Analysis of Hypertension using Vitamin C Level as a Key Predictor Along with Four Other Potential Risk Factors

Scarlett He and Zi Jie (Jay) Wei

2022-02-02

Setup and Data Ingest

Initial Setup and Package Loads

```
library(knitr); library(rmdformats)
library(janitor); library(magrittr); library(naniar)
library(broom); library(patchwork)
library(readxl)
library(Epi)
library(GGally)
library(MASS)
library(car)
library(equatiomatic)
library(mosaic)
library(Hmisc)
library(ggpubr)
library(nhanesA)
library(tidyverse)
## Load Love-boost
source("data/Love-boost.R")
## Global options
opts_chunk$set(comment=NA)
opts_knit$set(width=75)
theme_set(theme_bw())
```

Loading the Raw Data into R

We will be using the 2017-18 NHANES data for this project. We specifically selected the Demographic Variables and Sample Weights data file, Blood Pressure data file, and the Vitamin C data file. They are the following:

- `DEMO_J`
- `BPX_J`
- `WHQ_J`
- `VIC_J`

We called raw data from the Demographic Variables and Sample Weights data file demo_raw.

We called raw data from the Blood Pressure data file bp_raw.

We called raw data from the Weight History data file whq_raw.

We called raw data from the Vitamin C data file vitc_raw.

Codes

```
demo_raw <- nhanes('DEMO_J') %>% tibble()%>% clean_names()
```

```
Processing SAS dataset DEMO_J    ..
```

```
bp_raw <- nhanes('BPX_J') %>% tibble()%>% clean_names()
```

```
Processing SAS dataset BPX_J     ..
```

```
whq_raw <- nhanes('WHQ_J') %>% tibble()%>% clean_names()
```

```
Processing SAS dataset WHQ_J     ..
```

```
vitc_raw <- nhanes('VIC_J') %>% tibble() %>% clean_names()
```

```
Processing SAS dataset VIC_J     ..
```

Cleaning the Data

Contents of the Raw Tibbles

- `demo_raw` contains 46 variables and 9254 observations.

```
dim(demo_raw)
```

```
[1] 9254   46
```

- `bp_raw` contains 21 variables and 8704 observations.

```
dim(bp_raw)
```

```
[1] 8704   21
```

- `whq_raw` contains 37 variables and 6161 observations.

```
dim(whq_raw)
```

```
[1] 6161   37
```

Contents of the Raw Tibbles (Continued)

- `vitc_raw` contains 4 variables and 7435 observations.

```
dim(vitc_raw)
```

```
[1] 7435    4
```

Filtering out all variables that will be used

We will be using the following variables: respondent sequence number (SEQN), interview/examination status (RIDSTATR), age in years at screening (RIDAGEYR), country of birth (DMDBORN4), sex (RIAGENDR), first systolic blood pressure reading (BPXSY1), consideration on current weight (WHQ030), attempt to lose weight in the past year (WHQ070), race origin (RIDRETH3), and vitamin C level (LBXVIC).

We have limited our subjects to people who completed both interviews and examinations and who are considered as middle-aged adults (ages 36-55 years).

We also ensured that our final dataset had only complete data without any missing values.

We have called my filtered datasets: `demo1`, `bp1`, `whq1`, and `vitc1`.

```
demo1 <- demo_raw %>% select(seqn,ridstatr,ridageyr,riagendr,ridreth3, dmdborn4) %>%
  filter(complete.cases(.)) %>%
  filter(ridstatr==2) %>%
  filter(ridageyr %in% (36:55))

bp1 <- bp_raw %>% select(seqn, bpxsy1)%>% filter(complete.cases(.))

whq1 <-whq_raw %>% select(seqn, whq030,whq070)%>% filter(complete.cases(.))

vitc1 <- vitc_raw %>% select(seqn,lbxvic) %>% filter(complete.cases(.))
```

Three Merging Steps

1. We first merged the two datasets, `demo1` and `bp1`, by the respondent sequence number and obtained a new tibble, which we called `temp_1`.

```
temp_1 <- left_join(demo1, bp1, by="seqn")
```

1. We then merged another dataset, `whq1` with the new tibble `temp_1` by the respondent sequence number and obtained the final tibble, which we called `temp_2`.

```
temp_2 <- left_join(temp_1, whq1,by="seqn") %>% filter(complete.cases(.))
```

1. We then merged the last dataset, `vitc1` with the new tibble `temp_2` by the respondent sequence number and obtained the final tibble, which we called `merged_2`.

```
merged_2 <- left_join(temp_2, vitc1,by="seqn") %>% filter(complete.cases(.))
```

Final Merged Tibble

- `merged_2` contains 10 variables and 1143 observations. Nine variables are seqn, ridstatr, ridageyr, dmdborn4, riagendr, bpxsy1, ridreth3, whq030, whq070, and lbxvic.

```
dim(merged_2)
```

```
[1] 1143   10
```

Checking the Merge

The number of distinct respondent sequence number should match the number of rows. The output of the code below is true so the two values are identical.

```
identical(merged_2 %$% n_distinct(seqn),
          merged_2 %>% nrow())
```

```
[1] TRUE
```

Since we added new variable, the code here should output false, and it does output false.

```
identical(names(merged_2),
          names(demo1))
```

```
[1] FALSE
```

Since both checks are successful, we think our merge was correct.

Checking our Outcome and Key Predictor

Our outcome variable is first systolic blood pressure readings (bpxsy1) and our key predictor variable is vitamin C level in mg/dL (lbxvic). We have no missing values because we only selected complete cases. We have 1142 subjects for analysis. The ranges of our outcome and predictor variables look plausible, so we think it is safe to continue. The only additional thing we have done here is to remove the labeller for each variable by mutating all our variables to numeric variables. We will change some of them back into factors in the later steps.

```
df_stats(~bpxsy1+lbxvic,data=merged_2)
```

```
  response    min       Q1  median     Q3    max        mean         sd    n
1   bpxsy1 88.000 112.0000 122.000 130.00 216.00 122.5774278 16.2453522 1143
2   lbxvic  0.033   0.4925   0.835   1.12   2.76   0.8313141  0.4445638 1143
  missing
1       0
2       0
```

```
merged_2 <- merged_2 %>% mutate_at(vars(riagendr:lbxvic), as.numeric)
```

Checking the Categorical Variables

`riagendr`: Sex (Male/Female)

```
merged_2 <- merged_2 %>%
    mutate(riagendr = fct_recode(factor(riagendr),"Male" = "1", "Female" ="2"))
merged_2 %>% tabyl(riagendr)
```

```
 riagendr   n   percent
     Male 546 0.4776903
   Female 597 0.5223097
```

```
merged_2 <- merged_2 %>% mutate(riagendr =fct_relevel(factor(riagendr),"Female","Male"))
```

- We will first convert `riagendr` into a factor.
- The order of `riagendr` goes from "Male" to "Female". We want female to be the reference group because male tends to have higher systolic blood pressure based on research. Thus, we will use the `fct_relevel` function to fix this issue.
- We have 546 male subjects and 597 female subjects

Checking the Categorical Variables (Continued)

`whq070`: Tried to lose weight in past year? (Attempt to lose weight in the past year? yes/no)

```
merged_2 <- merged_2 %>%
    mutate(whq070 = fct_recode(factor(whq070),
                               "Attempted" = "1", "Not Attempted" = "2"))
merged_2 %>% tabyl(whq070)
```

```
        whq070   n  percent
     Attempted 521 0.455818
 Not Attempted 622 0.544182
```

```
merged_2 <- merged_2 %>% mutate(whq070 =fct_relevel(factor(whq070),"Not Attempted","Attempted"))
```

- We will first convert `whq070` into a factor. We then changed “Yes” to “Attempted” and “No” to “Not Attempted” for a better understanding.
- The order of `riagendr` goes from “Attempted” to “Not Attempted.” We want to reorder so that “Not Attempted” is the reference group, and the new order should be: (“Not Attempted”, “Attempted”).
- The number of subjects for each subgroup is greater than 30.

