Final Exam Review ALM 2019

(Section 1) Predicting ear infections by age, sex, race-ethnicity

The data management, exploratory data analysis, and statistical modeling in this section is used to answer the research question:

Does knowing the age, sex, and race-ethnicity of a person help to predict whether or not the person has ever had 3 or more ear infections?

Codebook (NHANES 2011-2012)

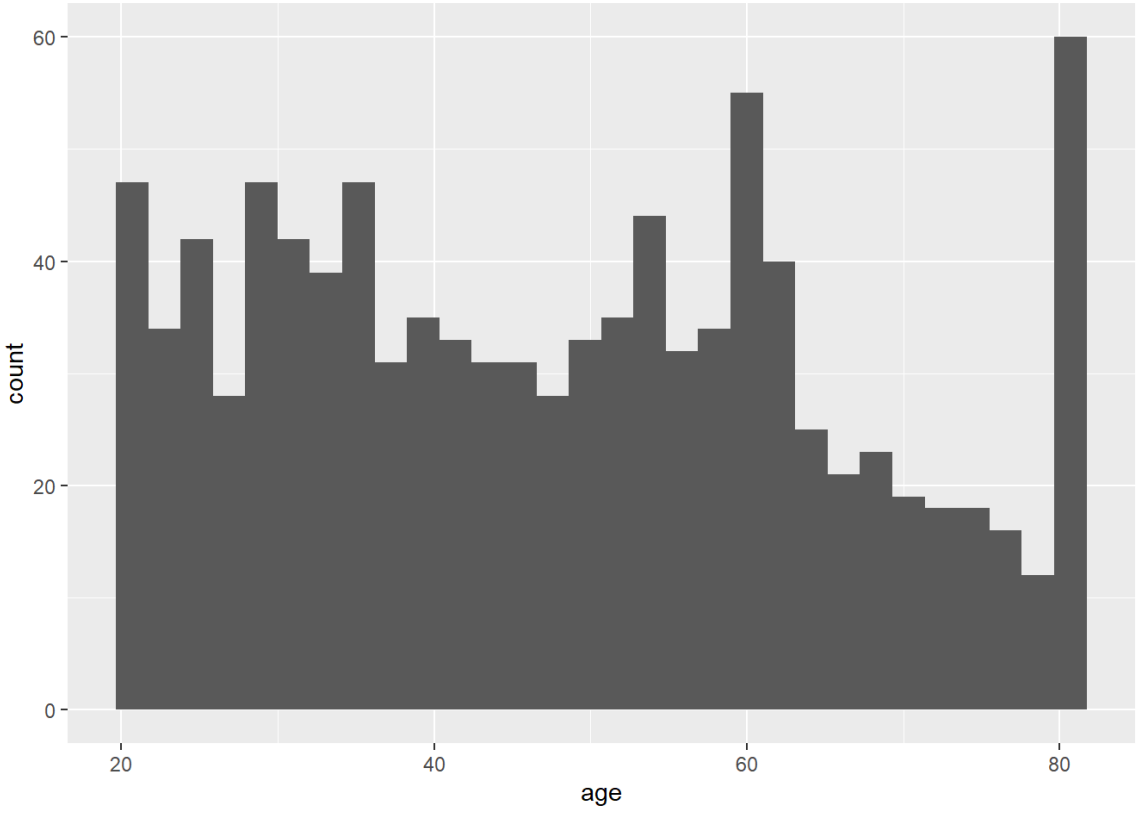
- AUQ136: Ever had 3 or more ear infections
  - 1 = Yes, 2 = No, 7 = Refused, 9 = Don't know
- RIAGENDR: Gender of the participant
  - 1 = Male, 2 = Female
- RIDAGEYR: Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
- RIDRETH3: Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category
  - 1 = Mexican American, 2 = Other Hispanic, 3 = Non-Hispanic White, 4 = Non-Hispanic Black, 6 = Non-Hispanic Asian, 7 = Other Race - Including Multi-Racial

Data management

```
# import NHANES 2011-2012 data
library(package = RNHANES)
nhanes2012 <- nhanes_load_data(file_name = "AUQ_G",
                               year = '2011-2012',
                               demographics = TRUE)

# select variables for ear infections, age, sex, race
# add labels to ear infections, race, sex
library(package = "tidyverse")
nhanes2012.clean <- nhanes2012 %>%
  select(AUQ136, RIAGENDR, RIDAGEYR, RIDRETH3) %>%
  mutate(ear.infect = recode_factor(AUQ136, `1` = "Yes",
                                    `2` = "No",
                                    `7` = NA_character_,
                                    `9` = NA_character_)) %>%
  rename(age = RIDAGEYR) %>%
  mutate(race.eth = recode_factor(RIDRETH3, `1` = "Mexican American",
                            `2` = "Other Hispanic",
                            `3` = "Non-Hispanic White",
                            `4` = "Non-Hispanic Black",
                            `6` = "Non-Hispanic Asian",
                            `7` = "Other Race - Including Multi-Racial")) %>%
  mutate(sex = recode_factor(RIAGENDR, `1` = "Male", `2` = "Female")) %>%
  drop_na(ear.infect) %>%
  group_by(ear.infect) %>%
  sample_n(500) %>%
  ungroup() %>%
  select(ear.infect, age, race.eth, sex)

# change reference group to Non-Hispanic Black and Female
nhanes2012.clean$race.eth <- relevel(nhanes2012.clean$race.eth,
                                     ref = "Non-Hispanic Black")
nhanes2012.clean$sex <- relevel(nhanes2012.clean$sex,
                                     ref = "Female")
```

Exploratory data analysis

```
# check distribution of age
nhanes2012.clean %>%
  ggplot(aes(x = age)) +
  geom_histogram()
```




```
# make a table of descriptives
library(package = "tableone")
desc.table <- CreateTableOne(data = nhanes2012.clean, strata = "ear.infect",
                             vars = c('age', 'sex', 'race.eth'))
print(desc.table, nonnormal = "age", showAllLevels = TRUE)
```

```
##                     Stratified by ear.infect
##                      level                              Yes                 No                  p      test
##   n                                                      500                 500
##   age (median [IQR])                                 46.00 [32.00, 60.00] 49.00 [34.00, 64.00]  0.010 nonn
orm
##   sex (%)            Female                            278 (55.6)          254 (50.8)          0.145
##                      Male                              222 (44.4)          246 (49.2)
##   race.eth (%)       Non-Hispanic Black                 70 (14.0)          149 (29.8)         <0.001
##                      Mexican American                   55 (11.0)           51 (10.2)
##                      Other Hispanic                     46 ( 9.2)           50 (10.0)
##                      Non-Hispanic White                268 (53.6)          163 (32.6)
##                      Non-Hispanic Asian                 35 ( 7.0)           82 (16.4)
##                      Other Race - Including Multi-Racial    26 ( 5.2)            5 ( 1.0)
```

Estimate the model

```
# recode so yes = 1 and no = 0
nhanes2012.clean <- nhanes2012.clean %>%
  mutate(ear.infect.num = recode(ear.infect,
                                 `Yes` = 1,
                                 `No` = 0))

# check recoding
table(nhanes2012.clean$ear.infect.num, nhanes2012.clean$ear.infect)
```

```
##
##     Yes  No
##   0   0 500
##   1 500   0
```

```
# model ear infection by age, sex, race
library(package = "odds.n.ends")
ear.inf.mod <- glm(ear.infect.num ~ sex + age + race.eth,
                   data = nhanes2012.clean)
odds.n.ends(ear.inf.mod)
```

