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 – x quantitative and y categorical – that are listed below.

LIST HERE

Our response variable is pulse – which is the 60 second pulse rate of a participant.

Given many of our variables (such as _____) were only collected for participants 18 and older, we will be limiting the scope of our population to adults (>18 years old) who live in the US.

Age (quantitative)

- Age in years at screening of study participant. All subjects 80 and older were recorded as 80.

Testosterone (quantitative)

- Testosterone total (ng/dL), recorded for patients 6 and older. Note that no testosterone data for 2009-2010 was recorded.

Physactivedays (numerical discrete)

- Number of days in a typical week that partipant does moderate or vigorous-intensity activity. Reported in patients 12 years or older.

Pulse (quantitative)

- 60 second pulse rate

Gender (categorical)

- Gender (sex) of study particpant coded as "male" or "female".

SleepTrouble (categorical)

- Paritipant has told a doctor or other health professional that they had trouble sleeping. Reported in patients 16 and older. Either "yes" or "no".

Depressed (categorical)

- Self reported number of days where participant felt down, depressed or hopeless. Reported in patients 18 and olders with categories of "none", "several", "majority (more than half the days)", or "almostall".

Education (categorical)

- Educational level of study participant. Reported for ages 20 or older. Categories to choose from are "8thgrade", "9-11thgrade", "Highschool", "SomeCollege", or "CollegeGrad".

Work (categorical)

- Categorizes whether study participant is "working", "not working" or no data was collected.

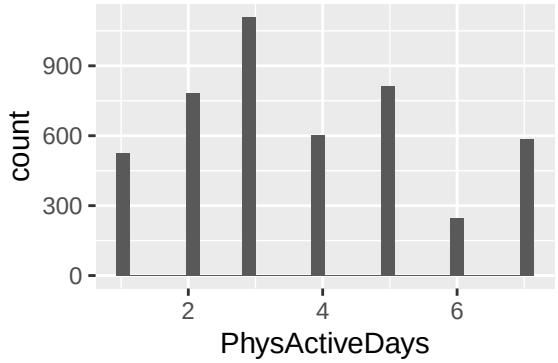
HHIncomeMid (numerical)

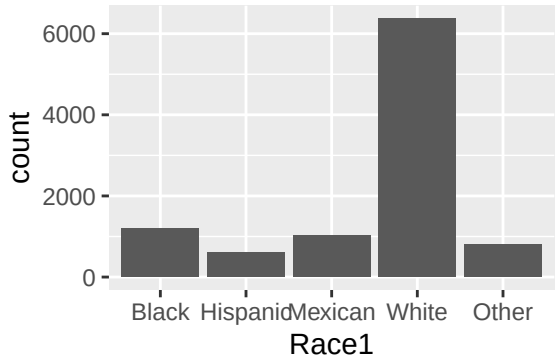
- Total annual gross income for the household in US dollars, derived from the median of each partition. Variable was partitioned into blocks with the smallest one being (0,4999) and the largest block being (100,000 or more).