Checking the Categorical Variables (Continued)

`whq030`: How do you consider your weight? (Consideration on current weight. Overweight/underweight/about the right weight)

```
merged_2 <- merged_2 %>%
  mutate(whq030 = fct_recode(factor(whq030),
"Overweight" = "1", "Underweight" = "2", "About the right weight" = "3")) %>% filter(whq030 != "9") %>%
  droplevels()

merged_2 %>% tabyl(whq030)
```

```
                whq030   n    percent
            Overweight 654 0.57267951
           Underweight  46 0.04028021
 About the right weight 442 0.38704028
```

```
merged_2 <- merged_2 %>%
mutate(whq030 =fct_relevel(factor(whq030),"About the right weight", "Underweight", "Overweight"))
```

Checking the Categorical Variables (Continued)

`whq030`: How do you consider your weight? (Consideration on current weight. Overweight/underweight/about the right weight)

- We will first convert `whq070` into a factor. We then changed “1” to “Overweight”, “2” to “Underweight, and”3" to “About the right weight” for a better understanding/clarification.
- We dropped the level where whq030 equals “9” for subjects who did not know about their current weight status.
- Instead of (“Overweight”, “Underweight”, “About the right weight”), we want the new order to be: (“About the right weight”, “Underweight”, “Overweight”). Therefore, we used `fct_revel` to fix this.
- The number of subjects for each subgroup is greater than 30.

Checking the Categorical Variables (Continued)

`ridreth3`: Race/Hispanic origin w/ NH Asian (Mexican American/Other Hispanic/Non-Hispanic White/Non-Hispanic Black/Non-Hispanic Asian/Other Race - Including Multi-Racial)

```
merged_2 <- merged_2 %>% mutate(ridreth3 = fct_recode(factor(ridreth3),
"Mexican American" = "1", "Other Hispanic" = "2", "Non-Hispanic White" = "3",
"Non-Hispanic Black" = "4", "Non-Hispanic Asian" = "6", "Other Race" = "7"))

merged_2 %>% tabyl(ridreth3)
```

```
           ridreth3   n    percent
   Mexican American 184 0.16112084
     Other Hispanic 108 0.09457093
 Non-Hispanic White 341 0.29859895
 Non-Hispanic Black 228 0.19964974
 Non-Hispanic Asian 221 0.19352014
         Other Race  60 0.05253940
```

```
merged_2 <- merged_2 %>% mutate(ridreth3 =fct_relevel(factor(ridreth3),"Non-Hispanic White"))
```

Checking the Categorical Variables (Continued)

`ridreth3`: Race/Hispanic origin w/ NH Asian (Mexican American/Other Hispanic/Non-Hispanic White/Non-Hispanic Black/Non-Hispanic Asian/Other Race - Including Multi-Racial)

- We will first convert `ridreth3` into a factor. We then changed “1” to “Mexican American”, “2” to “Other Hispanic”, “3” to “Non-Hispanic White”, “4” to “Non-Hispanic Black”, “6” to “Non-Hispanic Asian”, and “7” to “Other Race” for a better understanding/clarification. -Instead of this current order, we want “Non-Hispanic White” to be the reference group (or on the top) because in most research papers, non-Hispanic white is the reference group.
- The number of subjects for each subgroup is greater than 30.

Checking the Categorical Variables (Continued)

What about the subjects?

We checked here to make sure that the numbers of our unique code `seqn` matches the number of rows in the dataset. We have a total of 1142 subjects.

```
nrow(merged_2)
```

```
[1] 1142
```

```
n_distinct(merged_2 %>% select(seqn))
```

```
[1] 1142
```

We will select only the variables that we will be using for the analysis and call the new tibble `final_2`. We think we can proceed to the next step from here.

```
final_2 <- merged_2 %>%
  select(seqn,riagendr,ridreth3,bpxsy1,whq030,whq070,lbxvic)
```

Codebook and Data Description

The Codebook

The 7 variables in the `final_2` tibble as the following:

Variable	Type	Description / Levels
seqn	ID	The respondent sequence number (Between 93718-102956)
bpxsy1	Quant	**Outcome variable**; First systolic blood pressure (SBP) reading in mmHg with minimum of 72 and maximum of 216.
lbxvic	Quant	**Key predictor**; Vitamin C level in mg/dL with minimum of 0.033 and maximum of 2.760.
riagendr	Cat-2	yes, no: Male or female?
whq030	Cat-3	Overweight, Underweight, About the right weight: How do you consider your weight?
whq070	Cat-2	yes, no: Tried to lose weight in past year?
ridreth3	Cat-6	Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other Race - Including Multi-Racial: Race/Hispanic origin w/ NH Asian

Analytic Tibble

We have proven here that we have a tibble.

```
final_2
```

```
# A tibble: 1,142 x 7
   seqn       riagendr ridreth3           bpxsy1 whq030         whq070    lbxvic
   <labelled> <fct>    <fct>               <dbl> <fct>          <fct>      <dbl>
 1 93718      Male     Non-Hispanic Black    128 About the rig~ Not Atte~  2.05
 2 93728      Male     Non-Hispanic Black    122 Overweight     Not Atte~  0.204
 3 93729      Male     Non-Hispanic Black    130 About the rig~ Not Atte~  0.396
 4 93735      Male     Other Hispanic        154 Overweight     Not Atte~  0.071
 5 93758      Female   Non-Hispanic White    154 Overweight     Not Atte~  0.104
 6 93760      Female   Mexican American      132 About the rig~ Not Atte~  0.615
 7 93761      Male     Non-Hispanic Asian    132 About the rig~ Attempted  0.13
 8 93763      Male     Non-Hispanic White    120 Overweight     Not Atte~  0.209
 9 93766      Female   Other Hispanic        148 Overweight     Not Atte~  0.67
10 93770      Male     Other Hispanic        130 Overweight     Attempted  0.078
# ... with 1,132 more rows
```

```
is_tibble(final_2)
```

```
[1] TRUE
```

Numerical Data Description

We have no missing values. The orders of our categorical variables are correct. We should be fine to proceed to our research question.

```
final_2 %>%
  select(-seqn) %$%
  Hmisc::describe(.)
```

```
.

 6  Variables      1142  Observations
---------------------------------------------------------------------------
riagendr
       n  missing distinct
    1142        0        2

Value      Female   Male
Frequency     597    545
Proportion  0.523  0.477
---------------------------------------------------------------------------
ridreth3
       n  missing distinct
    1142        0        6

lowest : Non-Hispanic White Mexican American   Other Hispanic     Non-Hispanic Black Non-Hispanic Asian
highest: Mexican American   Other Hispanic     Non-Hispanic Black Non-Hispanic Asian Other Race

Value      Non-Hispanic White   Mexican American     Other Hispanic
Frequency                 341                184                108
Proportion              0.299              0.161              0.095

Value      Non-Hispanic Black Non-Hispanic Asian         Other Race
Frequency                 228                221                 60
Proportion              0.200              0.194              0.053
```

Our Research Question

Background

For this research, we will be using the National Health and Nutrition Examination Surveys (NHANES) data from 2017 to 2018. According to the CDC website, every year, there are approximately 5,000 individuals of all ages interviewed in their homes and complete the health examination component of the survey. For this analysis, we are specifically interested in studying the effects of hypertension in middle-aged adults (ages 36-55 years; 1142 subjects). We want to explore the risk factors of hypertension to promote better health care guidelines and prevention measures prior to these middle-aged adults turning into older adults (60+). Hypertension (or high blood pressure) is a common disease in older adults. Untreated hypertension can lead to more serious cardiovascular disease and other health conditions. Additionally, we are interested in the effect of using vitamin C levels to predict blood pressure levels. In some past research articles, higher vitamin C level is associated with lower blood pressure. Thus, our research question is as the following:

Question

How effectively can we predict the subjects' first systolic BP readings using their vitamin C levels, and is the quality of prediction meaningfully improved when I adjust for four other predictors (gender, subjects' consideration of their current weight, attempt to lose weight in the past year, and race origin) in the `final_2` data?