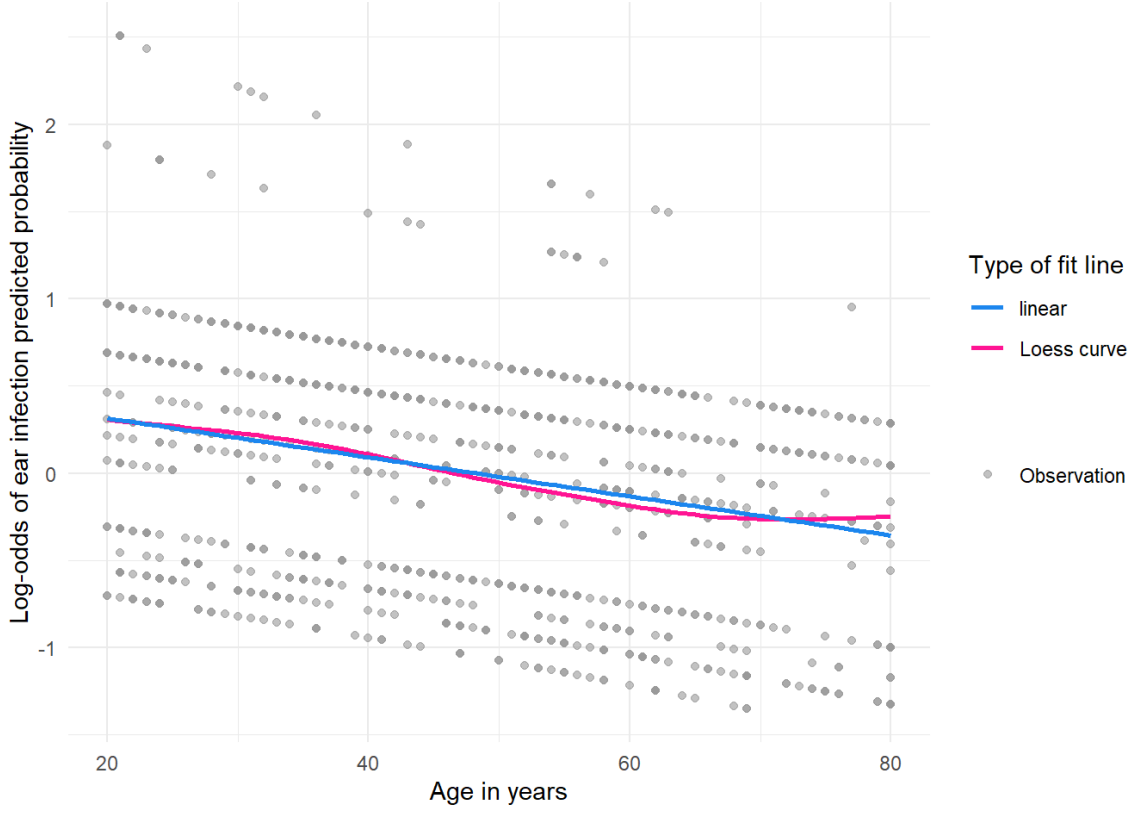
```
## $`Logistic regression model significance`
## Chi-squared        d.f.           p
##      24.634       7.000       0.001
##
## $`Contingency tables (model fit): percent predicted`
##                   Percent observed
## Percent predicted     1     0   Sum
##               1   0.349 0.219 0.568
##               0   0.151 0.281 0.432
##               Sum 0.500 0.500 1.000
##
## $`Contingency tables (model fit): frequency predicted`
##                 Number observed
## Number predicted    1    0  Sum
##              1    349  219  568
##              0    151  281  432
##              Sum  500  500 1000
##
## $`Predictor odds ratios and 95% CI`
##                                                    OR 2.5 % 97.5 %
## (Intercept)                                     1.607 1.441  1.793
## sexMale                                         0.942 0.888  1.000
## age                                             0.997 0.996  0.999
## race.ethMexican American                        1.209 1.082  1.351
## race.ethOther Hispanic                          1.166 1.040  1.307
## race.ethNon-Hispanic White                      1.352 1.251  1.461
## race.ethNon-Hispanic Asian                      0.968 0.870  1.078
## race.ethOther Race - Including Multi-Racial 1.655 1.383  1.981
##
## $`Model sensitivity`
## [1] 0.698
##
## $`Model specificity`
## [1] 0.562
```

Assumptions & Diagnostics

```
# get the VIFs
car::vif(ear.inf.mod)
```

```
##          GVIF Df GVIF^(1/(2*Df))
## sex      1.01  1            1.00
## age      1.02  1            1.01
## race.eth 1.02  5            1.00
```

```
# make a variable of the log odds of the predicted probabilities
logodds.use <- log(x = ear.inf.mod$fitted.values/(1-ear.inf.mod$fitted.values))

# make a small data frame with the log odds variable and the age predictor
graph.data <- data.frame(logodds.use, age = ear.inf.mod$model$age)

# graph the logit with age
graph.data %>%
  ggplot(aes(x = age, y = logodds.use))+
  geom_point(aes(size = "Observation"), color = "gray60", alpha = .6) +
  geom_smooth(se = FALSE, aes(color = "Loess curve")) +
  geom_smooth(method = lm, se = FALSE, aes(color = "linear")) +
  theme_minimal() +
  labs(x = "Age in years", y = "Log-odds of ear infection predicted probability") +
  scale_color_manual(name="Type of fit line", values=c("dodgerblue2","deeppink")) +
  scale_size_manual(values = 1.5, name = "")
```



```
# change the cutoff for leverage to reflect the 8 parameters in the model
nhanes.cleaned.diag <- nhanes2012.clean %>%
drop_na() %>%
mutate(standardres = rstandard(model = ear.inf.mod)) %>%
mutate(cooks.dist = cooks.distance(model = ear.inf.mod)) %>%
mutate(lever = hatvalues(model = ear.inf.mod)) %>%
mutate(outlier.infl = as.numeric(x = lever > 2*8/n()) +
as.numeric(x = cooks.dist > 4/n()) +
as.numeric(x = abs(x = standardres) > 1.96))

# examine the outliers & influential
nhanes.cleaned.diag %>%
  select(ear.infect, age, race.eth, sex, outlier.infl) %>%
  filter(outlier.infl >= 2)
```

```
## # A tibble: 5 x 5
##   ear.infect   age race.eth                             sex    outlier.infl
##   <fct>      <dbl> <fct>                                <fct>         <dbl>
## 1 No            56 Other Race - Including Multi-Racial Male              2
## 2 No            58 Other Race - Including Multi-Racial Male              2
## 3 No            36 Other Race - Including Multi-Racial Female            2
## 4 No            28 Other Race - Including Multi-Racial Male              2
## 5 No            24 Other Race - Including Multi-Racial Male              2
```

(Section 2) Examining mean systolic blood pressure by sex and race-ethnicity

The data management, exploratory data analysis, and statistical modeling in this section is used to answer the research question:

Is there a difference in mean systolic blood pressure by race-ethnicity, sex, and the interaction between the two?

Codebook

- BPXSY1: Systolic blood pressure in mmHg
- RIAGENDR: Gender of the participant
  - 1 = Male, 2 = Female
- RIDRETH3: Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category
  - 1 = Mexican American, 2 = Other Hispanic, 3 = Non-Hispanic White, 4 = Non-Hispanic Black, 6 = Non-Hispanic Asian, 7 = Other Race - Including Multi-Racial

```
# import nhanes 2013-2014 BPX file with demographics
nhanes.2014 <- nhanes_load_data(file_name = "BPX",
                                      year = "2013-2014",
                                      demographics = TRUE)
```

