

An Analysis of NHANES Data

Sara Colando and Ian Horsburgh

22 February 2022

Introduction

The National Health and Nutrition Exam Survey (NHANES) is administered by the National Center for Health Statistics, which is a branch of the Center for Disease Control in the United States. The target population of NHANES is people who live in the United States (of any age). Each year, NHANES randomly selects 7,000 residents in the US to participate in NHANES. The participation of those selected is confidential and voluntary. Once selected residents opt-in to NHANES, they undergo a personal interview and “standardized health examination”, where researchers can collect health information.

The NHANES data that we are analyzing has 10,000 observations and 75 variables which were collected from 2009-2012. The observational unit is a resident of the United States that is six months or older. For our project, we will only be using 10 variables – 7 quantitative and 4 categorical – that are listed below.

Quantitative

1. Pulse (*60 second pulse rate*)
2. Testosterone (*recorded in ng/dL for patients 6 and older; no data for 2009-2010 was recorded*)
3. Poverty (*A ratio of family income to poverty guidelines. Smaller numbers indicate more poverty*)
4. Age (*all subjective 80 or older are recorded as 80*)
5. Physactivedays (*Number of days in a typical week that participant does moderate or vigorous-intensity activity. Reported in patients 12 years or older*)
6. AlcoholDay (*Average number of drinks consumed on days that participant drank alcoholic beverages. Reported for participants aged 18 years or older*)
7. HHIncomeMid (*Variable was partitioned into blocks with the smallest one being (0,4999) and the largest block being 100,000 or more, the median of each block was then used to estimate income*)

Categorical

8. Gender (*sex of study participant coded as “male” or “female”*)
9. Race1 (*Reported race of study participant: Mexican, Hispanic, White, Black, or Other*)
10. Education (*Reported for ages 20 or older. Categories to choose from are “8thgrade”, “9-11thgrade”, “Highschool”, “SomeCollege”, or “CollegeGrad”*)
11. Depressed (*Self reported number of days where participant felt down, depressed or hopeless. Reported in patients 18 and older with categories of “none”, “several”, “majority (more than half the days)”, or “almostall”*)

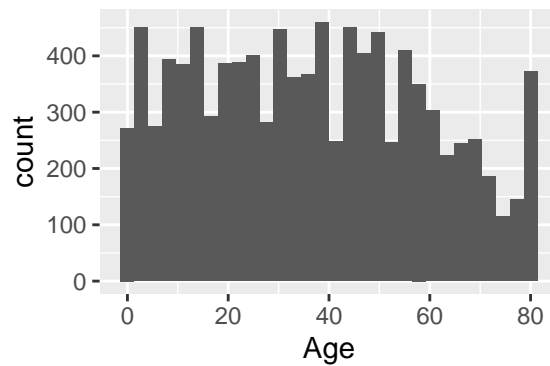
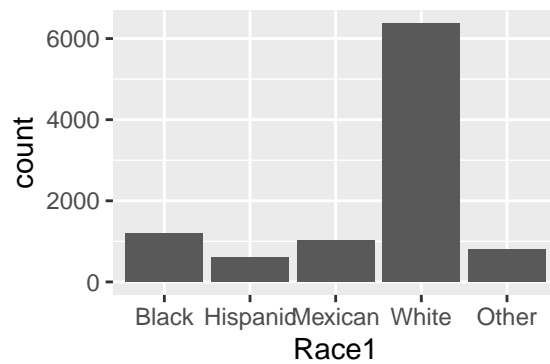
Our response variable will be pulse – which is the 60 second pulse rate of a participant. We will run pulse against a variety of explanatory variables to assess linear fit.