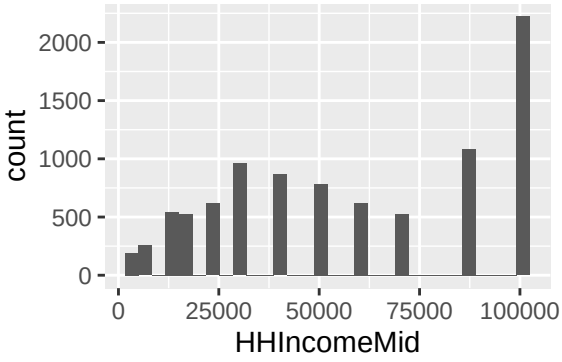
Summary Statistics

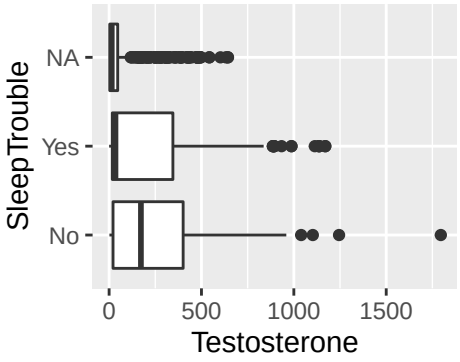
```
## # A tibble: 11 x 5
##    skim_variable  numeric.mean numeric.sd complete_rate numeric.p50
##    <chr>                 <dbl>      <dbl>         <dbl>       <dbl>
##  1 Gender                NA         NA            1            NA
##  2 SleepTrouble          NA         NA            0.777        NA
##  3 Depressed             NA         NA            0.667        NA
##  4 Education             NA         NA            0.722        NA
##  5 Race1                 NA         NA            1            NA
##  6 Age                   36.7       22.4          1            36
##  7 HHIncomeMid        57206.     33020.           0.919     50000
##  8 Testosterone         198.       227.           0.413        43.8
##  9 PhysActiveDays         3.74       1.84         0.466         3
## 10 Pulse                 73.6       12.2          0.856        72
## 11 Poverty                2.80       1.68         0.927         2.7
```