Partitioning the Data

Partitioning

We have created the training sample `training_2` with a randomly selected 70% of the data from `final2` and the test sample `test_2` with the remaining 30% of the data from `final2`. The set.seed we used here is 4312021.

```
#We are using set.seed here to make sure we can get the same result later.
set.seed(4312021)

training_2 <- final_2 %>%
  slice_sample(., prop = .70)
test_2 <- anti_join(final_2, training_2, by = "seqn")
```

Partitioning (Continued)

```
#We are checking the number of rows and columns in final_2
dim(final_2)
```

```
[1] 1142    7
```

```
#We are checking the number of rows and columns in training data to make sure we have 70% of final_2 here
dim(training_2)
```

```
[1] 799   7
```

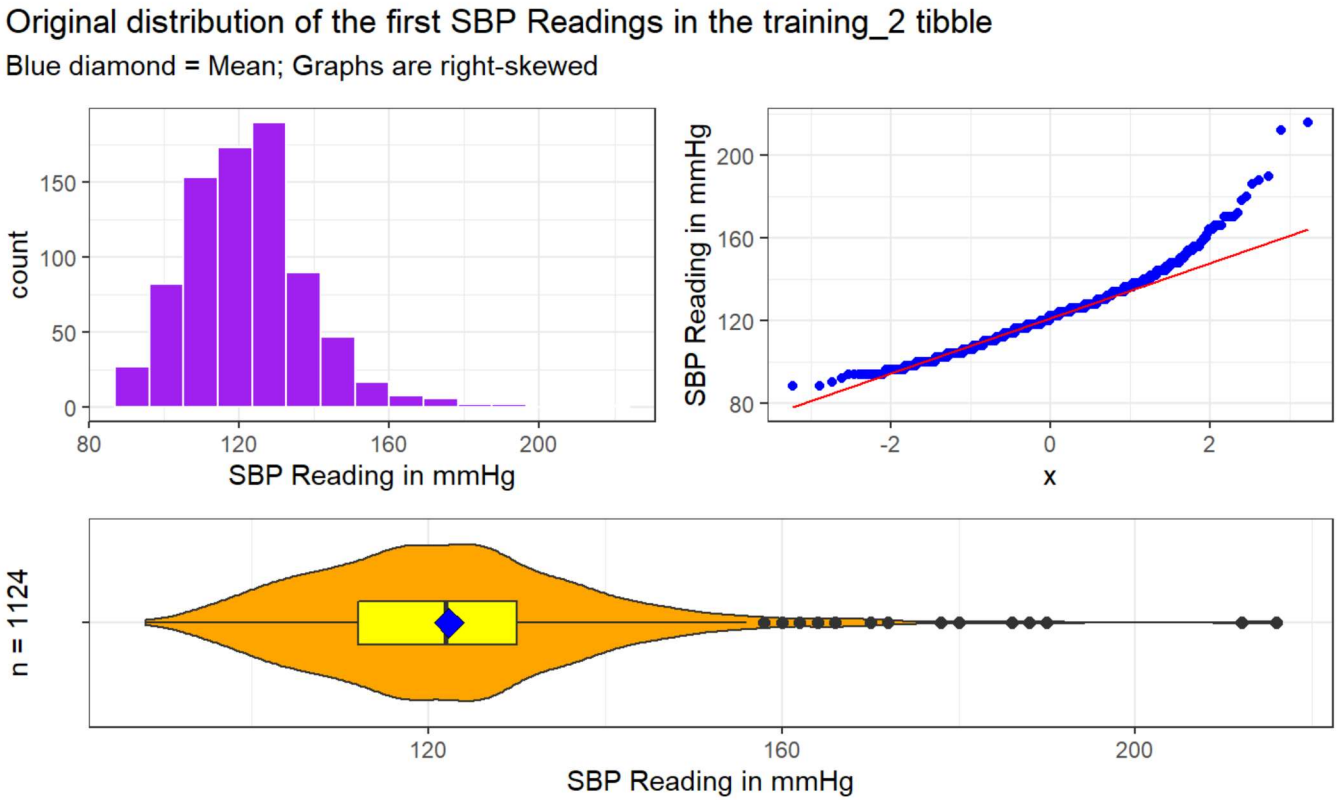
```
#We are checking the number of rows and columns in test data to make sure we have the rest(30%) of final_2 here
dim(test_2)
```

```
[1] 343   7
```

The total number of rows based on results above: 799+343=1142, which is same as the total number of rows in final_2.

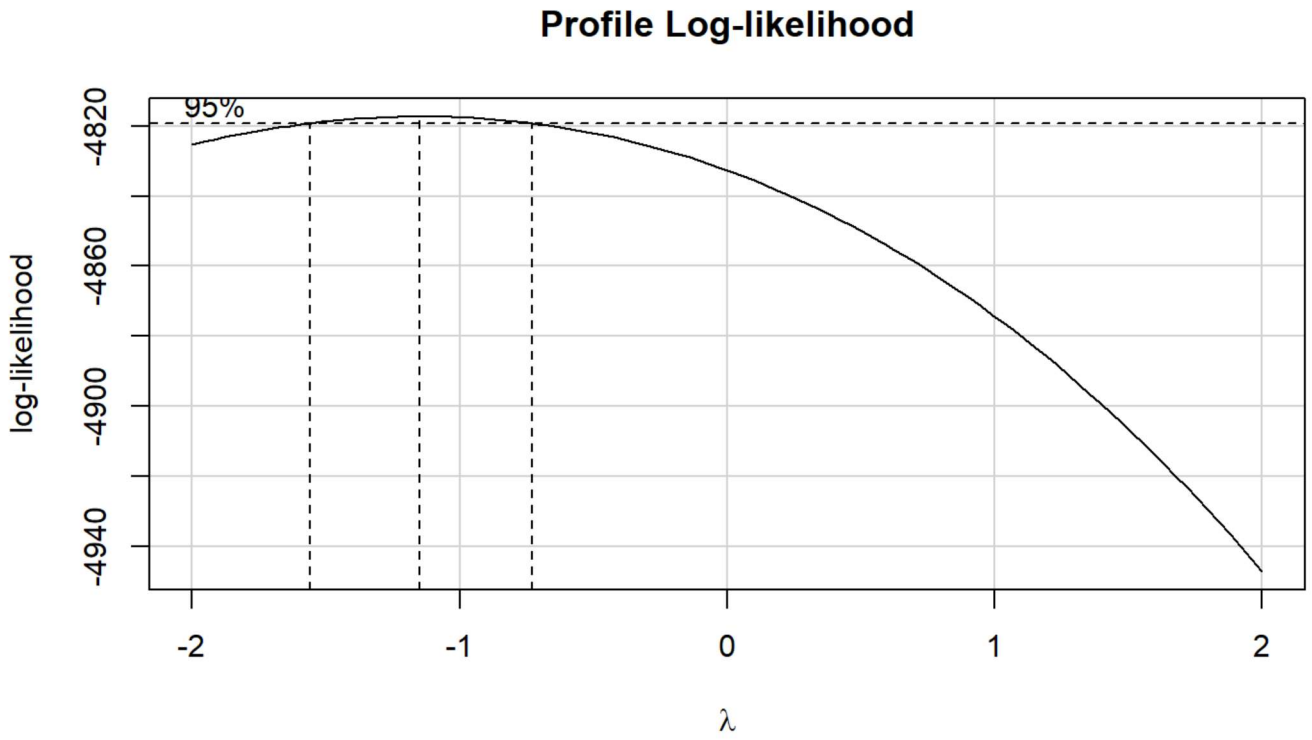
Transforming the Outcome

Visualizing the Outcome Distribution



Based on this initial visualization of the data outcome, we can see that the data is very right-skewed. We need to do transformation.