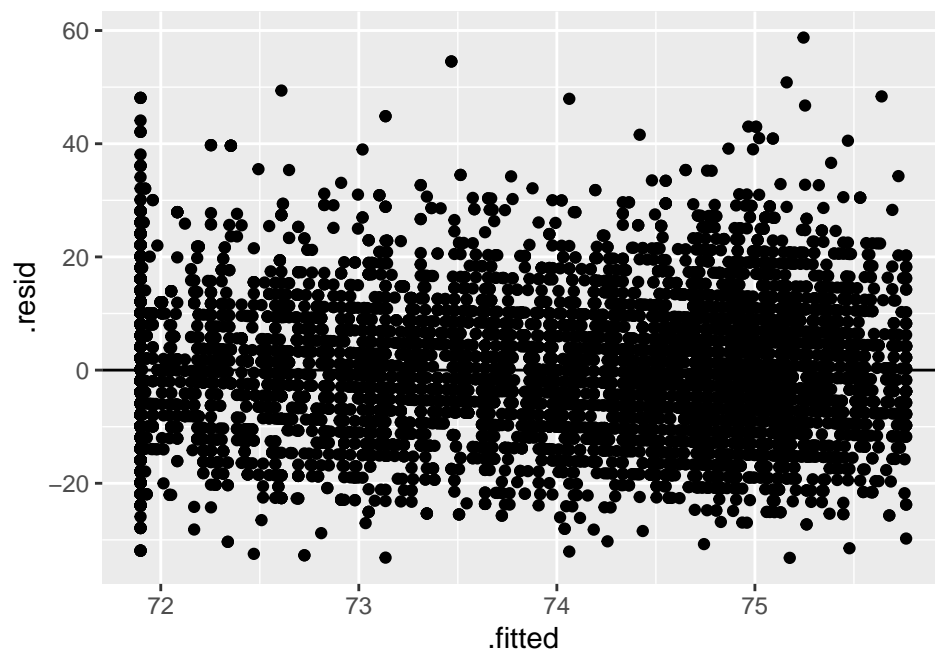
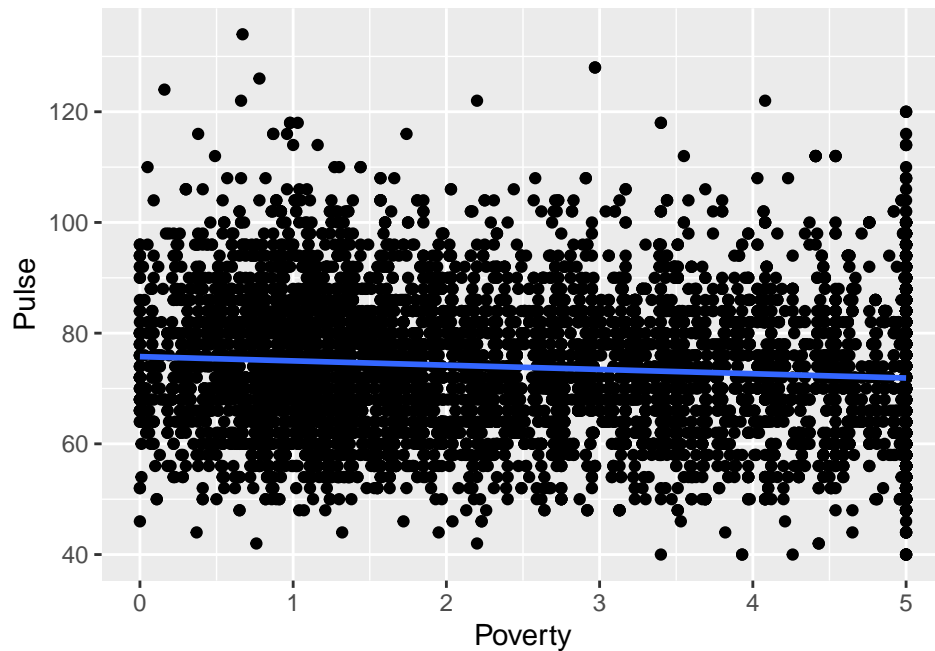
Given many of our variables (such as depressed, alcohol, and education) were not collected for adolescents, we will be limiting the scope of our population to adults (>18 years old) who live in the US.