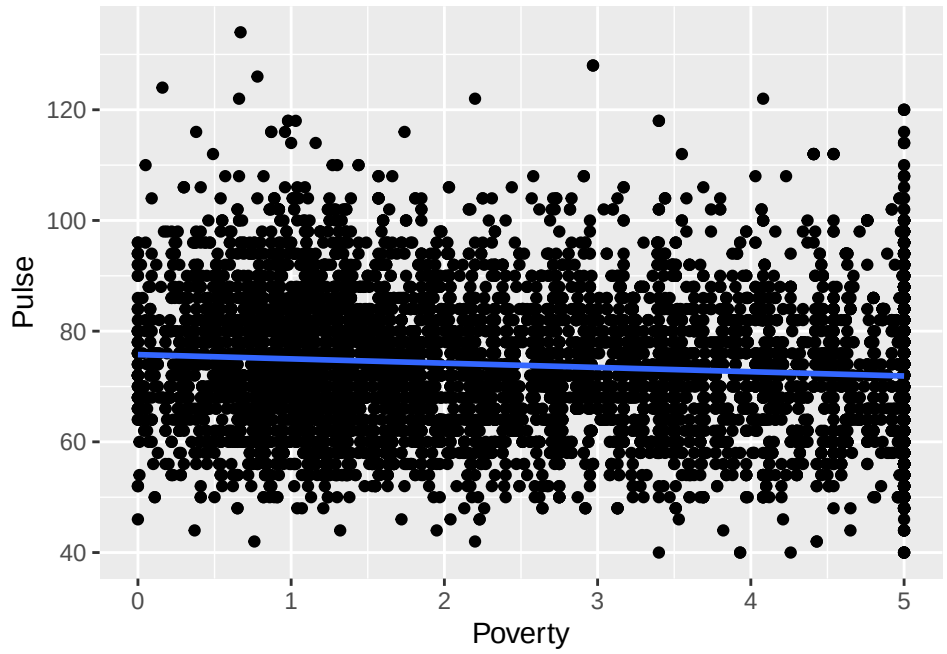
Graphs

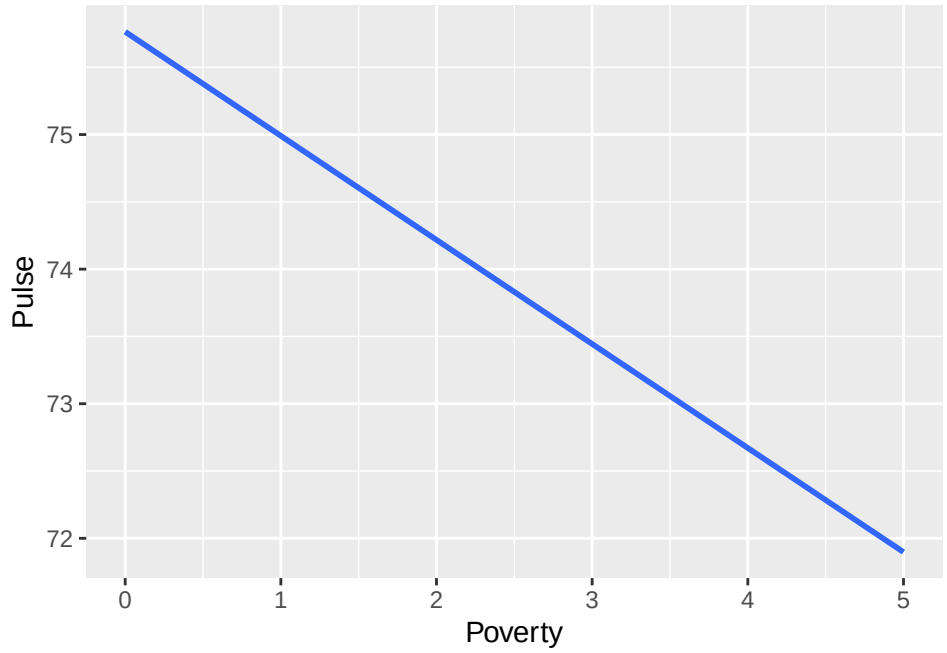

900
600
300
0
count
2
4
6
PhysActiveDays


6000
4000
2000
0
count
Black
Hispanic
Mexican
White
Other
Race1


2000
1500
1000
500
0
count
0
25000
50000
75000
100000
HHIncomeMid


NA
Yes
No
SleepTrouble
0
500
1000
1500
Testosterone



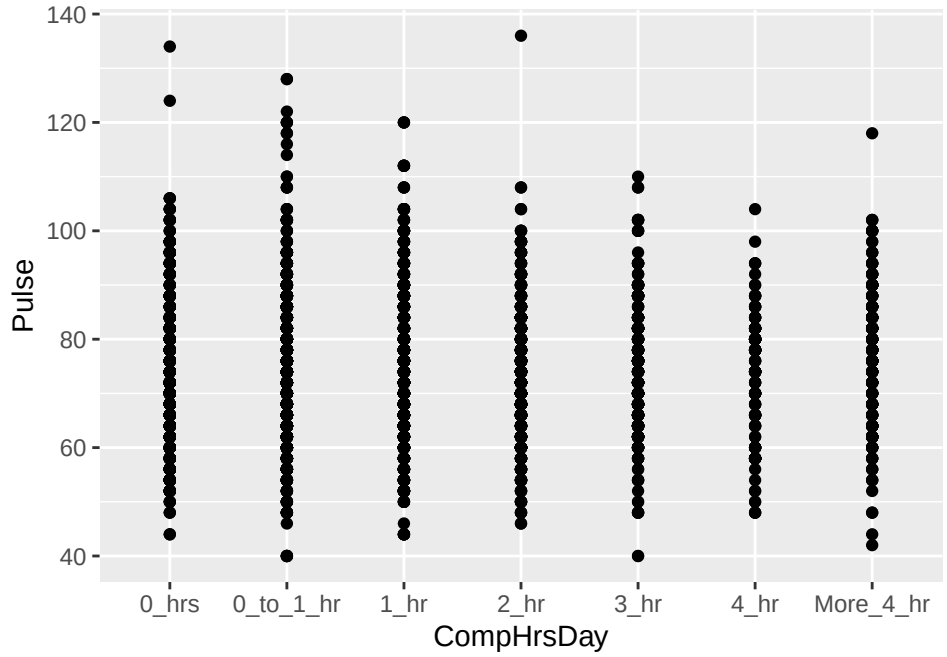



```
OurData %>%
 lm(Pulse ~ as.factor(HHIncomeMid), data = .) %>%
 tidy()
```

```
## # 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
```