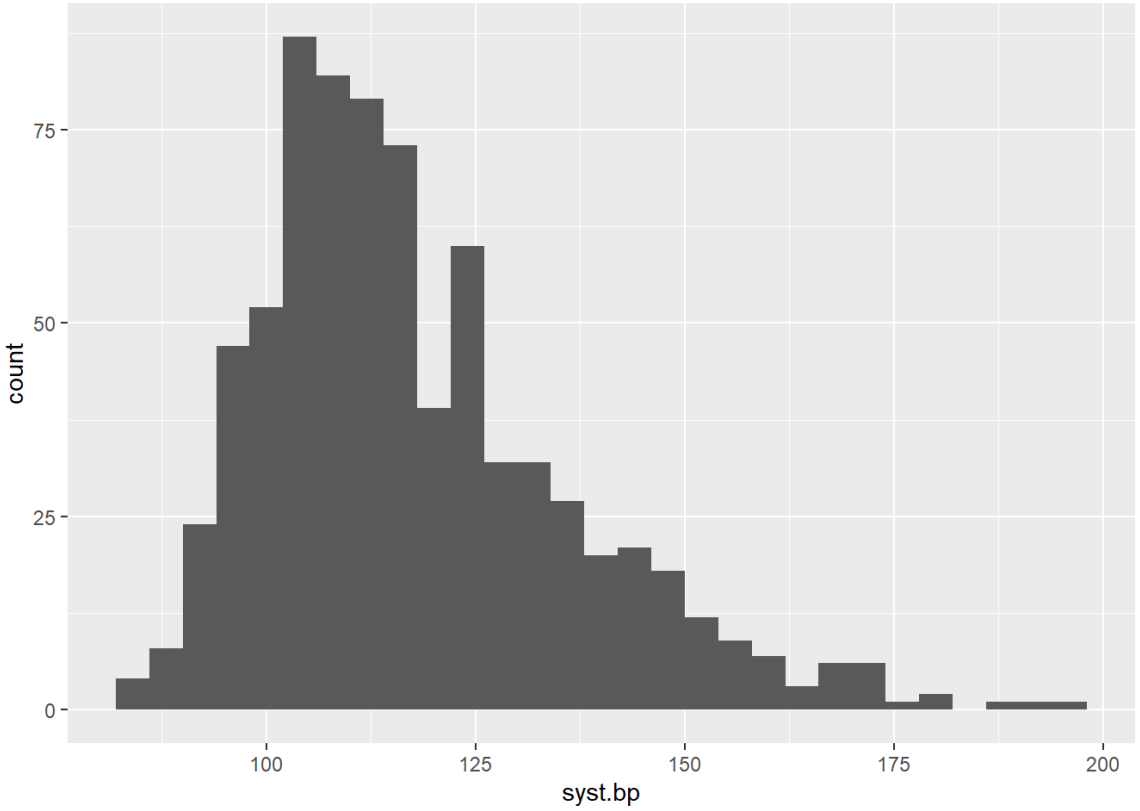
Data management

```
# add labels to factors
# make better variable names
# take a sample of 1000 observations
nhanes.2014.clean <- nhanes.2014 %>%
  select(RIAGENDR, RIDRETH3, BPXSY1) %>%
  mutate(race.eth = recode_factor(RIDRETH3, `1` = "Mexican American",
                           `2` = "Other Hispanic",
                           `3` = "Non-Hispanic White",
                           `4` = "Non-Hispanic Black",
                           `6` = "Non-Hispanic Asian",
                           `7` = "Other Race - Including Multi-Racial")) %>%
  mutate(sex = recode_factor(RIAGENDR, `1` = "Male", `2` = "Female")) %>%
  rename(syst.bp = BPXSY1) %>%
  sample_n(1000) %>%
  select(syst.bp, race.eth, sex)

# check out the data
summary(object = nhanes.2014.clean)
```

```
##     syst.bp                                   race.eth       sex
##  Min.   : 82   Mexican American                  :157   Male  :474
##  1st Qu.:106   Other Hispanic                   : 96   Female:526
##  Median :114   Non-Hispanic White               :411
##  Mean   :119   Non-Hispanic Black               :211
##  3rd Qu.:128   Non-Hispanic Asian               : 85
##  Max.   :198   Other Race - Including Multi-Racial: 40
##  NA's   :246
```

Exploratory data analysis

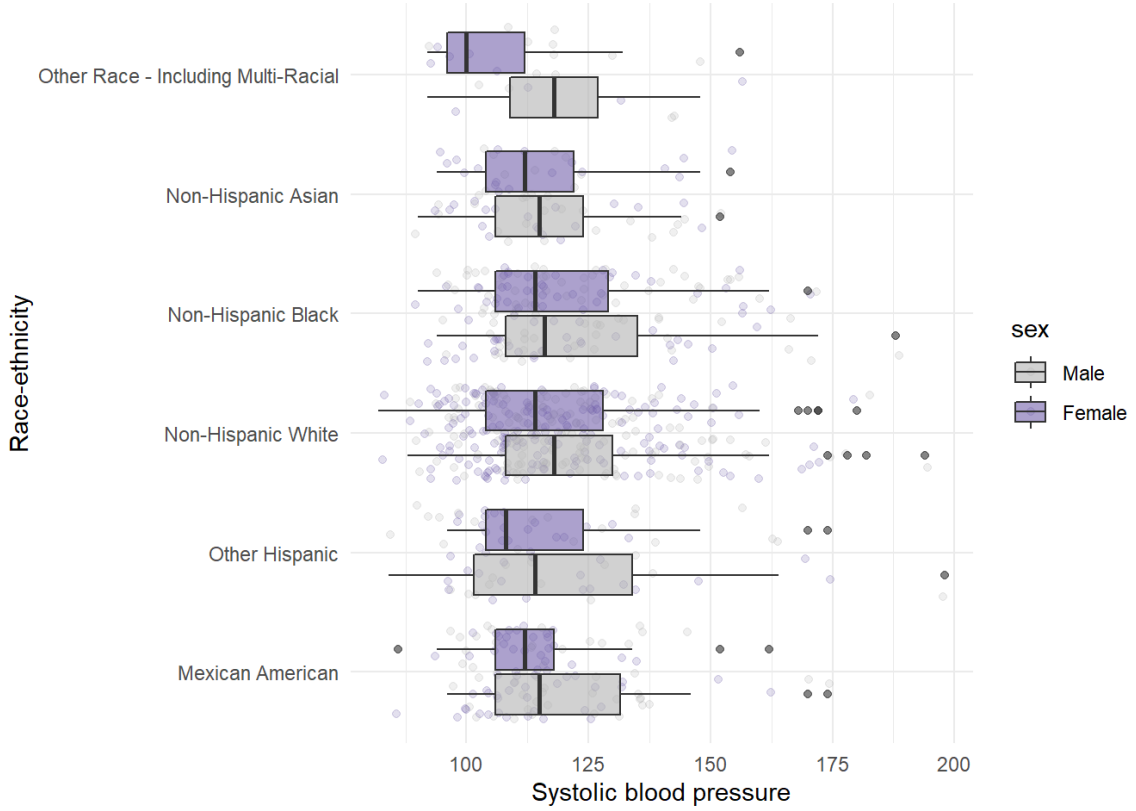
```
# distribution of blood pressure
nhanes.2014.clean %>%
  ggplot(aes(x = syst.bp)) +
  geom_histogram()
```




```
# table of descriptives
CreateTableOne(data = nhanes.2014.clean)
```

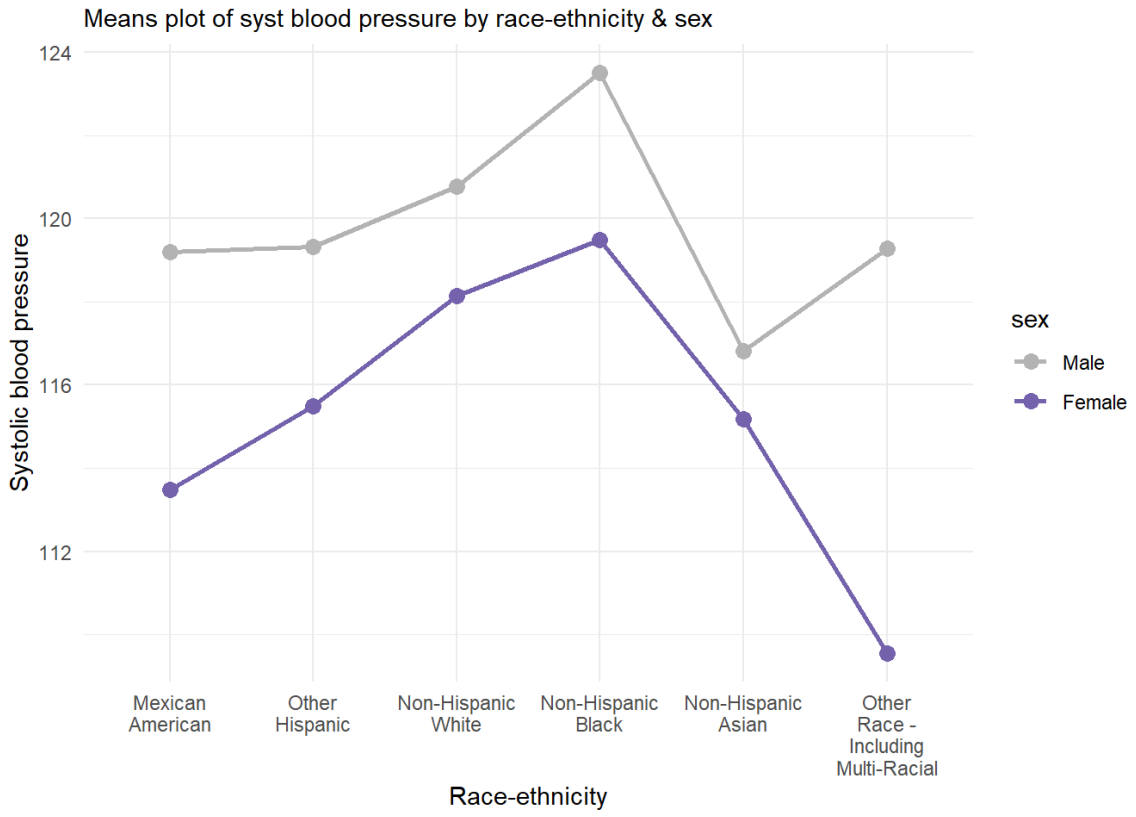
```
##
##                                      Overall
##   n                                    1000
##   syst.bp (mean (SD))                118.70 (18.66)
##   race.eth (%)
##      Mexican American                   157 (15.7)
##      Other Hispanic                      96 ( 9.6)
##      Non-Hispanic White                 411 (41.1)
##      Non-Hispanic Black                 211 (21.1)
##      Non-Hispanic Asian                  85 ( 8.5)
##      Other Race - Including Multi-Racial 40 ( 4.0)
##   sex = Female (%)                      526 (52.6)
```

```
# examine blood pressure by sex & race-eth
nhanes.2014.clean %>%
ggplot(aes(y = syst.bp, x = race.eth)) +
  geom_jitter(aes(color = sex), alpha = .2) +
  geom_boxplot(aes(fill = sex), alpha = .6) +
  scale_fill_manual(values = c("gray70", "#7463AC")) +
  scale_color_manual(values = c("gray70", "#7463AC")) +
  theme_minimal() +
  coord_flip() +
  labs(x = "Race-ethnicity", y = "Systolic blood pressure")
```