Summary Statistics

```
## # A tibble: 11 x 5
##   skim_variable  numeric.mean numeric.sd complete_rate numeric.p50
##   <chr>          <dbl>      <dbl>         <dbl>      <dbl>
## 1 Gender          NA         NA             1         NA
## 2 Depressed        NA         NA          0.667         NA
## 3 Education        NA         NA          0.722         NA
## 4 Race1           NA         NA             1         NA
## 5 Age             36.7       22.4             1         36
## 6 HHIncomeMid     57206.    33020.          0.919    50000
## 7 Testosterone    198.       227.           0.413     43.8
## 8 PhysActiveDays   3.74       1.84           0.466         3
## 9 Pulse           73.6      12.2           0.856         72
##10 Poverty          2.80       1.68           0.927         2.7
##11 AlcoholDay       2.91       3.18           0.491         2
```

Graphs





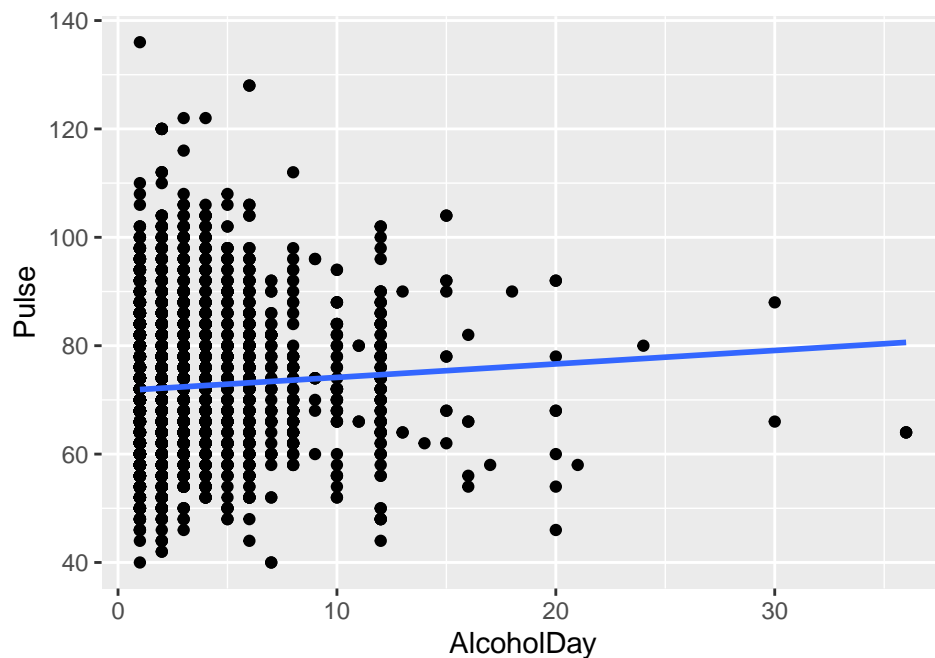
```
OurData %>%
  lm(Pulse ~ as.factor(HHIncomeMid), data = .) %>%
  tidy()
```

```
## # A tibble: 12 x 5
##   term                                estimate std.error statistic p.value
##   <chr>                                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)                        73.7       1.02     72.6     0
## 2 as.factor(HHIncomeMid)7500         1.78       1.33      1.34    0.180
## 3 as.factor(HHIncomeMid)12500        1.33       1.16      1.14    0.254
## 4 as.factor(HHIncomeMid)17500       -0.0500     1.17    -0.0428  0.966
```