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

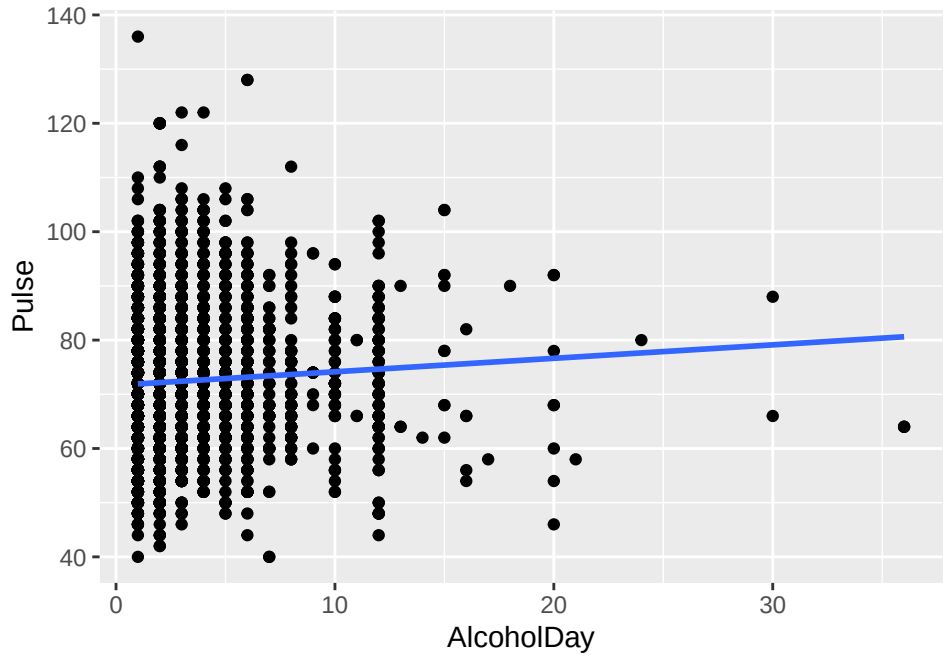



```
NHANES %>%
 lm(Pulse ~ CompHrsDay, data = .) %>%
  tidy()
```

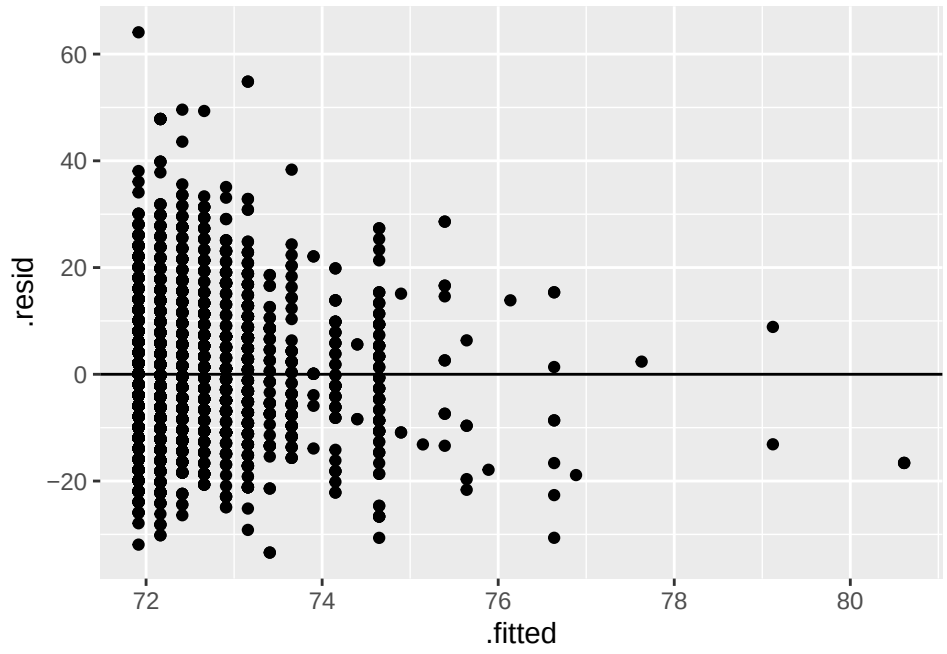
```
## # A tibble: 7 x 5
##   term                estimate std.error statistic   p.value
##   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           72.4       0.415   174.    0
## 2 CompHrsDay0_to_1_hr    1.36      0.540     2.51  0.0121
## 3 CompHrsDay1_hr         1.49      0.575     2.59  0.00952
## 4 CompHrsDay2_hr         0.981     0.663     1.48  0.139
## 5 CompHrsDay3_hr         1.93      0.786     2.45  0.0142
## 6 CompHrsDay4_hr         0.684     1.05      0.648 0.517
## 7 CompHrsDayMore_4_hr    3.79      0.889     4.26  0.0000205
```

```
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)
```

