911
 White   :1667   Fair     :338
 Other   : 366   Poor     : 60
                 NA's     :887
```

Today's Data: Step 4

- Make Race1 and HealthGen into factors, leave ID as character.
- Restrict Age to 21-79, and require DBP $>=$ 40 mm Hg.

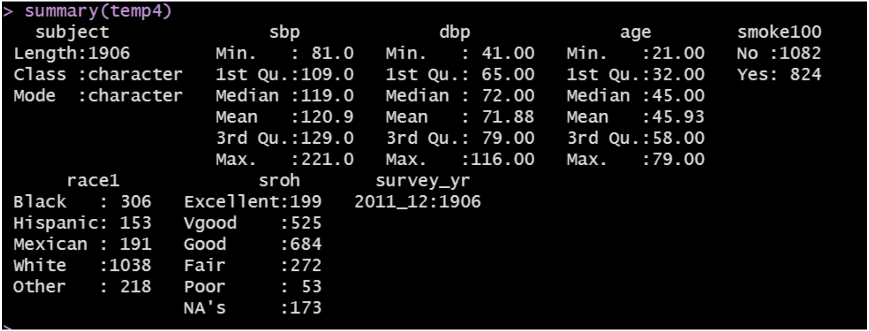
```
temp4 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, BPSysAve, BPDiaAve, Age, Smoke100,
           Race1, HealthGen, SurveyYr) %>%
    rename(Subject = ID, SBP = BPSysAve, DBP = BPDiaAve,
           SROH = HealthGen) %>%
    clean_names() %>%
    mutate(across(where(is.character), as_factor)) %>%
    mutate(subject = as.character(subject)) %>%
    filter(age > 20 & age < 80) %>%
    filter(dbp > 39) %>%
    distinct() %>%
    droplevels()
```

The temp4 tibble

```
temp4
```

```
# A tibble: 1,906 x 8
   subject   sbp   dbp   age smoke100 race1    sroh 
   <chr>   <int> <int> <int> <fct>    <fct>    <fct>
 1 62172     103    72    43 Yes      Black    Good 
 2 62176     107    69    34 No       White    Vgood
 3 62180     107    66    35 No       White    Good 
 4 62199     110    65    57 Yes      White    Vgood
 5 62205     122    87    28 No       White    Good 
 6 62206     106    50    35 No       White    <NA> 
 7 62208     105    59    38 No       Hispanic Good 
 8 62209     108    57    62 No       Mexican  Fair 
 9 62220     120    71    31 No       Black    Good 
10 62222     104    73    32 No       White    Good 
# ... with 1,896 more rows, and 1 more variable:
#   survey_yr <fct>
```

Summarizing the `temp4` tibble



```
> summary(temp4)
   subject               sbp             dbp              age          smoke100  
 Length:1906        Min.   : 81.0   Min.   : 41.00   Min.   :21.00   No :1082  
 Class :character   1st Qu.:109.0   1st Qu.: 65.00   1st Qu.:32.00   Yes: 824  
 Mode  :character   Median :119.0   Median : 72.00   Median :45.00             
                    Mean   :120.9   Mean   : 71.88   Mean   :45.93             
                    3rd Qu.:129.0   3rd Qu.: 79.00   3rd Qu.:58.00             
                    Max.   :221.0   Max.   :116.00   Max.   :79.00             
      race1             sroh        survey_yr   
 Black   : 306   Excellent:199   2011_12:1906  
 Hispanic: 153   Vgood    :525                 
 Mexican : 191   Good     :684                 
 White   :1038   Fair     :272                 
 Other   : 218   Poor     : 53                 
                 NA's     :173                 
```

Today's Data: Select random sample of 700

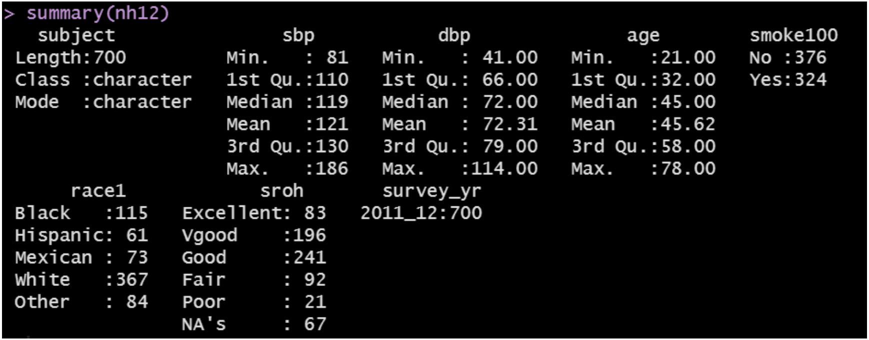
```
set.seed(20210930)
nh12 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, BPSysAve, BPDiaAve, Age, Smoke100,
           Race1, HealthGen, SurveyYr) %>%
    rename(Subject = ID, SBP = BPSysAve, DBP = BPDiaAve,
           SROH = HealthGen) %>%
    clean_names() %>%
    mutate(across(where(is.character), as_factor)) %>%
    mutate(subject = as.character(subject)) %>%
    filter(age > 20 & age < 80) %>%
    filter(dbp > 39) %>%
    distinct() %>%
    slice_sample(n = 700) %>%
    droplevels()
```

The nh12 tibble