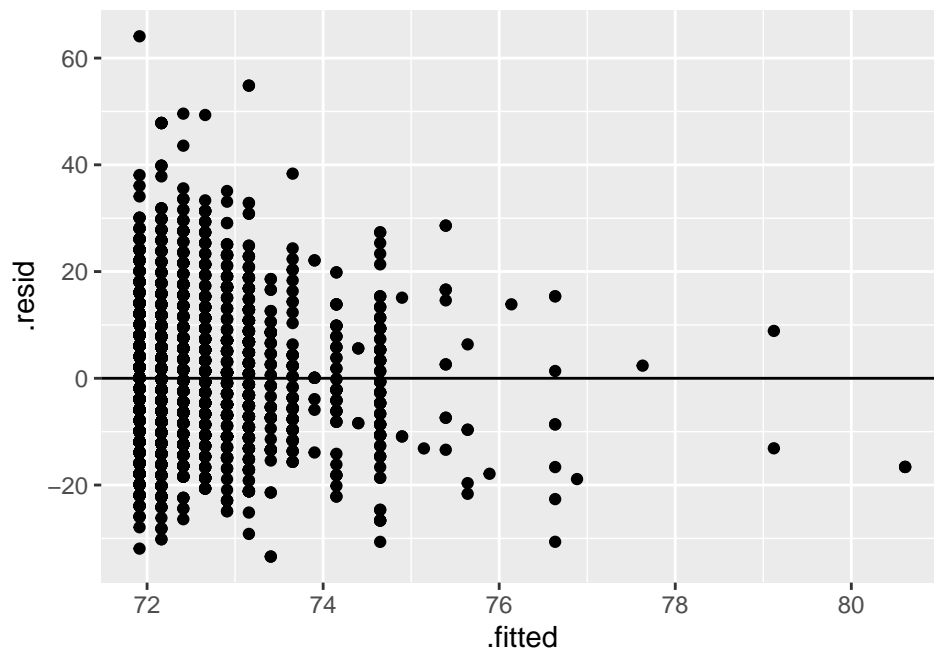
```
## 5 as.factor(HHIncomeMid)22500 2.31 1.15 2.01 0.0441
## 6 as.factor(HHIncomeMid)30000 0.652 1.10 0.592 0.554
## 7 as.factor(HHIncomeMid)40000 0.591 1.11 0.534 0.594
## 8 as.factor(HHIncomeMid)50000 -1.06 1.12 -0.949 0.343
## 9 as.factor(HHIncomeMid)60000 -1.45 1.14 -1.28 0.202
## 10 as.factor(HHIncomeMid)70000 0.106 1.16 0.0919 0.927
## 11 as.factor(HHIncomeMid)87500 0.0613 1.09 0.0563 0.955
## 12 as.factor(HHIncomeMid)100000 -1.56 1.05 -1.49 0.137
```

```
Alcohol_new <- NHANES %>%
  filter(AlcoholDay <= 40)
```

```
Alcohol_new %>%
  lm(Pulse ~ AlcoholDay, data = .) %>%
  ggplot(aes(x=AlcoholDay, y=Pulse)) +
  geom_point() +
  geom_smooth(method= "lm", se= FALSE)
```



```
Alcohol_new %>%
  lm(Pulse ~ AlcoholDay, data = .) %>%
  augment() %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```

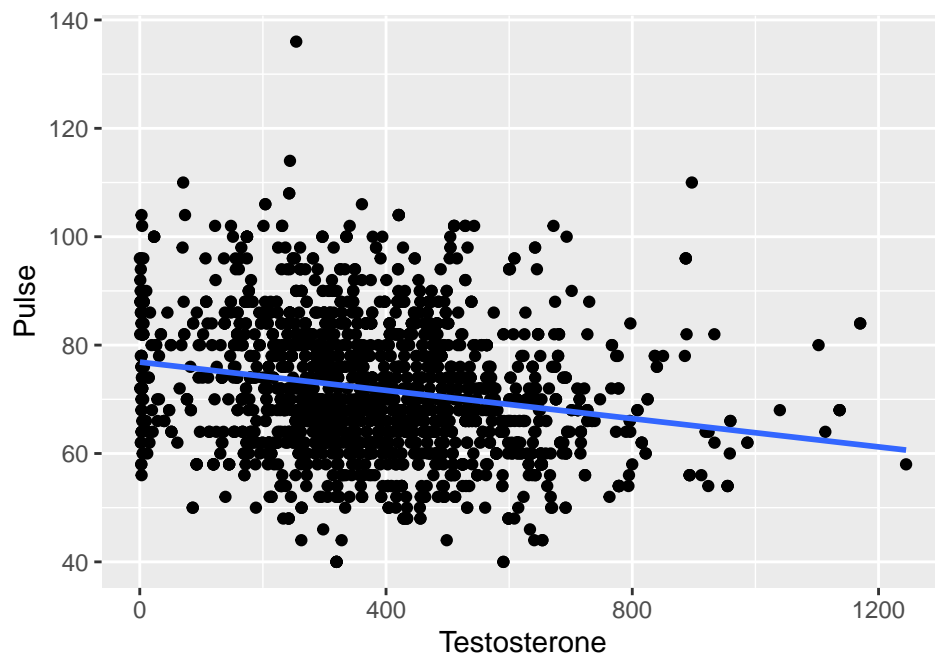