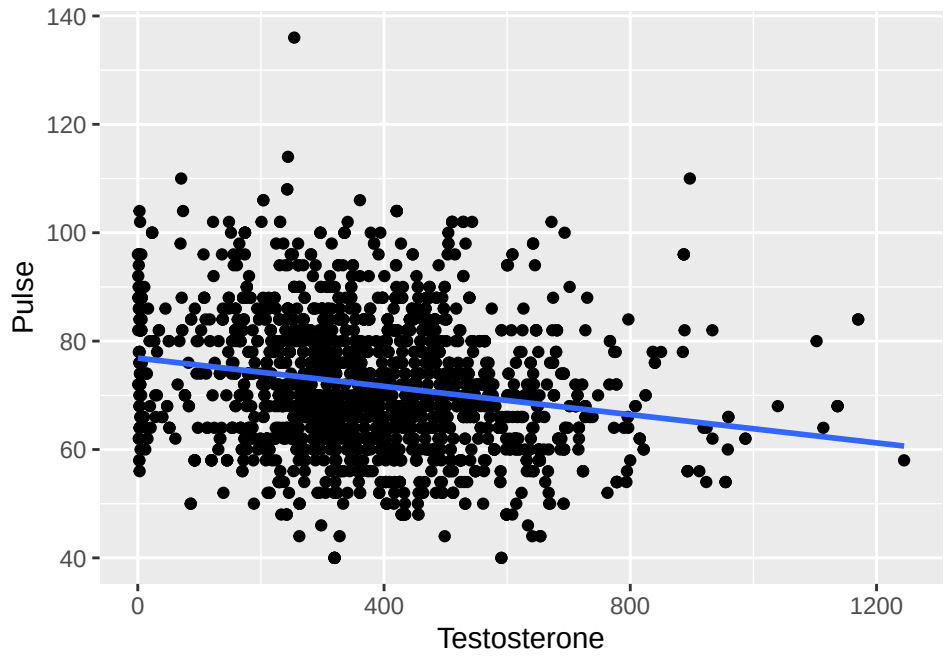



```
NHANES %>%
 lm(Pulse ~ AlcoholDay, data = .) %>%
  tidy()
```

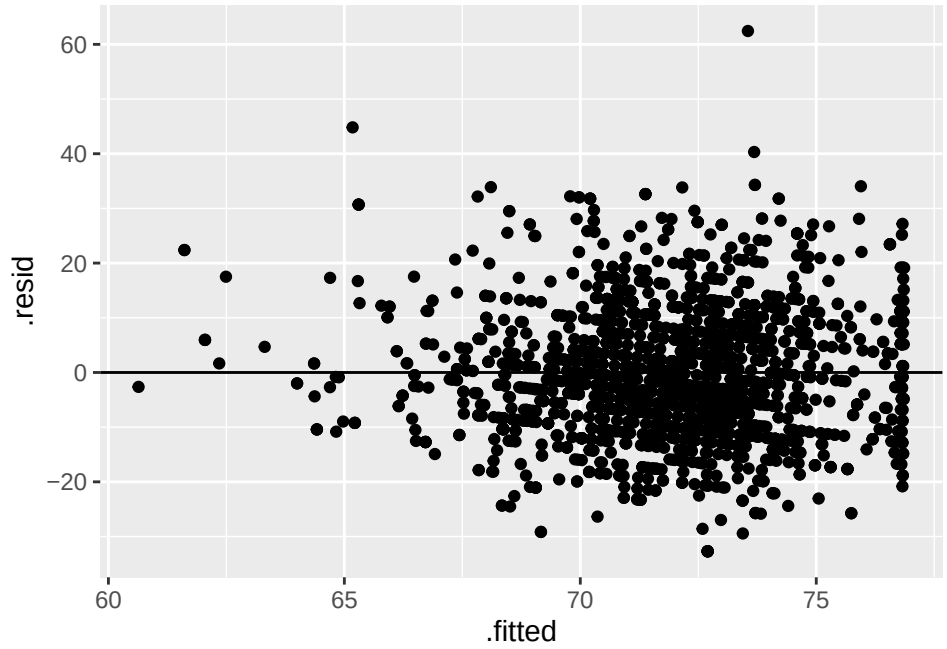
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   71.8      0.227     317.   0
## 2 AlcoholDay     0.197    0.0526      3.75 0.000177
```

```
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>
```