Means plot

```
library(scales)
# means plot of race-eth, sex, blood press
nhanes.2014.clean %>%
  ggplot(aes(y = syst.bp, x = race.eth, color = sex)) +
  stat_summary(fun.y = mean, geom="point", size = 3) +
  stat_summary(fun.y = mean, geom="line", aes(group = sex), size = 1) +
  scale_color_manual(values = c("gray70", "#7463AC")) +
  theme_minimal() +
  labs(x = "Race-ethnicity",
       y = "Systolic blood pressure",
       subtitle = "Means plot of syst blood pressure by race-ethnicity & sex") +
    scale_x_discrete(labels = wrap_format(10))
```




```
# means table
syst.bp.stats.2 <- nhanes.2014.clean %>%
  group_by(race.eth, sex) %>%
  drop_na(syst.bp) %>%
  summarize(m.bp = mean(x = syst.bp),
            sd.bp = sd(x = syst.bp))
syst.bp.stats.2
```

```
## # A tibble: 12 x 4
## # Groups:   race.eth [6]
##    race.eth                           sex     m.bp sd.bp
##    <fct>                              <fct>  <dbl> <dbl>
##  1 Mexican American                   Male    119.  17.2
##  2 Mexican American                   Female  113.  12.5
##  3 Other Hispanic                     Male    119.  25.2
##  4 Other Hispanic                     Female  116.  18.0
##  5 Non-Hispanic White                 Male    121.  18.7
##  6 Non-Hispanic White                 Female  118.  19.3
##  7 Non-Hispanic Black                 Male    123.  20.4
##  8 Non-Hispanic Black                 Female  119.  18.7
##  9 Non-Hispanic Asian                 Male    117.  15.4
## 10 Non-Hispanic Asian                 Female  115.  16.9
## 11 Other Race - Including Multi-Racial Male    119.  16.1
## 12 Other Race - Including Multi-Racial Female  110.  21.3
```

```
# means table by sex only
syst.bp.stats.3 <- nhanes.2014.clean %>%
  group_by(sex) %>%
  drop_na(syst.bp) %>%
  summarize(m.bp = mean(x = syst.bp),
            sd.bp = sd(x = syst.bp))
syst.bp.stats.3
```

```
## # A tibble: 2 x 3
##   sex     m.bp sd.bp
##   <fct>  <dbl> <dbl>
## 1 Male    121.  19.2
## 2 Female  117.  18.0
```

```
# means table by race-eth only
syst.bp.stats.4 <- nhanes.2014.clean %>%
  group_by(race.eth) %>%
  drop_na(syst.bp) %>%
  summarize(m.bp = mean(x = syst.bp),
            sd.bp = sd(x = syst.bp))
syst.bp.stats.4
```

```
## # A tibble: 6 x 3
##   race.eth                              m.bp sd.bp
##   <fct>                                <dbl> <dbl>
## 1 Mexican American                      116.  15.0
## 2 Other Hispanic                        117.  21.4
## 3 Non-Hispanic White                    119.  19.0
## 4 Non-Hispanic Black                    122.  19.7
## 5 Non-Hispanic Asian                    116.  16.1
## 6 Other Race - Including Multi-Racial   115.  18.5
```

```
# blood pressure by race-eth and sex
bp.by.raceth.sex <- aov(formula = syst.bp ~ race.eth + sex + race.eth * sex,
                        data = nhanes.2014.clean)
summary(bp.by.raceth.sex)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race.eth       5   3167     633    1.83 0.1039
## sex            1   2424    2424    7.02 0.0082 **
## race.eth:sex   5    480      96    0.28 0.9253
## Residuals    742 256241     345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 246 observations deleted due to missingness
```

```
# get effect size
library(package = "sjstats")
omega_sq(model = bp.by.raceth.sex)
```

```
##           term omegasq
## 1     race.eth   0.005
## 2          sex   0.008
## 3 race.eth:sex  -0.005
```

```
# Tukey's HSD post-hoc test
TukeyHSD(x = bp.by.raceth.sex)
```