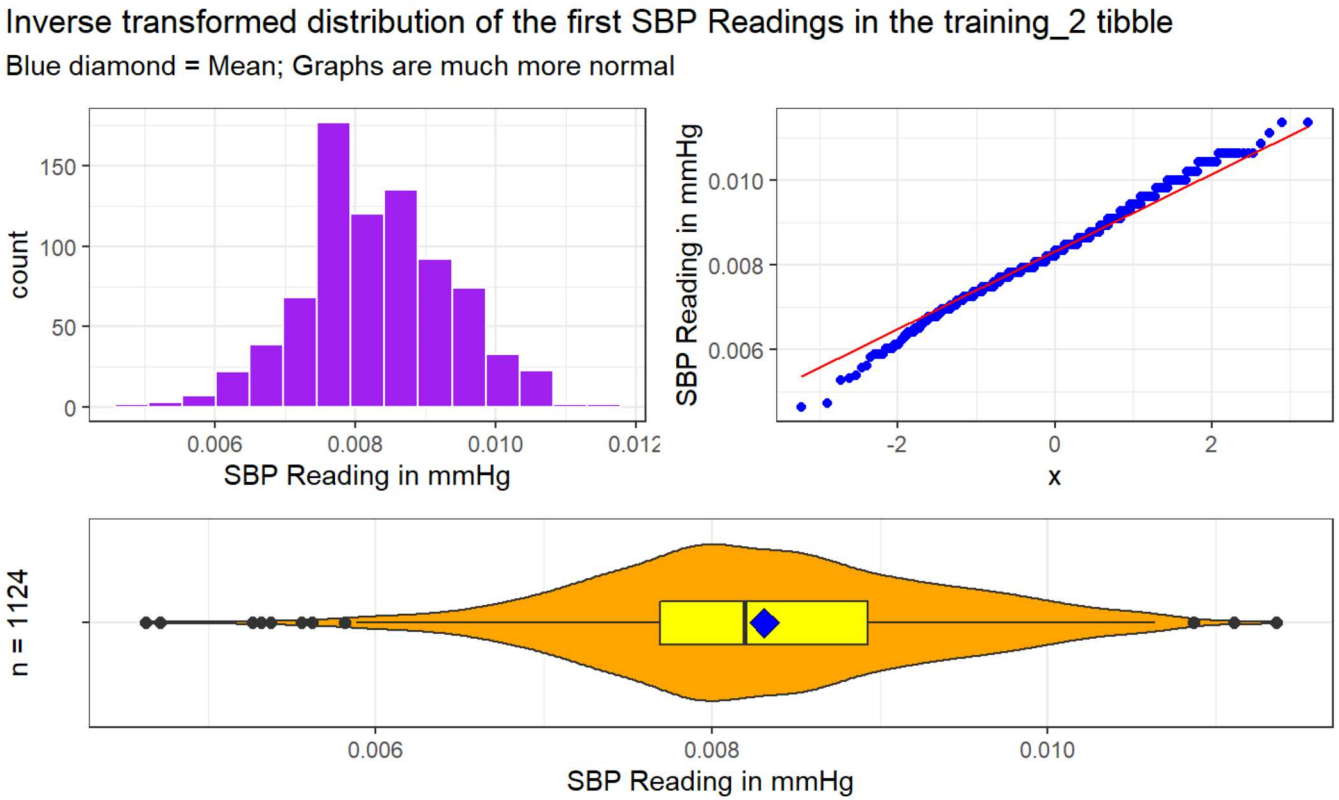
Assessing Transformation using **boxCox** function




```
Estimated transformation parameter
       Y1
-1.140491
```

The estimated transformation parameter is -1.14, which is close to -1. Thus, we will use inverse transformation for our data outcome.

Visualizing the Transformed Outcome



Based on the plot results above, we can see the transformed outcome is much more normally-distributed. Thus, we will use inverse transformation for our analysis.

- We created a new variable called `inv_bpxsy1` (inverse transformed bpxsy1).

Numerical Summary of the Outcome

We have provided the summary of our outcome variable `bpxsy1` (first systolic blood pressure) in both the original form and the inverse transformed form. For the original form, the minimum is 88, the median is 122, the mean is 122.3, and the maximum is 216. For the transformed form, the minimum is 0.0046, the median is 0.0082, the mean is 0.0083, and the maximum is 0.011. We have 799 subjects.

```
favstats(~ bpxsy1, data = training_2)
```

```
 min  Q1 median  Q3 max     mean       sd   n missing
  88 112    122 130 216 122.3204 16.33998 799       0
```

```
favstats(~ 1/bpxsy1, data = training_2)
```

```
        min          Q1      median          Q3         max        mean
 0.00462963 0.007692308 0.008196721 0.008928571 0.01136364 0.008311222
          sd   n missing
 0.001042635 799       0
```

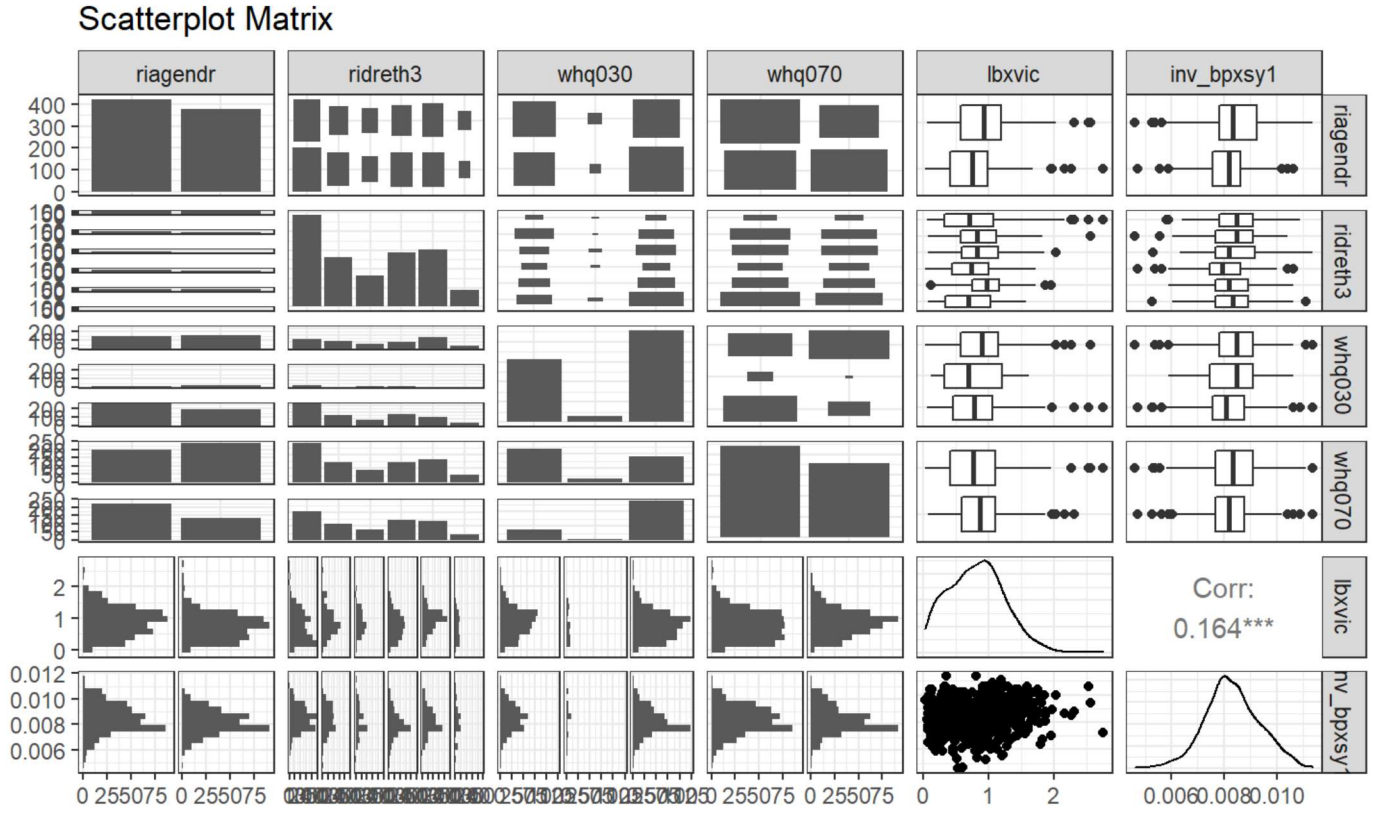
Numerical Summaries of the Predictors

We have provided the summary of our predictor variables: `lbxvic`, `riagendr`, `whq030`, `whq070`, and `ridreth3`. The mean and median for vitamin C level are 0.83 and 0.82, respectively. We have a total of 799 subjects.

```
categorical variables:
      name  class levels   n missing
1 riagendr factor      2 799       0
2 ridreth3 factor      6 799       0
3   whq030 factor      3 799       0
4   whq070 factor      2 799       0
                                distribution
1 Female (52.8%), Male (47.2%)
2 Non-Hispanic White (30.5%) ...
3 Overweight (57.2%) ...
4 Not Attempted (55.3%) ...

quantitative variables:
           name   class        min          Q1      median          Q3
...1     lbxvic numeric 0.03800000 0.500000000 0.822000000 1.110000000
...2 inv_bpxsy1 numeric 0.00462963 0.007692308 0.008196721 0.008928571
            max        mean          sd   n missing
...1 2.76000000 0.826501877 0.439313425 799       0
...2 0.01136364 0.008311222 0.001042635 799       0
```

Scatterplot Matrix

```
training_2 %>%
  select(riagendr,ridreth3,whq030,whq070,lbxvic,inv_bpxsy1) %>%
  ggpairs(., title = "Scatterplot Matrix",
          lower = list(combo = wrap("facethist", bins = 15)))
```



Scatterplot Matrix (Continued)

- We have provided a scatterplot matrix here for some general analyses.
- We should specifically look at the rightest column in the graph.
- The correlation between the inverse of our outcome variable and the key predictor is positive. This is surprising because as vitamin C level increases, systolic blood pressure level also increases.
- For the rest of the categorical predictors, it is hard for us to determine if there are any differences between each subgroup of the categorical predictors. However, from the graph, each subgroup appears to be similar to each other.