```
nh12
```

```
# A tibble: 700 x 8
   subject   sbp   dbp   age smoke100 race1    sroh
   <chr>   <int> <int> <int> <fct>    <fct>    <fct>
 1 71420     126    69    54 No       Mexican  Good
 2 64368     136    74    70 Yes      Black    Vgood
 3 62546     150    84    64 Yes      Mexican  Good
 4 70531     110    73    49 No       Black    Excelle~
 5 62974      98    74    30 No       White    Good
 6 66294     143    77    76 Yes      White    Good
 7 68762     104    76    34 Yes      White    <NA>
 8 70758     125    64    21 Yes      Hispanic <NA>
 9 71315     132    84    44 Yes      White    Good
10 66600     137    72    64 No       Black    Vgood
# ... with 690 more rows, and 1 more variable:
#   survey_yr <fct>
```

Summary of the `nh12` tibble

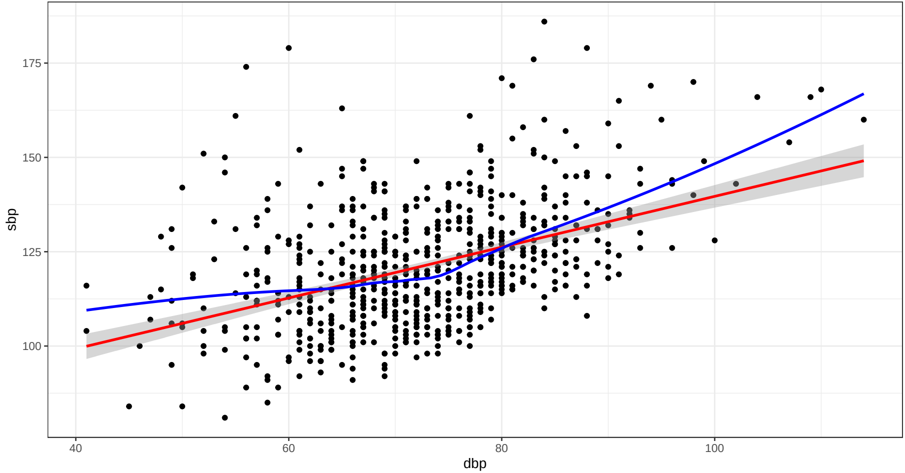


```
> summary(nh12)
   subject               sbp           dbp              age           smoke100
 Length:700         Min.   : 81   Min.   : 41.00   Min.   :21.00   No :376
 Class :character   1st Qu.:110   1st Qu.: 66.00   1st Qu.:32.00   Yes:324
 Mode  :character   Median :119   Median : 72.00   Median :45.00
                    Mean   :121   Mean   : 72.31   Mean   :45.62
                    3rd Qu.:130   3rd Qu.: 79.00   3rd Qu.:58.00
                    Max.   :186   Max.   :114.00   Max.   :78.00
      race1            sroh        survey_yr
 Black   :115   Excellent: 83   2011_12:700
 Hispanic: 61   Vgood    :196
 Mexican : 73   Good     :241
 White   :367   Fair     : 92
 Other   : 84   Poor     : 21
                NA's     : 67
```

- Outcome (quantitative): `sbp`
- Quantitative predictors: `dbp`, `age`
- Binary predictor: `smoke100` (Yes/No)
- 5-category predictor: `race1` (White, Black, Hispanic, Mexican, Other)
- 5-category predictor with missing data: `sroh` (E, VG, G, F,)
- Identification code: `subject`

Building Regression Model `m1` for `sbp`

Visualizing sbp against dbp



Pearson correlation r = 0.444.

Model m1