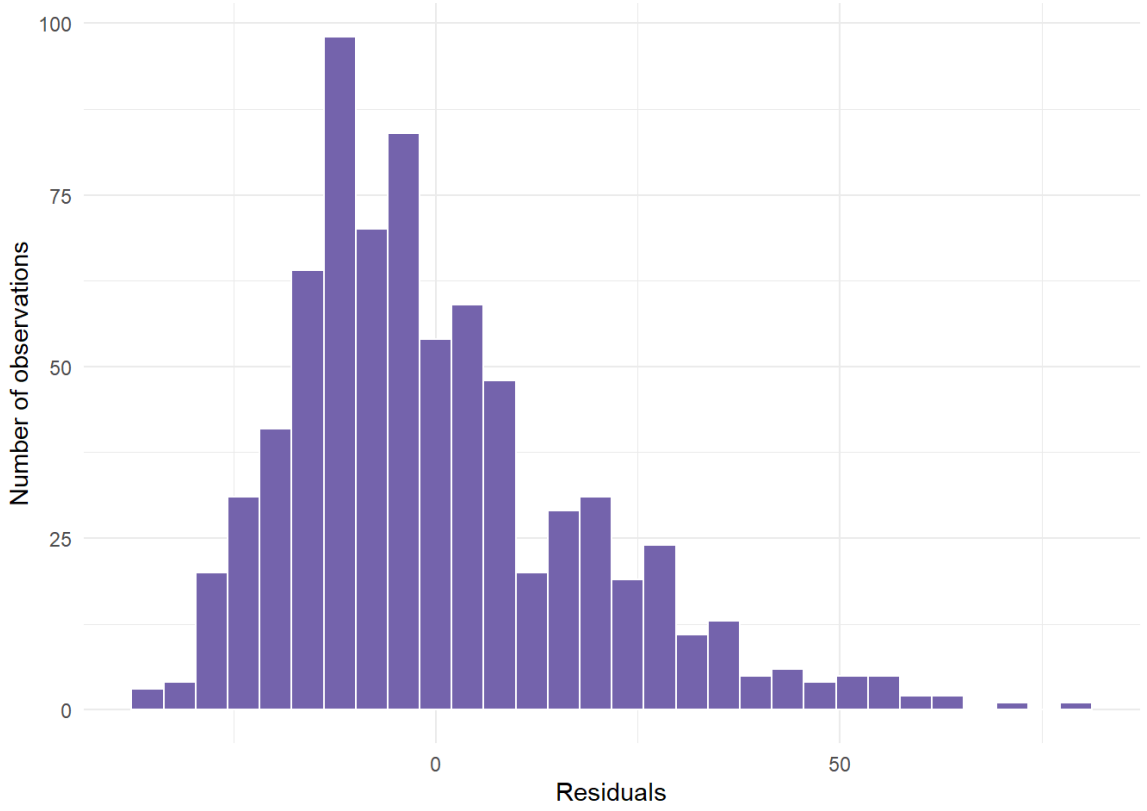
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = syst.bp ~ race.eth + sex + race.eth * sex, data = nhanes.2014.clean)
##
## $race.eth
##                                                      diff    lwr   upr p adj
## Other Hispanic-Mexican American                     1.140  -6.89  9.17 0.999
## Non-Hispanic White-Mexican American                 3.304  -2.54  9.15 0.589
## Non-Hispanic Black-Mexican American                 5.541  -1.03 12.12 0.155
## Non-Hispanic Asian-Mexican American                -0.112  -8.25  8.03 1.000
## Other Race - Including Multi-Racial-Mexican American   -0.576 -12.74 11.59 1.000
## Non-Hispanic White-Other Hispanic                   2.164  -4.76  9.09 0.948
## Non-Hispanic Black-Other Hispanic                   4.400  -3.15 11.95 0.555
## Non-Hispanic Asian-Other Hispanic                  -1.252 -10.20  7.69 0.999
## Other Race - Including Multi-Racial-Other Hispanic     -1.716 -14.43 11.00 0.999
## Non-Hispanic Black-Non-Hispanic White               2.237  -2.92  7.40 0.818
## Non-Hispanic Asian-Non-Hispanic White              -3.416 -10.46  3.63 0.736
## Other Race - Including Multi-Racial-Non-Hispanic White -3.880 -15.34  7.58 0.928
## Non-Hispanic Asian-Non-Hispanic Black              -5.653 -13.31  2.01 0.284
## Other Race - Including Multi-Racial-Non-Hispanic Black -6.117 -17.97  5.73 0.681
## Other Race - Including Multi-Racial-Non-Hispanic Asian -0.464 -13.25 12.32 1.000
##
## $sex
##              diff   lwr   upr p adj
## Female-Male -3.58 -6.24 -0.92 0.008
##
## $`race.eth:sex`
##                                                                                  diff    lwr    upr p a
dj
## Other Hispanic:Male-Mexican American:Male                                       0.1125 -13.68 13.905 1.0
00
## Non-Hispanic White:Male-Mexican American:Male                                   1.5651  -8.39 11.523 1.0
00
## Non-Hispanic Black:Male-Mexican American:Male                                   4.2940  -6.61 15.201 0.9
80
## Non-Hispanic Asian:Male-Mexican American:Male                                  -2.3875 -16.18 11.405 1.0
00
## Other Race - Including Multi-Racial:Male-Mexican American:Male                  0.0857 -18.34 18.508 1.0
00
## Mexican American:Female-Mexican American:Male                                 -5.7246 -17.35  5.898 0.9
04
## Other Hispanic:Female-Mexican American:Male                                   -3.7000 -16.62  9.224 0.9
99
## Non-Hispanic White:Female-Mexican American:Male                               -1.0605 -10.85  8.728 1.0
00
## Non-Hispanic Black:Female-Mexican American:Male                                0.2933 -10.83 11.417 1.0
00
## Non-Hispanic Asian:Female-Mexican American:Male                               -4.0108 -17.22  9.201 0.9
98
## Other Race - Including Multi-Racial:Female-Mexican American:Male              -9.6444 -31.71 12.416 0.9
57
## Non-Hispanic White:Male-Other Hispanic:Male                                    1.4526 -10.42 13.323 1.0
00
## Non-Hispanic Black:Male-Other Hispanic:Male                                    4.1815  -8.50 16.859 0.9
95
## Non-Hispanic Asian:Male-Other Hispanic:Male                                   -2.5000 -17.73 12.732 1.0
00
## Other Race - Including Multi-Racial:Male-Other Hispanic:Male                  -0.0268 -19.55 19.496 1.0
00
## Mexican American:Female-Other Hispanic:Male                                   -5.8371 -19.14  7.461 0.9
55
```

```
## Other Hispanic:Female-Other Hispanic:Male                                    -3.8125 -18.26 10.637 0.9
99
## Non-Hispanic White:Female-Other Hispanic:Male                                -1.1730 -12.90 10.557 1.0
00
## Non-Hispanic Black:Female-Other Hispanic:Male                                 0.1808 -12.68 13.045 1.0
00
## Non-Hispanic Asian:Female-Other Hispanic:Male                                -4.1233 -18.83 10.585 0.9
99
## Other Race - Including Multi-Racial:Female-Other Hispanic:Male               -9.7569 -32.74 13.231 0.9
65
## Non-Hispanic Black:Male-Non-Hispanic White:Male                               2.7289  -5.62 11.074 0.9
96
## Non-Hispanic Asian:Male-Non-Hispanic White:Male                              -3.9526 -15.82  7.918 0.9
95
## Other Race - Including Multi-Racial:Male-Non-Hispanic White:Male             -1.4794 -18.51 15.552 1.0
00
## Mexican American:Female-Non-Hispanic White:Male                              -7.2897 -16.55  1.971 0.2
93
## Other Hispanic:Female-Non-Hispanic White:Male                                -5.2651 -16.11  5.584 0.9
12
## Non-Hispanic White:Female-Non-Hispanic White:Male                            -2.6256  -9.44  4.193 0.9
83
## Non-Hispanic Black:Female-Non-Hispanic White:Male                            -1.2718  -9.90  7.354 1.0
00
## Non-Hispanic Asian:Female-Non-Hispanic White:Male                            -5.5759 -16.77  5.615 0.8
96
## Other Race - Including Multi-Racial:Female-Non-Hispanic White:Male          -11.2095 -32.12  9.704 0.8
40
## Non-Hispanic Asian:Male-Non-Hispanic Black:Male                              -6.6815 -19.36  5.996 0.8
55
## Other Race - Including Multi-Racial:Male-Non-Hispanic Black:Male             -4.2083 -21.81 13.395 1.0
00
## Mexican American:Female-Non-Hispanic Black:Male                             -10.0186 -20.29  0.256 0.0
64
## Other Hispanic:Female-Non-Hispanic Black:Male                                -7.9940 -19.72  3.733 0.5
25
## Non-Hispanic White:Female-Non-Hispanic Black:Male                            -5.3544 -13.50  2.788 0.5
82
## Non-Hispanic Black:Female-Non-Hispanic Black:Male                            -4.0006 -13.71  5.706 0.9
72
## Non-Hispanic Asian:Female-Non-Hispanic Black:Male                            -8.3048 -20.35  3.739 0.5
06
## Other Race - Including Multi-Racial:Female-Non-Hispanic Black:Male          -13.9384 -35.32  7.443 0.5
96
## Other Race - Including Multi-Racial:Male-Non-Hispanic Asian:Male              2.4732 -17.05 21.996 1.0
00
## Mexican American:Female-Non-Hispanic Asian:Male                              -3.3371 -16.64  9.961 1.0
00
## Other Hispanic:Female-Non-Hispanic Asian:Male                                -1.3125 -15.76 13.137 1.0
00
## Non-Hispanic White:Female-Non-Hispanic Asian:Male                             1.3270 -10.40 13.057 1.0
00
## Non-Hispanic Black:Female-Non-Hispanic Asian:Male                             2.6808 -10.18 15.545 1.0
00
## Non-Hispanic Asian:Female-Non-Hispanic Asian:Male                            -1.6233 -16.33 13.085 1.0
00
## Other Race - Including Multi-Racial:Female-Non-Hispanic Asian:Male           -7.2569 -30.24 15.731 0.9
97
## Mexican American:Female-Other Race - Including Multi-Racial:Male             -5.8103 -23.87 12.245 0.9
96
## Other Hispanic:Female-Other Race - Including Multi-Racial:Male               -3.7857 -22.71 15.134 1.0
00
## Non-Hispanic White:Female-Other Race - Including Multi-Racial:Male           -1.1462 -18.08 15.787 1.0
```