We think we had done enough for the visualization of the data. We will proceed to the next step.

Collinearity Checking

We have no other numeric candidate predictor other than our key predictor. Therefore, we do not think there are any problems with collinearity now. We will run a generalized VIF calculation in the later steps.

The Big Model

How We Build Our Big Model

- We will build the big model using all our predictor variables.
- We will use the inverse transformed outcome.
- We will use a 90% confidence level for the analysis.

Fitting/Summarizing the Kitchen Sink model

```
model_big <- lm(inv_bpxsy1 ~ lbxvic+riagendr+ridreth3+whq030+whq070,
                data = training_2)
summary(model_big)
```

```

Call:
lm(formula = inv_bpxsy1 ~ lbxvic + riagendr + ridreth3 + whq030 +
    whq070, data = training_2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0039150 -0.0005971 -0.0000152  0.0006210  0.0031956

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.537e-03  1.199e-04  71.182  < 2e-16 ***
lbxvic                      3.254e-04  8.462e-05   3.845 0.000130 ***
riagendrMale               -2.735e-04  7.368e-05  -3.712 0.000220 ***
ridreth3Mexican American   -1.653e-04  1.100e-04  -1.502 0.133471
ridreth3Other Hispanic     -2.098e-04  1.288e-04  -1.629 0.103675
ridreth3Non-Hispanic Black -4.673e-04  1.061e-04  -4.403 1.22e-05 ***
ridreth3Non-Hispanic Asian -3.371e-04  1.074e-04  -3.139 0.001761 **
ridreth3Other Race         -1.131e-04  1.638e-04  -0.690 0.490116
whq030Underweight          -4.211e-05  1.921e-04  -0.219 0.826505
whq030Overweight           -3.052e-04  8.148e-05  -3.745 0.000193 ***
whq070Attempted             3.193e-05  7.883e-05   0.405 0.685505
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001006 on 788 degrees of freedom
Multiple R-squared:  0.08058,   Adjusted R-squared:  0.06891
F-statistic: 6.906 on 10 and 788 DF,  p-value: 2.062e-10
```

Fitting/Summarizing the Kitchen Sink model (Continued)

Our big model predicts the inverse of `bpxsy1` using the predictors: `lbxvic`, `riagendr`, `whq030`, `whq070`, and `ridreth3`. We have provided a summary of our big model below. Based on the p-values, we can see that vitamin C level, sex (male), Non-Hispanic Black origin, Non-Hispanic Asian origin, and subjects who considered themselves overweight significantly predict first systolic blood pressure reading.

Effect Sizes: Coefficient Estimates

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	0.008537	0.000120	0.008339	0.008734	0.000000
lbxvic	0.000325	0.000085	0.000186	0.000465	0.000130
riagendrMale	-0.000273	0.000074	-0.000395	-0.000152	0.000220
ridreth3Mexican American	-0.000165	0.000110	-0.000346	0.000016	0.133471
ridreth3Other Hispanic	-0.000210	0.000129	-0.000422	0.000002	0.103675
ridreth3Non-Hispanic Black	-0.000467	0.000106	-0.000642	-0.000293	0.000012
ridreth3Non-Hispanic Asian	-0.000337	0.000107	-0.000514	-0.000160	0.001761
ridreth3Other Race	-0.000113	0.000164	-0.000383	0.000157	0.490116
whq030Underweight	-0.000042	0.000192	-0.000358	0.000274	0.826505
whq030Overweight	-0.000305	0.000081	-0.000439	-0.000171	0.000193
whq070Attempted	0.000032	0.000079	-0.000098	0.000162	0.685505

Effect Sizes: Coefficient Estimates (Continued)

- We used a 90% confidence level.
- Again, based on the p-values, we can see that vitamin C level, sex (male), Non-Hispanic Black origin, Non-Hispanic Asian origin, and Subjects who considered themselves overweight significantly predict first systolic blood pressure reading.
- The estimated coefficient of inverse transformed vitamin C level in predicting SBP reading is 0.000325.
- The estimated coefficient of being a male in predicting SBP reading is -0.000274.
- The estimated coefficient of being a Non-Hispanic Black in predicting SBP reading is -0.000467.
- The estimated coefficient of being a Non-Hispanic Asian in predicting SBP reading is -0.000337.
- The estimated coefficient of subjects who considered themselves overweight in predicting SBP reading is -0.000305.

Describing the Equation

We have provided the equation for predicting our outcome variable using our predictors below.

```
extract_eq(model_big, use_coefs = TRUE, coef_digits = 6,
           terms_per_line = 2, wrap = TRUE, ital_vars = TRUE)
```

$$\begin{aligned}\widehat{inv\_bpxsy1} = {} & 0.008537 + 0.000325(lbxvic) - \\ & 0.000273(riagendr_{Male}) - 0.000165(ridreth3_{Mexican\ American}) - \\ & 0.00021(ridreth3_{Other\ Hispanic}) - 0.000467(ridreth3_{Non-Hispanic\ Black}) - \\ & 0.000337(ridreth3_{Non-Hispanic\ Asian}) - 0.000113(ridreth3_{Other\ Race}) - \\ & 4.2e-05(whq030_{Underweight}) - 0.000305(whq030_{Overweight}) + \\ & 3.2e-05(whq070_{Attempted})\end{aligned}$$

Describing the Equation (Continued)

For this model, I will interpret all the variables that have p-value less than 0.1.

- For the key predictor, for every increase of one point in `lbxvic`, we expect an increase in the outcome or inverse of `bpxsy1` by 0.000325 (1/mmHg), with 90% confidence interval (0.000186, 0.000465).
- When holding all other predictors constant, we can say that for every increase of one point in `riagendrMale`, we expect an decrease in the outcome or inverse of `bpxsy1` by 0.000273 (1/mmHg), with 90% confidence interval (-0.000395, -0.000152).
- When holding all other predictors constant, we can say that for every increase of one point in `ridreth3Non-Hispanic Black`, we expect an decrease in the outcome or inverse of `bpxsy1` by 0.000467 (1/mmHg), with 90% confidence interval (-0.000642, -0.000293).
- When holding all other predictors constant, we can say that for every increase of one point in `ridreth3Non-Hispanic Asian`, we expect an decrease in the outcome or inverse of `bpxsy1` by 0.000337 (1/mmHg), with 90% confidence interval (-0.000514, -0.000160).
- When holding all other predictors constant, we can say that for every increase of one point in `whq030Overweight`, we expect an increase in the outcome or inverse of `bpxsy1` by 0.000305 (1/mmHg), with 90% confidence interval (-0.000439, -0.000171).

The Smaller Model

How We Built the Small Model

- We will build a smaller model using a subset of our big model predictors, chosen to maximize predictive value within our training sample.
- We will use the inverse transformed outcome.
- We will use a 90% confidence level for the analysis.