```
m1 <- lm(sbp ~ dbp, data = nh12)

tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	72.33	3.76	66.13	78.52
dbp	0.67	0.05	0.59	0.76

```
glance(m1) %>%
  select(r.squared, adj.r.squared, sigma, AIC, BIC, nobs) %>%
  kable(digits = c(3,3,1,1,1,0))
```

r.squared	adj.r.squared	sigma	AIC	BIC	nobs
0.197	0.196	14.3	5717.6	5731.2	700

Model `m1`

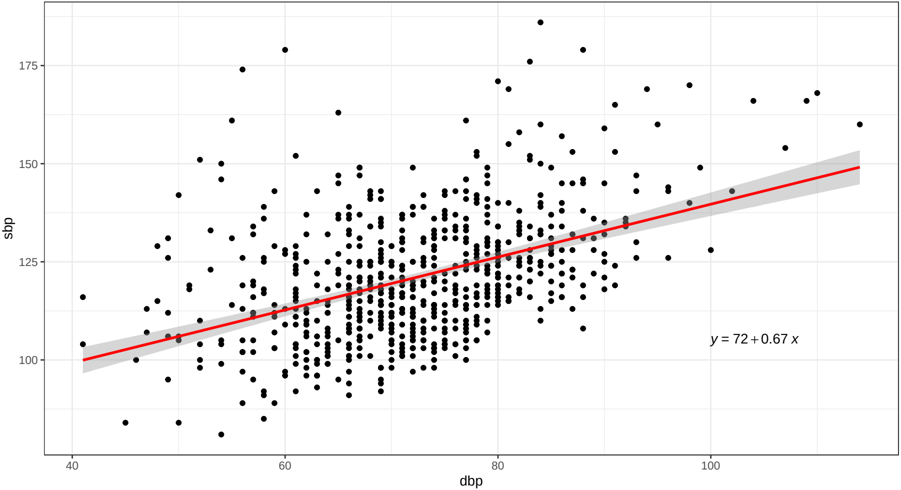
```r
extract_eq(m1, use_coefs = TRUE, coef_digits = 3)
```

$$\widehat{\text{sbp}} = 72.326 + 0.673(\text{dbp})$$

To include the equation in the scatterplot, I might use `stat_regline_equation()` from the ggpubr package.

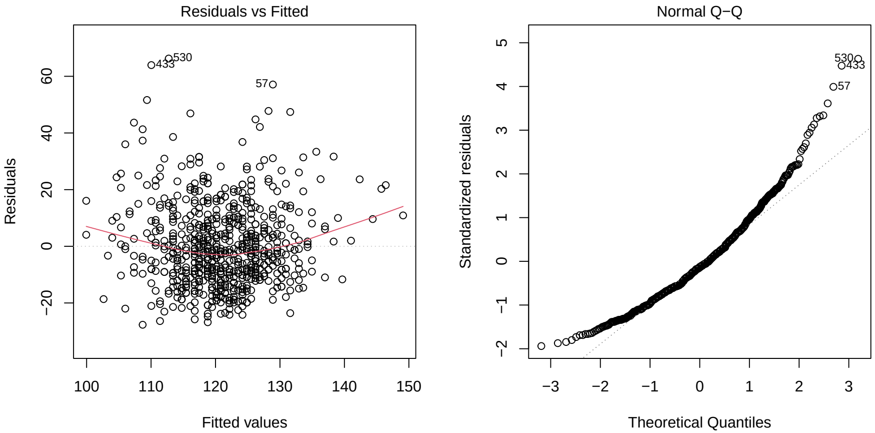
```r
ggplot(nh12, aes(x = dbp, y = sbp)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", formula = y ~ x) +
  stat_regline_equation(label.x = 100, label.y = 105)
```

Including the Equation in the Scatterplot

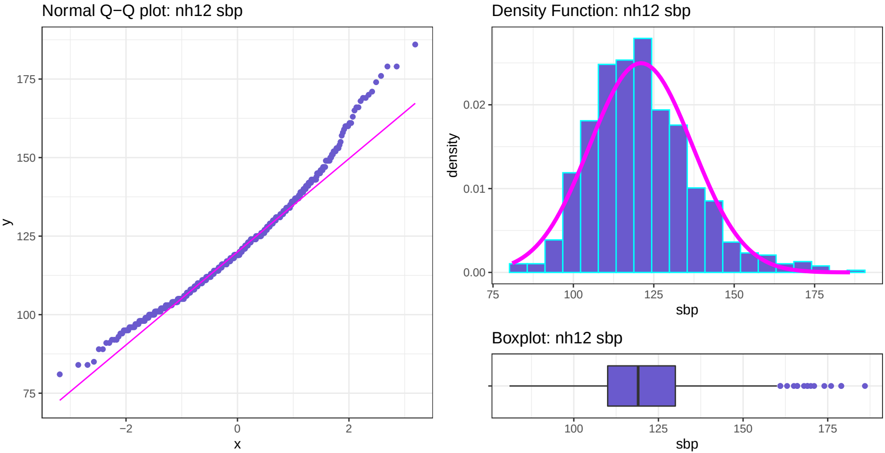


Quick Residual Plots for Model `m1`

```
par(mfrow = c(1,2))
plot(m1, which = c(1:2)); par(mfrow = c(1,1))
```