```
00
## Non-Hispanic Black:Female-Other Race - Including Multi-Racial:Male                    0.2076 -17.53 17.946 1.0
00
## Non-Hispanic Asian:Female-Other Race - Including Multi-Racial:Male                   -4.0965 -23.21 15.021 1.0
00
## Other Race - Including Multi-Racial:Female-Other Race - Including Multi-Racial:Male  -9.7302 -35.76 16.300 0.9
87
## Other Hispanic:Female-Mexican American:Female                                         2.0246 -10.37 14.420 1.0
00
## Non-Hispanic White:Female-Mexican American:Female                                     4.6641  -4.42 13.743 0.8
75
## Non-Hispanic Black:Female-Mexican American:Female                                     6.0179  -4.49 16.522 0.7
73
## Non-Hispanic Asian:Female-Mexican American:Female                                     1.7138 -10.98 14.409 1.0
00
## Other Race - Including Multi-Racial:Female-Mexican American:Female                   -3.9199 -25.68 17.836 1.0
00
## Non-Hispanic White:Female-Other Hispanic:Female                                       2.6395  -8.06 13.334 1.0
00
## Non-Hispanic Black:Female-Other Hispanic:Female                                       3.9933  -7.94 15.922 0.9
95
## Non-Hispanic Asian:Female-Other Hispanic:Female                                      -0.3108 -14.21 13.586 1.0
00
## Other Race - Including Multi-Racial:Female-Other Hispanic:Female                     -5.9444 -28.42 16.533 0.9
99
## Non-Hispanic Black:Female-Non-Hispanic White:Female                                   1.3538  -7.08  9.784 1.0
00
## Non-Hispanic Asian:Female-Non-Hispanic White:Female                                  -2.9503 -13.99  8.091 0.9
99
## Other Race - Including Multi-Racial:Female-Non-Hispanic White:Female                 -8.5840 -29.42 12.249 0.9
72
## Non-Hispanic Asian:Female-Non-Hispanic Black:Female                                  -4.3041 -16.54  7.936 0.9
92
## Other Race - Including Multi-Racial:Female-Non-Hispanic Black:Female                 -9.9378 -31.43 11.555 0.9
36
## Other Race - Including Multi-Racial:Female-Non-Hispanic Asian:Female                 -5.6336 -28.28 17.011 1.0
00
```

Check assumptions

```
# make a data frame
bp.raceth.sex <- data.frame(bp.by.raceth.sex$residuals)

# plot the residuals
bp.raceth.sex %>%
ggplot(aes(x = bp.by.raceth.sex.residuals)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
  labs(x = "Residuals",
       y = "Number of observations")
```




```
# Levene test
library(package = "car")
leveneTest(y = syst.bp ~ race.eth*sex,
               data = nhanes.2014.clean)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  11    1.82  0.046 *
##       742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(3) Predicting systolic blood pressure by age, sex, and race-ethnicity

The data management, exploratory data analysis, and statistical modeling in this section is used to answer the research question:

Do age, sex, and race-ethnicity help to predict systolic blood pressure?

Codebook

- BPXSY1: Systolic blood pressure in mmHg
- RIDAGEYR: Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
- RIAGENDR: Gender of the participant
  - 1 = Male, 2 = Female
- RIDRETH3: Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category
  - 1 = Mexican American, 2 = Other Hispanic, 3 = Non-Hispanic White, 4 = Non-Hispanic Black, 6 = Non-Hispanic Asian, 7 = Other Race - Including Multi-Racial

Data cleaning

```
# make a smaller data frame
nhanes.2014.clean <- nhanes.2014 %>%
  select(RIAGENDR, RIDRETH3, BPXSY1, RIDAGEYR)

# check out the data
summary(object = nhanes.2014.clean)
```

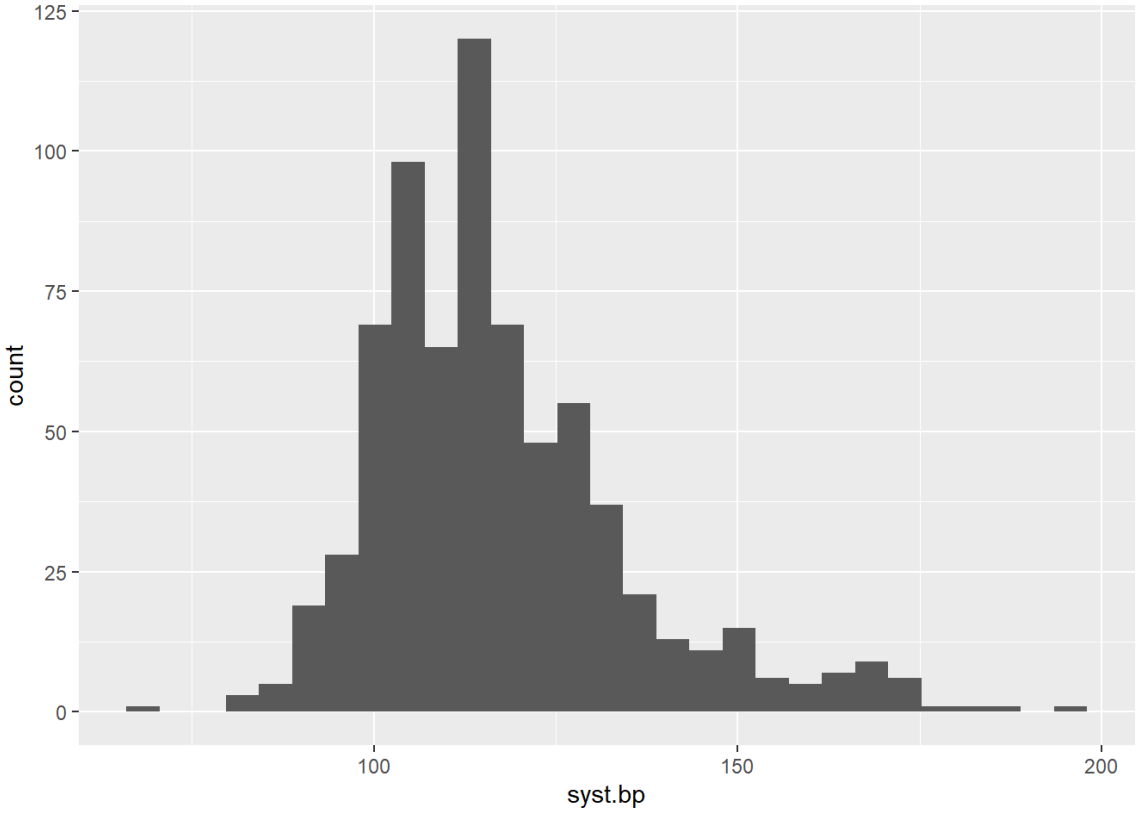
```
##     RIAGENDR       RIDRETH3        BPXSY1        RIDAGEYR
##  Min.   :1.00   Min.   :1.00   Min.   : 66    Min.   : 0.0
##  1st Qu.:1.00   1st Qu.:2.00   1st Qu.:106    1st Qu.:10.0
##  Median :2.00   Median :3.00   Median :116    Median :27.0
##  Mean   :1.51   Mean   :3.28   Mean   :118    Mean   :31.6
##  3rd Qu.:2.00   3rd Qu.:4.00   3rd Qu.:128    3rd Qu.:52.0
##  Max.   :2.00   Max.   :7.00   Max.   :228    Max.   :80.0
##                                NA's   :2641
```

```
# add labels to factors
# make better variable names
nhanes.2014.clean <- nhanes.2014 %>%
  select(RIAGENDR, RIDRETH3, BPXSY1, RIDAGEYR) %>%
  mutate(race.eth = recode_factor(RIDRETH3, `1` = "Mexican American",
                           `2` = "Other Hispanic",
                           `3` = "Non-Hispanic White",
                           `4` = "Non-Hispanic Black",
                           `6` = "Non-Hispanic Asian",
                           `7` = "Other Race - Including Multi-Racial")) %>%
  mutate(sex = recode_factor(RIAGENDR, `1` = "Male", `2` = "Female")) %>%
  rename(syst.bp = BPXSY1) %>%
  rename(age = RIDAGEYR) %>%
  sample_n(1000) %>%
  select(syst.bp, race.eth, sex, age)

# change reference group to Non-Hispanic Black and Female
nhanes.2014.clean$race.eth <- relevel(nhanes.2014.clean$race.eth,
                                      ref = "Non-Hispanic White")
nhanes.2014.clean$sex <- relevel(nhanes.2014.clean$sex,
                                      ref = "Female")

# check out the data
summary(object = nhanes.2014.clean)
```