Backwards Stepwise Elimination

The backwards selection stepwise method suggests us to consider a model using predictors: `lbxvic`, `riagendr`, `ridreth3`, and `whq030`.

```
step(model_big)
```

```
Start:  AIC=-11018
inv_bpxsy1 ~ lbxvic + riagendr + ridreth3 + whq030 + whq070

           Df  Sum of Sq        RSS    AIC
- whq070    1 1.6610e-07 0.00079776 -11020
<none>                  0.00079760 -11018
- whq030    2 1.4541e-05 0.00081214 -11008
- riagendr  1 1.3944e-05 0.00081154 -11006
- ridreth3  5 2.2953e-05 0.00082055 -11005
- lbxvic    1 1.4964e-05 0.00081256 -11005

Step:  AIC=-11019.83
inv_bpxsy1 ~ lbxvic + riagendr + ridreth3 + whq030

           Df  Sum of Sq        RSS    AIC
<none>                  0.00079776 -11020
- whq030    2 1.5611e-05 0.00081337 -11008
- riagendr  1 1.4520e-05 0.00081228 -11007
- ridreth3  5 2.2787e-05 0.00082055 -11007
- lbxvic    1 1.5360e-05 0.00081312 -11007


Call:
lm(formula = inv_bpxsy1 ~ lbxvic + riagendr + ridreth3 + whq030,
    data = training_2)

Coefficients:
              (Intercept)                    lbxvic
```

Fitting the "small" model

We have provided a summary of our small model below. Based on the p-values, we can see that vitamin C level, sex (male), Non-Hispanic Black origin, Non-Hispanic Asian origin, and Subjects who considered themselves overweight significantly predict first systolic blood pressure reading.

```
Call:
lm(formula = inv_bpxsy1 ~ lbxvic + riagendr + ridreth3 + whq030,
    data = training_2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0039230 -0.0005921 -0.0000116  0.0006231  0.0032071

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.542e-03  1.192e-04  71.653  < 2e-16 ***
lbxvic                      3.284e-04  8.425e-05   3.898 0.000105 ***
riagendrMale               -2.770e-04  7.311e-05  -3.790 0.000162 ***
ridreth3Mexican American   -1.640e-04  1.099e-04  -1.492 0.136072
ridreth3Other Hispanic     -2.074e-04  1.286e-04  -1.613 0.107109
ridreth3Non-Hispanic Black -4.643e-04  1.058e-04  -4.388  1.3e-05 ***
ridreth3Non-Hispanic Asian -3.339e-04  1.070e-04  -3.119 0.001882 **
ridreth3Other Race         -1.108e-04  1.636e-04  -0.677 0.498499
whq030Underweight          -4.598e-05  1.917e-04  -0.240 0.810529
whq030Overweight           -2.932e-04  7.595e-05  -3.861 0.000122 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001006 on 789 degrees of freedom
Multiple R-squared:  0.08039,   Adjusted R-squared:  0.0699
F-statistic: 7.663 on 9 and 789 DF,  p-value: 7.668e-11
```

Effect Sizes: Coefficient Estimates

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	0.008542	0.000119	0.008346	0.008738	0.000000
lbxvic	0.000328	0.000084	0.000190	0.000467	0.000105
riagendrMale	-0.000277	0.000073	-0.000397	-0.000157	0.000162
ridreth3Mexican American	-0.000164	0.000110	-0.000345	0.000017	0.136072
ridreth3Other Hispanic	-0.000207	0.000129	-0.000419	0.000004	0.107109
ridreth3Non-Hispanic Black	-0.000464	0.000106	-0.000639	-0.000290	0.000013
ridreth3Non-Hispanic Asian	-0.000334	0.000107	-0.000510	-0.000158	0.001882
ridreth3Other Race	-0.000111	0.000164	-0.000380	0.000159	0.498499
whq030Underweight	-0.000046	0.000192	-0.000362	0.000270	0.810529
whq030Overweight	-0.000293	0.000076	-0.000418	-0.000168	0.000122

Effect Sizes: Coefficient Estimates (Continued)

- We used a 90% confidence level.
- Again, based on the p-values, we can see that vitamin C level, sex (male), Non-Hispanic Black origin, Non-Hispanic Asian origin, and Subjects who considered themselves overweight significantly predict first systolic blood pressure reading.
- The estimated coefficient of inverse transformed vitamin C level in predicting SBP reading is 0.000328.
- The estimated coefficient of being a male in predicting SBP reading is -0.000277.
- The estimated coefficient of being a Non-Hispanic Black in predicting SBP reading is -0.000464.
- The estimated coefficient of being a Non-Hispanic Asian in predicting SBP reading is -0.000334.
- The estimated coefficient of subjects who considered themselves overweight in predicting SBP reading is -0.000293.

Small Model Regression Equation

We have provided the equation for predicting our outcome variable using our predictors below.

```
extract_eq(model_small, use_coefs = TRUE, coef_digits = 6,
           terms_per_line = 2, wrap = TRUE, ital_vars = TRUE)
```

$$\begin{aligned}\widehat{inv\_bpxsy1} = {} & 0.008542 + 0.000328(lbxvic) - \\ & 0.000277(riagendr_{Male}) - 0.000164(ridreth3_{Mexican\ American}) - \\ & 0.000207(ridreth3_{Other\ Hispanic}) - 0.000464(ridreth3_{Non-Hispanic\ Black}) - \\ & 0.000334(ridreth3_{Non-Hispanic\ Asian}) - 0.000111(ridreth3_{Other\ Race}) - \\ & 4.6e-05(whq030_{Underweight}) - 0.000293(whq030_{Overweight})\end{aligned}$$

Small Model Regression Equation (Continued)

For the small model, I will also interpret all the variables that have p-value less than 0.1.

- For the key predictor, for every increase of one point in `lbxvic`, we expect an increase in the outcome or inverse of `bpxsy1` by 0.000328 (1/mmHg), with 90% confidence interval (0.000190, 0.000467).
- When holding all other predictors constant, we can say that for every increase of one point in `riagendrMale`, we expect an decrease in the outcome or inverse of `bpxsy1` by 0.000277 (1/mmHg), with 90% confidence interval (-0.000397, -0.000157).
- When holding all other predictors constant, we can say that for every increase of one point in `ridreth3Non-Hispanic Black`, we expect an decrease in the outcome or inverse of `bpxsy1` by 0.000464 (1/mmHg), with 90% confidence interval (-0.000639, -0.000290).
- When holding all other predictors constant, we can say that for every increase of one point in `ridreth3Non-Hispanic Asian`, we expect an decrease in the outcome or inverse of `bpxsy1` by 0.000334 (1/mmHg), with 90% confidence interval (-0.000510, -0.000158).
- When holding all other predictors constant, we can say that for every increase of one point in `whq030Overweight`, we expect an increase in the outcome or inverse of `bpxsy1` by 0.000293 (1/mmHg), with 90% confidence interval (-0.000418, -0.000168).

In-Sample Comparison

Quality of Fit

We will compare the two models built from our training sample using adjusted R-squared, the residual standard error, AIC and BIC.

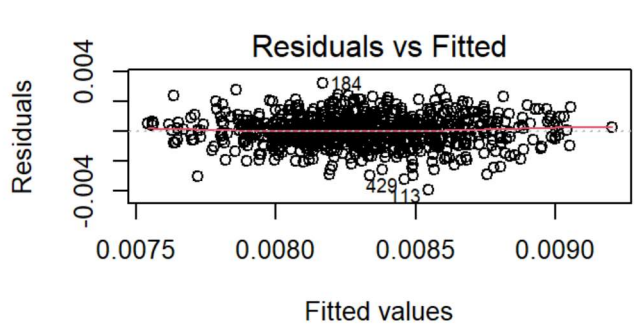
modelname	nobs	df	AIC	BIC	r.squared	adj.r.squared	sigma	statistic	p.value	df.residual
big	799	10	-8748.535	-8692.334	0.081	0.069	0.001	6.906	0	788
small	799	9	-8750.368	-8698.851	0.080	0.070	0.001	7.663	0	789

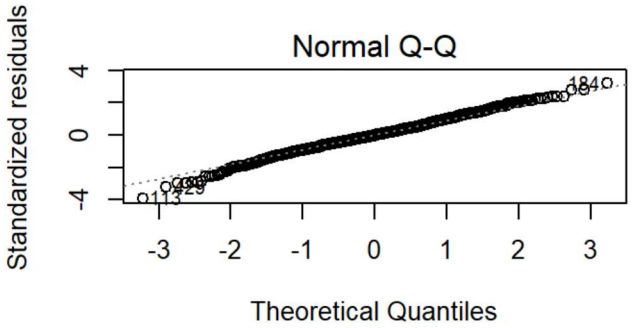
Quality of Fit(Continued)

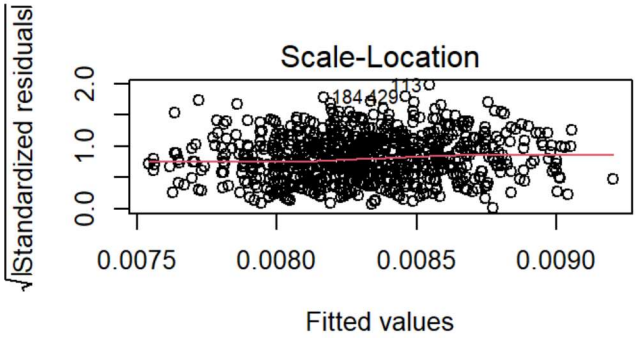
- The small model is better in terms of AIC and BIC because it has smaller values.
- The big model is slightly better in terms of R-squared value by 0.001 but the small model is slightly better in terms of adjusted R-squared value by 0.001.
- Overall, the small model with four predictors: `lbxvic`, `riagendr`, `ridreth3`, and `whq030`, performs slightly better in the training sample.