```
NHANES %>%
  lm(Pulse ~ AlcoholDay, data = .) %>%
  tidy()
```

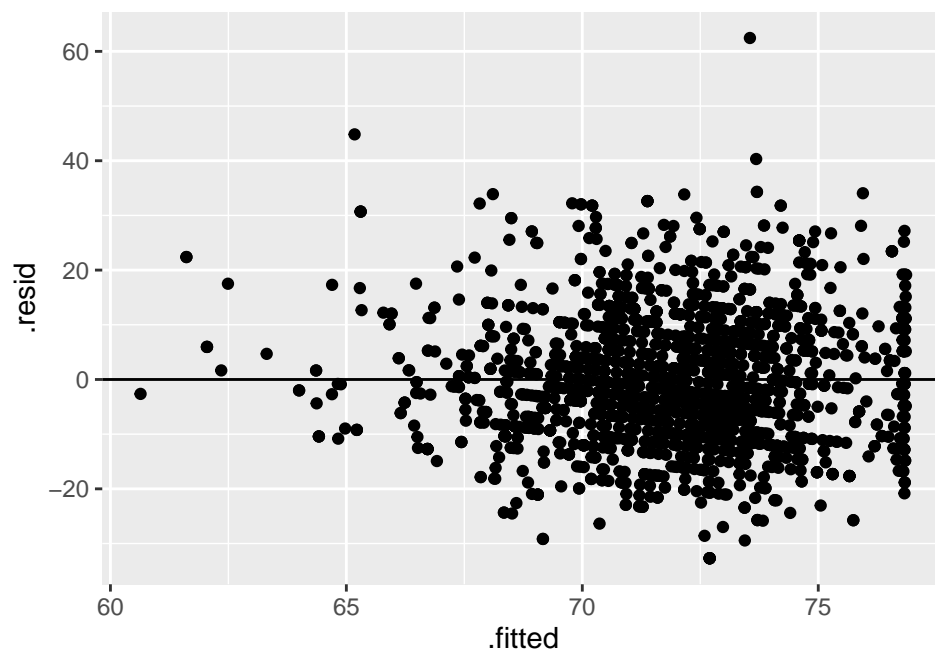
```
## # A tibble: 2 x 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)  71.8      0.227    317.      0
## 2 AlcoholDay   0.197     0.0526     3.75 0.000177
```

```
testosterone_new <- OurData %>%
  filter(Testosterone < 1400) %>%
  filter(Gender == "male")
```

```
testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  ggplot(aes(x=Testosterone, y=Pulse)) + geom_point() + geom_smooth(method= "lm", se= FALSE)
```



```
testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  augment() %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



```
testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  tidy(conf.int = TRUE, conf.level = 0.95)
```

```
## # A tibble: 2 x 7
##   term          estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)    76.9      0.586     131.    0        75.7     78.0
## 2 Testosterone  -0.0130    0.00137    -9.52 5.08e-21 -0.0157 -0.0103
```

```
testosterone_new %>%
  lm(Pulse ~ Testosterone, data = .) %>%
  glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC    BIC
##   <dbl>      <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   0.0438      0.0433  11.9     90.5 5.08e-21     1 -7694. 15394. 15411.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Oversampling and NHANES

NHANES “over-samples persons 60 and older, African Americans, Asians, and Hispanics” to create more reliable and representative samples of the population.