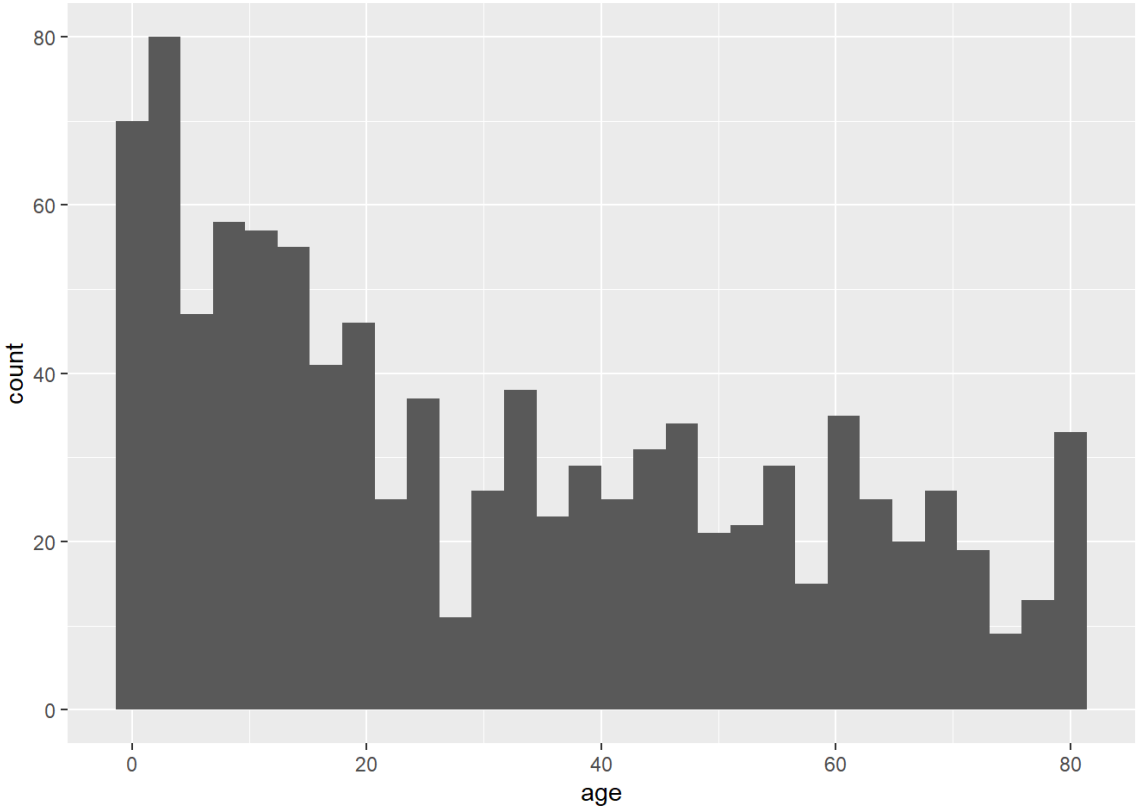
```
##     syst.bp                                  race.eth       sex           age
##  Min.   : 66   Non-Hispanic White                :328   Female:491   Min.   : 0.0
##  1st Qu.:106   Mexican American                  :172   Male  :509   1st Qu.: 9.0
##  Median :114   Other Hispanic                   : 97                Median :25.0
##  Mean   :118   Non-Hispanic Black               :232                Mean   :30.8
##  3rd Qu.:126   Non-Hispanic Asian               :108                3rd Qu.:51.0
##  Max.   :198   Other Race - Including Multi-Racial: 63               Max.   :80.0
##  NA's   :286
```

Exploratory data analysis

```
# distribution of blood pressure
nhanes.2014.clean %>%
  ggplot(aes(x = syst.bp)) +
  geom_histogram()
```




```
# distribution of age
nhanes.2014.clean %>%
  ggplot(aes(x = age)) +
  geom_histogram()
```



```
# table of descriptives
desc.table <- CreateTableOne(data = nhanes.2014.clean)
print(desc.table, nonnormal = 'age')
```

```
##
##                                        Overall
##   n                                      1000
##   syst.bp (mean (SD))                  117.46 (17.79)
##   race.eth (%)
##      Non-Hispanic White                   328 (32.8)
##      Mexican American                     172 (17.2)
##      Other Hispanic                        97 ( 9.7)
##      Non-Hispanic Black                   232 (23.2)
##      Non-Hispanic Asian                   108 (10.8)
##      Other Race - Including Multi-Racial   63 ( 6.3)
##   sex = Male (%)                          509 (50.9)
##   age (median [IQR])                    25.00 [9.00, 51.00]
```

Estimate the model

```
# systolic blood pressure by age, race-eth, sex
bp.age.race.sex <- lm(formula = syst.bp ~ age +
                        sex + race.eth,
                        data = nhanes.2014.clean,
                        na.action = na.exclude)
summary(object = bp.age.race.sex)
```

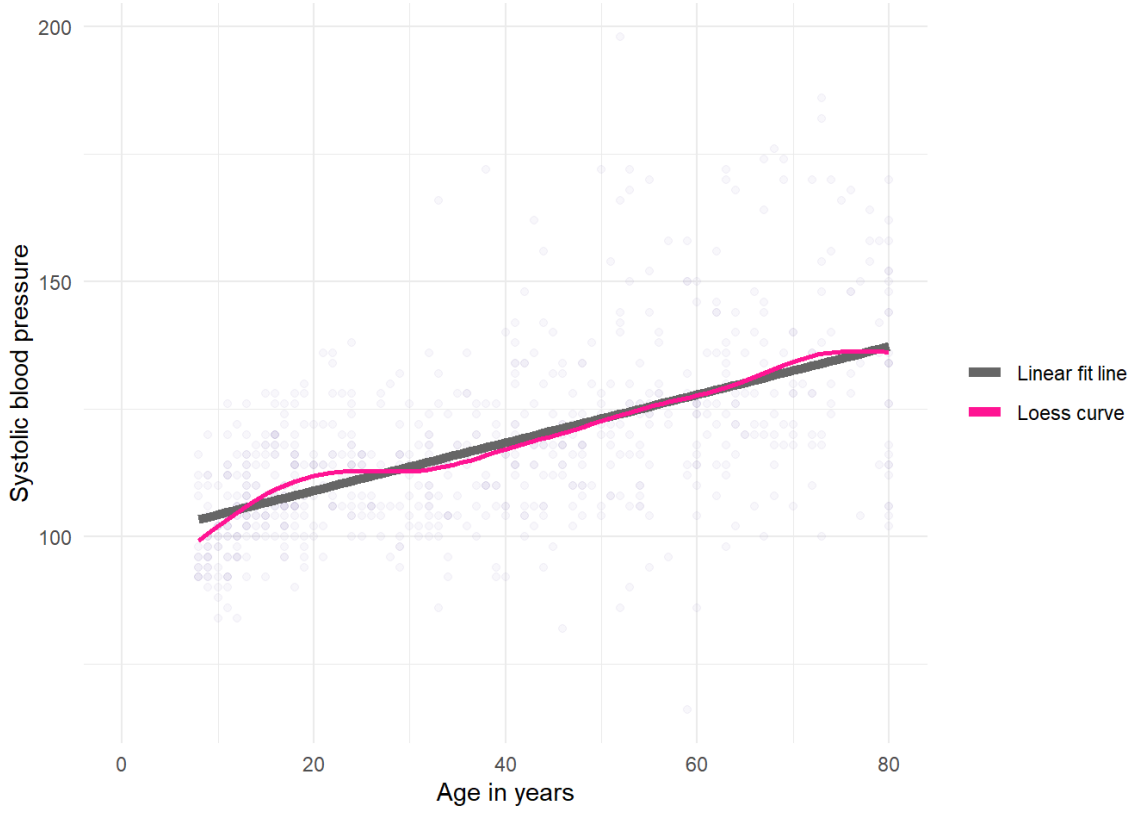
```
##
## Call:
## lm(formula = syst.bp ~ age + sex + race.eth, data = nhanes.2014.clean,
##     na.action = na.exclude)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.39  -8.30  -1.51   7.17  70.59
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               95.1721     1.5264   62.35  < 2e-16 ***
## age                                        0.4820     0.0256   18.84  < 2e-16 ***
## sexMale                                    3.7811     1.0850    3.48  0.00052 ***
## race.ethMexican American                   2.2879     1.6639    1.37  0.16957
## race.ethOther Hispanic                     3.3924     1.9435    1.75  0.08134 .
## race.ethNon-Hispanic Black                 5.1941     1.4593    3.56  0.00040 ***
## race.ethNon-Hispanic Asian                 1.1308     1.8317    0.62  0.53721
## race.ethOther Race - Including Multi-Racial   1.9054     2.6809    0.71  0.47749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.4 on 706 degrees of freedom
##   (286 observations deleted due to missingness)
## Multiple R-squared:  0.351,  Adjusted R-squared:  0.345
## F-statistic: 54.6 on 7 and 706 DF,  p-value: <2e-16
```

```
# confint for new model
confint(object = bp.age.race.sex)
```

```
##                                              2.5 % 97.5 %
## (Intercept)                                 92.175 98.169
## age                                          0.432  0.532
## sexMale                                      1.651  5.911
## race.ethMexican American                    -0.979  5.555
## race.ethOther Hispanic                      -0.423  7.208
## race.ethNon-Hispanic Black                   2.329  8.059
## race.ethNon-Hispanic Asian                  -2.465  4.727
## race.ethOther Race - Including Multi-Racial -3.358  7.169
```