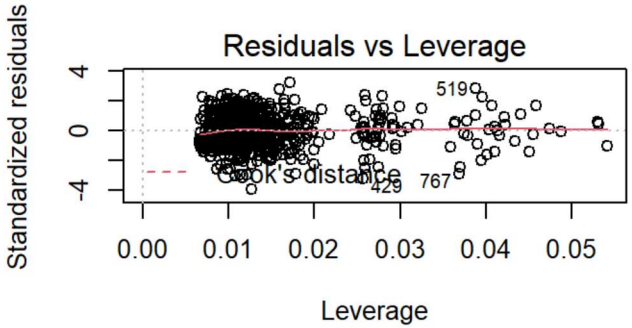
Assessing Assumptions

Residual Plots for the Big Model









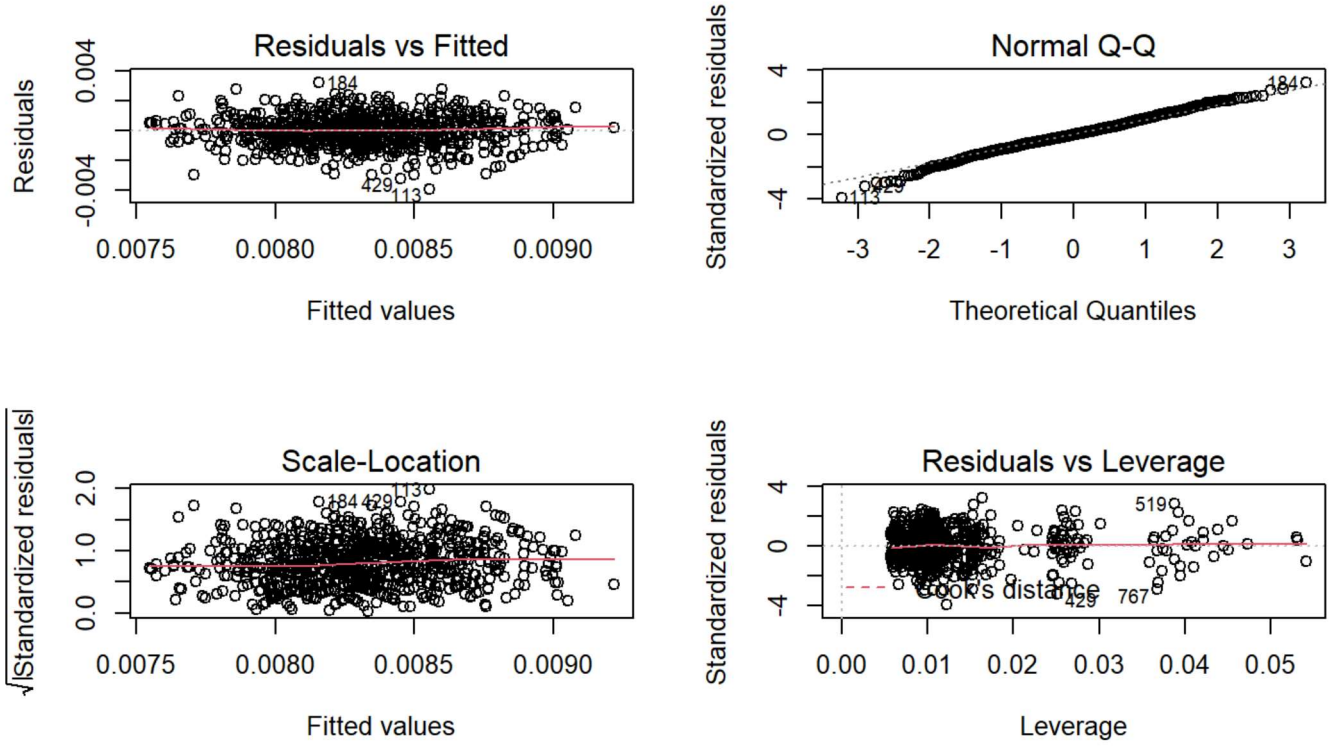
Residual Plots for the Big Model (Continued)

From the graphs we can conclude:

- From the top left graph, we see no substantial problems assuming linearity of the transformed data.
- From the top right graph, we see no substantial problems assuming normality of the transformed data.
- From the bottom left graph, we see no substantial problems assuming constant variance.
- From the bottom right graph, we see no highly leveraged points.

Residual Plots for the Small Model

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



Residual Plots for the Small Model (Continued)

Again, from the graphs we can conclude: - From the top left graph, we see no substantial problems assuming linearity of the transformed data. - From the top right graph, we see no substantial problems assuming normality of the transformed data. - From the bottom left graph, we see no substantial problems assuming constant variance. - From the bottom right graph, we see no highly leveraged points.

We did not see any substantial problems in terms of assumptions for both models.

Does collinearity have a meaningful impact?

```
car::vif(model_big)
```

```
             GVIF Df GVIF^(1/(2*Df))
lbxvic   1.089630  1        1.043853
riagendr 1.067850  1        1.033368
ridreth3 1.102919  5        1.009844
whq030   1.259695  2        1.059416
whq070   1.212397  1        1.101089
```

Since none of the generalized variance inflation factors is above 5, there is no two variables that are highly correlated with each other. Thus, we are not concerned about collinearity.

Comparing the Models

For the training sample, our small model performs slightly better in terms of adjusted R-squared, AIC, and BIC. Both models show no substantial problems with regression assumptions.

Model Validation

Which Model?

- We will use our two models: the big model and the small model, to predict the value of our outcome variable in the test sample, `test_2`.

Calculating Prediction Errors

Big Model: Back-Transformation and Calculating Fits/Residuals

- `.fitted` here tries to predict inverse of bpxsy1 instead of bpxsy1. Therefore, we inverse the `.fitted` value to back out of our inverse transformation.
- `bpxsy1_fit` is the estimated value using the big model for each subject in the test sample.
- `bpxsy1_res` is the prediction errors (observed bpxsy1 - estimated bpxsy1).

```
# A tibble: 3 x 11
  seqn   mod_name bpxsy1 bpxsy1_fit bpxsy1_res riagendr ridreth3  whq030  whq070
  <labe> <chr>     <dbl>      <dbl>      <dbl> <fct>    <fct>     <fct>   <fct>
1 93718  big         128       118.       9.84 Male     Non-Hisp~ About ~ Not A~
2 93761  big         132       125.       7.01 Male     Non-Hisp~ About ~ Attem~
3 93775  big         108       117.      -9.38 Female   Non-Hisp~ About ~ Not A~
# ... with 2 more variables: lbxvic <dbl>, .fitted <dbl>
```

Small Model: Back-Transformation and Calculating Fits/Residuals

We did the same method above again for the small model.

- `.fitted` here tries to predict inverse of bpxsy1 instead of bpxsy1. Therefore, we inverse the `.fitted` value to back out of our inverse transformation.
- `bpxsy1_fit` here is the estimated value using the small model for each subject in the test sample.
- `bpxsy1_res` is the prediction errors (observed bpxsy1 - estimated bpxsy1).

```
# A tibble: 3 x 11
  seqn   mod_name bpxsy1 bpxsy1_fit bpxsy1_res riagendr ridreth3  whq030  whq070
  <labe> <chr>     <dbl>      <dbl>      <dbl> <fct>    <fct>     <fct>   <fct>
1 93718  small       128       118.       9.99 Male     Non-Hisp~ About ~ Not A~
2 93761  small       132       125.       6.59 Male     Non-Hisp~ About ~ Attem~
3 93775  small       108       117.      -9.22 Female   Non-Hisp~ About ~ Not A~
# ... with 2 more variables: lbxvic <dbl>, .fitted <dbl>
```