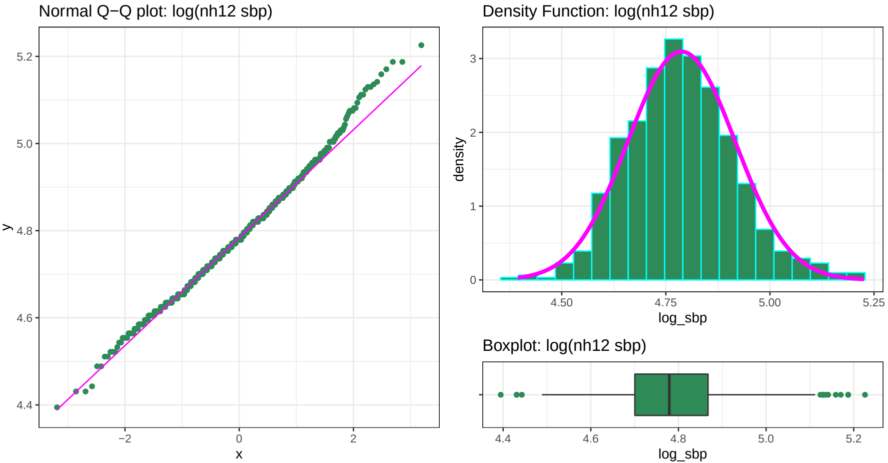


Should we think about transforming sbp here?



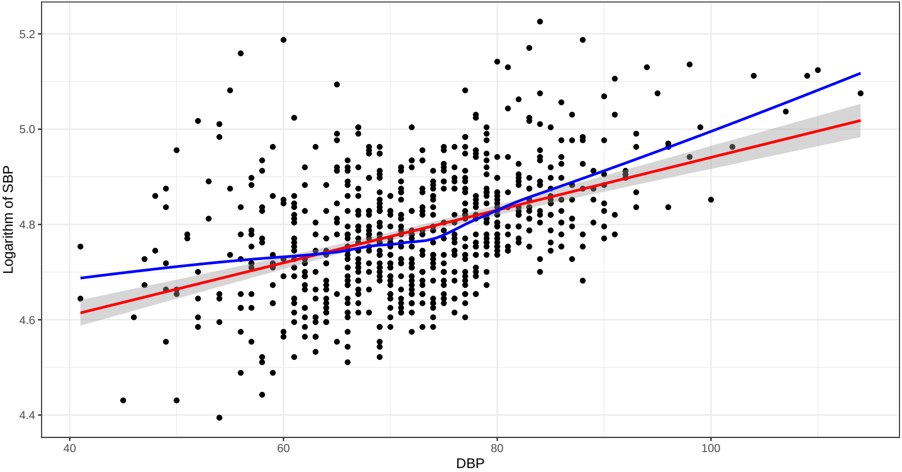
min	Q1	median	Q3	max	mean	sd	n	missing
81	110	119	130	186	121	16	700	0

Logarithm of sbp?



min	Q1	median	Q3	max	mean	sd	n	missing
4.4	4.7	4.8	4.9	5.2	4.8	0.1	700	0

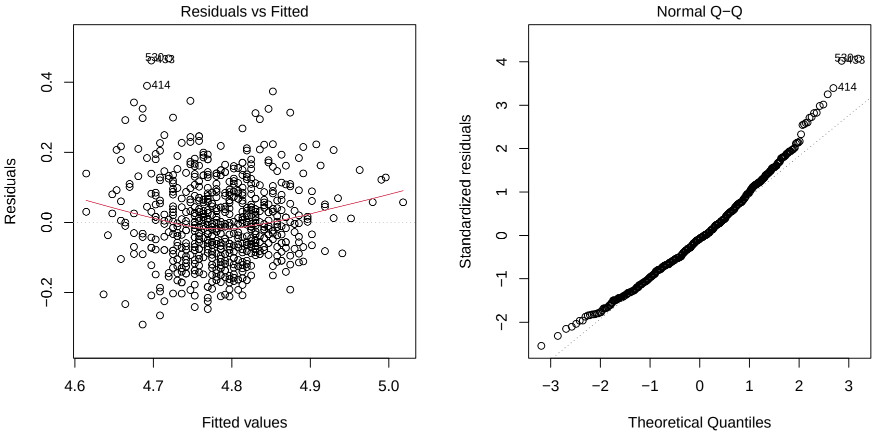
Scatterplot of log(sbp) vs. dbp



Pearson correlation r = 0.452.

Quick Residual Plots for Model m2

```
m2 <- lm(log(sbp) ~ dbp, data = nh12)
par(mfrow = c(1,2))
plot(m2, which = c(1:2)); par(mfrow = c(1,1))
```



Today's Agenda

1. Using data from NHANES
2. A complex data management challenge
3. Using `dbp` to predict `sbp` again
4. Considering a transformation of our outcome