Assumptions & Diagnostics

```
# systolic blood pressure and age
nhanes.2014.clean %>%
  ggplot(aes(x = age, y = syst.bp)) +
  geom_point(color = "#7463AC", alpha = .05) +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE, size = 2) +
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  theme_minimal() +
  labs(y = "Systolic blood pressure", x = "Age in years") +
  scale_color_manual(values = c("gray40", "deeppink"), name = "")
```



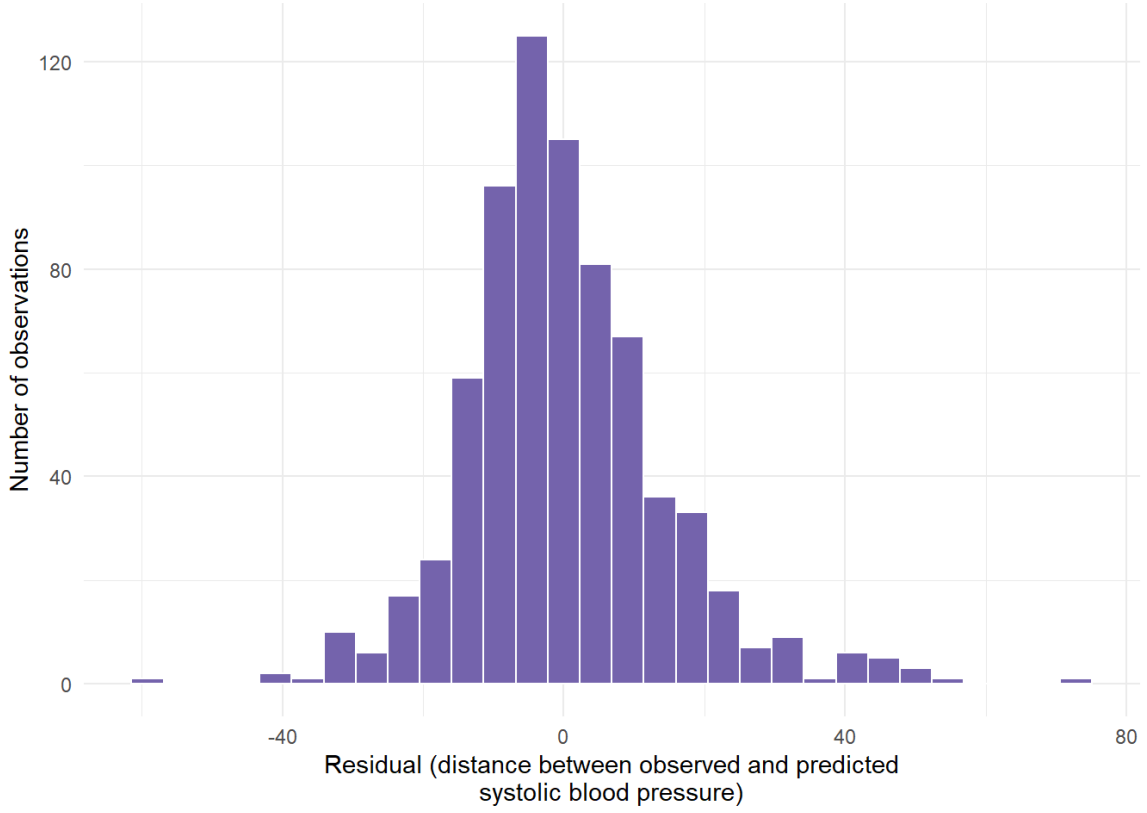

```
# Breusch-Pagan test
lmtest::bptest(formula = bp.age.race.sex)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bp.age.race.sex
## BP = 61, df = 7, p-value = 8e-11
```

```
# Durbin-Watson test
lmtest::dwtest(formula = bp.age.race.sex)
```

```
##
##  Durbin-Watson test
##
## data:  bp.age.race.sex
## DW = 2, p-value = 0.2
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# check plot of age and syst.bp
data.frame(bp.age.race.sex$residuals) %>%
  ggplot(aes(x = bp.age.race.sex.residuals)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
  labs(x = "Residual (distance between observed and predicted\nsystolic blood pressure)",
       y = "Number of observations")
```




```
# change the cutoff for leverage to reflect the 8 parameters in the model
nhanes.cleaned.diag.bp <- nhanes.2014.clean %>%
# drop_na() %>%
mutate(standardres = rstandard(model = bp.age.race.sex)) %>%
mutate(cooks.dist = cooks.distance(model = bp.age.race.sex)) %>%
mutate(lever = hatvalues(model = bp.age.race.sex)) %>%
mutate(outlier.infl = as.numeric(x = lever > 2*8/n()) +
as.numeric(x = cooks.dist > 4/n()) +
as.numeric(x = abs(x = standardres) > 1.96))

# examine the outliers & influential
nhanes.cleaned.diag.bp %>%
  select(syst.bp, age, race.eth, sex, outlier.infl) %>%
  filter(outlier.infl >= 2)
```

```
##    syst.bp age                          race.eth    sex outlier.infl
## 1      162  80                  Non-Hispanic White Female           2
## 2       86  52                  Non-Hispanic Black Female           2
## 3      162  43                  Non-Hispanic Black Female           2
## 4       86  60                    Mexican American Female           2
## 5      158  78                      Other Hispanic Female           2
## 6       86  33                      Other Hispanic   Male           2
## 7       92  40                  Non-Hispanic Black   Male           2
## 8      170  74                  Non-Hispanic Black Female           2
## 9      198  52                      Other Hispanic   Male           3
## 10     106  53                      Other Hispanic   Male           2
## 11     106  70                    Mexican American   Male           2
## 12     106  80                      Other Hispanic Female           3
## 13     130  19 Other Race - Including Multi-Racial   Male           2
## 14     104  77                  Non-Hispanic White   Male           2
## 15     174  69                  Non-Hispanic White Female           2
## 16     166  75                  Non-Hispanic Black Female           2
## 17     166  33                      Other Hispanic Female           2
## 18     170  72                  Non-Hispanic White Female           2
## 19      66  59                  Non-Hispanic White   Male           2
## 20     104  80                  Non-Hispanic White   Male           2
## 21     128  23 Other Race - Including Multi-Racial   Male           2
## 22     164  78                  Non-Hispanic White Female           2
## 23     168  76                      Other Hispanic Female           3
## 24     102  80                  Non-Hispanic White Female           2
## 25     170  63                  Non-Hispanic White Female           2
## 26     158  57                  Non-Hispanic Black Female           2
## 27     166  52                  Non-Hispanic White Female           2
## 28     186  73                  Non-Hispanic Asian   Male           3
## 29     172  50                  Non-Hispanic Black Female           2
## 30     148  42                      Other Hispanic Female           2
## 31     156  62                    Mexican American Female           2
## 32     176  68                    Mexican American   Male           2
## 33     174  67                  Non-Hispanic Asian   Male           2
## 34     172  53                  Non-Hispanic Black   Male           2
## 35     170  80                  Non-Hispanic White   Male           2
## 36     170  69                  Non-Hispanic Black   Male           2
## 37      82  46                  Non-Hispanic White Female           2
## 38      94  55                      Other Hispanic Female           3
## 39     156  44                  Non-Hispanic Black   Male           2
## 40      90  18 Other Race - Including Multi-Racial   Male           2
## 41     172  38                  Non-Hispanic White   Male           2
## 42     100  73                  Non-Hispanic White Female           2
## 43     168  53                  Non-Hispanic Black   Male           2
## 44     114  79                      Other Hispanic Female           2
## 45     168  64                  Non-Hispanic Black   Male           2
## 46     182  73                      Other Hispanic Female           3
## 47     172  63                    Mexican American Female           2
## 48     170  55                    Mexican American   Male           2
```