Combining the Results

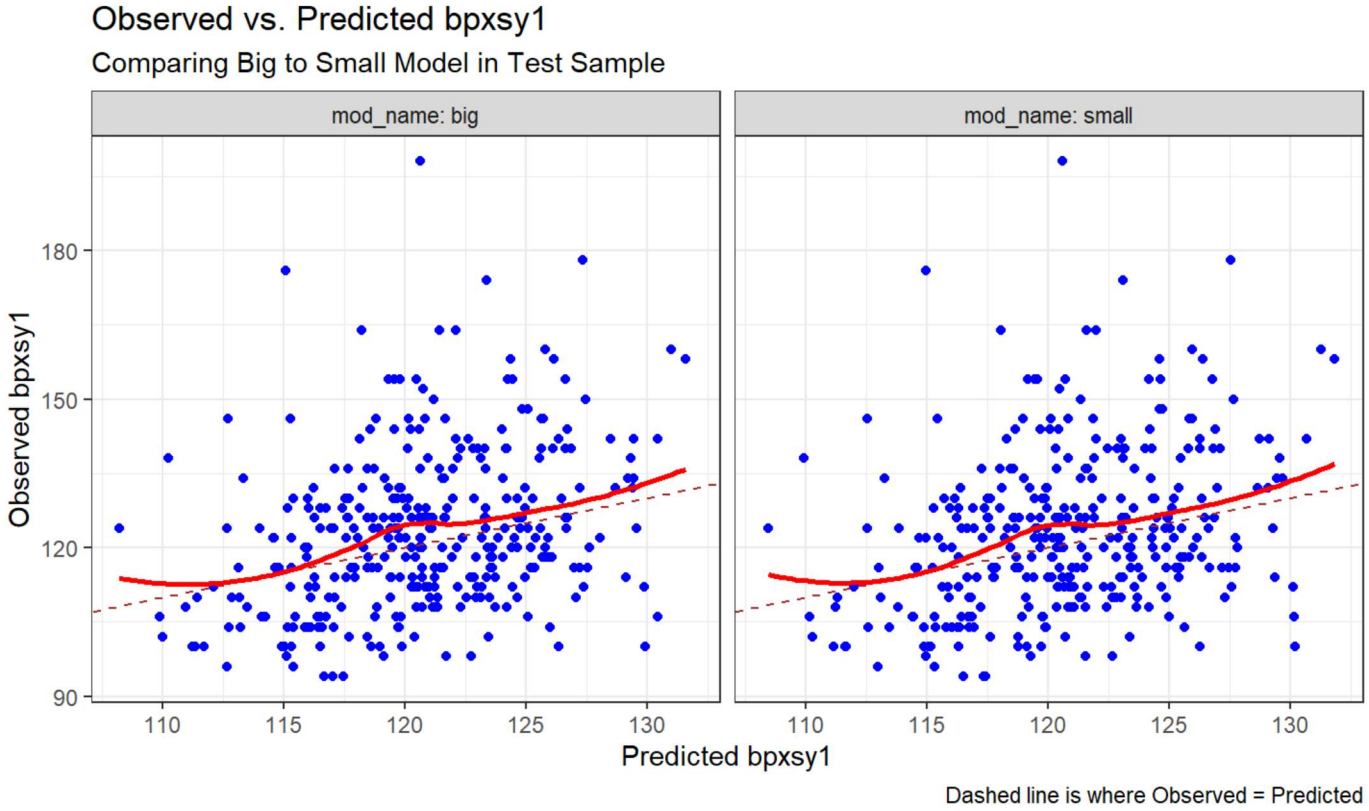
```
test_comp <- union(aug_big, aug_small) %>%
  arrange(seqn, mod_name)

test_comp %>% head()
```

```
# A tibble: 6 x 11
  seqn   mod_name bpxsy1 bpxsy1_fit bpxsy1_res riagendr ridreth3  whq030  whq070
  <labe> <chr>     <dbl>      <dbl>      <dbl> <fct>    <fct>     <fct>   <fct>
1 93718  big         128       118.       9.84 Male     Non-Hisp~ About ~ Not A~
2 93718  small       128       118.       9.99 Male     Non-Hisp~ About ~ Not A~
3 93761  big         132       125.       7.01 Male     Non-Hisp~ About ~ Attem~
4 93761  small       132       125.       6.59 Male     Non-Hisp~ About ~ Attem~
5 93775  big         108       117.      -9.38 Female   Non-Hisp~ About ~ Not A~
6 93775  small       108       117.      -9.22 Female   Non-Hisp~ About ~ Not A~
# ... with 2 more variables: lbxvic <dbl>, .fitted <dbl>
```

- We created a new tibble called `test_comp`. This tibble includes all the predictions and prediction errors using both the big and small models on our test data.
- We will then visualize the predictions by both models and compare their prediction errors to see which model performs better.

Visualizing the Predictions

From the two plots, we can see that the two models are fairly similar to each other in terms of predictions



and prediction errors.

Summarizing the Errors

- We have calculated the mean absolute prediction error (MAPE), the median absolute prediction error (medAPE), the maximum absolute prediction error (maxAPE), and the square root of the mean squared prediction error (RMSPE) for both models.

mod_name	n	MAPE	medAPE	maxAPE	RMSPE
big	343	11.79959	9.70106	77.36912	15.44282
small	343	11.79769	9.74910	77.39428	15.44203

Both models suggest an average error in predicting systolic blood pressure (using MAPE) of more than 11.80 mm Hg.

- In terms of MAPE and RMSPE, the small model performs slightly better.
- In terms of medAPE and maxAPE, the big model performs slightly better.
- We think MAPE is more important than medAPE for blood pressure prediction. Therefore, overall, we think the small model performs slightly better than the big model. However, both models do not differ from each other a lot.

Identify the largest errors

The first systolic reading of the subject with respondent sequence number: 97054 was poorly fitted by both models. The 97054 subject is a Non-Hispanic White male who considered himself about the right weight and did not attempt to lose weight in the past year.

```
temp1 <- aug_big %>%
  filter(abs(bpxsy1_res) == max(abs(bpxsy1_res)))

temp2 <- aug_small %>%
  filter(abs(bpxsy1_res) == max(abs(bpxsy1_res)))

bind_rows(temp1, temp2)
```

```
# A tibble: 2 x 11
  seqn   mod_name bpxsy1 bpxsy1_fit bpxsy1_res riagendr ridreth3  whq030  whq070
  <labe> <chr>     <dbl>      <dbl>      <dbl> <fct>    <fct>     <fct>   <fct>
1 97054  big         198       121.       77.4 Male     Non-Hisp~ About ~ Not A~
2 97054  small       198       121.       77.4 Male     Non-Hisp~ About ~ Not A~
# ... with 2 more variables: lbxvic <dbl>, .fitted <dbl>
```

Validated R-squared values

We calculated the R-squared values for both models, and the values do not vary a lot from each other. The big model has a slightly higher R-squared value, but this is possibly due to using more predictors.

```
aug_big %$% cor(bpxsy1, bpxsy1_fit)^2
```

```
[1] 0.0940192
```

```
aug_small %$% cor(bpxsy1, bpxsy1_fit)^2
```

```
[1] 0.09379321
```

Comparing the Models

In conclusion, we wanted to select the small model based on its performance on mean absolute prediction error, square root of the mean squared prediction error, AIC, BIC, and adjusted R-squared value.

Discussion

Chosen Model

Even though both models were similar, we chose the small model because of its performance on mean absolute prediction error, square root of the mean squared prediction error, AIC, BIC, and adjusted R-squared value.

Answering My Question

Vitamin C level, gender (male), subjects' consideration of their current weight (considered overweight), and race origin (Non-Hispanic Black and Non-Hispanic Asian) are all variables that can meaningfully and effectively predict the subjects' first systolic blood pressure readings in the `final_2` data. We were surprised by the association between SBP reading and vitamin C level because we anticipated that an increase in vitamin C level would decrease SBP level; however, our model showed that by holding other predictor variables constant, an increase in vitamin C level also increases SBP level.

Next Steps

One major limitation of our study was that we did not know if the association discovered was consistent over the years. We would want to do the same analysis in other years to see if we would get similar results. Additionally, even though there were no highly influential points in our data, we still had many outliers that could be removed from the data. This might affect findings in the study, but removing outliers would introduce bias into our study. A logical next step would be to categorize vitamin C levels into “low vitamin C level,” “optimal vitamin C level,” and “high vitamin C level” to fully understand the association between SBP levels and vitamin C levels.

Reflection

If we knew how to do multiple imputations prior to completing study 2, we would have kept all the missing values and done imputations on them. This way we will have a larger sample size and potentially better understand the association between systolic blood pressure and vitamin C level. Additionally, we might want to treat outliers as missing values and do imputations on them as well.