Analysis of NHANES Sleep Data

Adam Kapelner, Josh Magarick

(Special Thanks to Dylan Small)

Department of Statistics
The Wharton School, University of Pennsylvania

NHANES

Since the National Health Survey Act of 1956,

NHANES

Since the National Health Survey Act of 1956, the US thought it was important to get readings on the population's health status. In 1971, this was coined...

NHANES

Since the National Health Survey Act of 1956, the US thought it was important to get readings on the population's health status. In 1971, this was coined...

> *"The National Health and Nutrition Examination Survey (NHANES)... a program of studies designed to assess the health and nutritional status of adults and children in the United States."*

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

NHANES

Since the National Health Survey Act of 1956, the US thought it was important to get readings on the population's health status. In 1971, this was coined...

> *"The National Health and Nutrition Examination Survey (NHANES)... a program of studies designed to assess the health and nutritional status of adults and children in the United States."*

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Nationally representative sample of about 10,000 people every year by using a stratified sample of households from a sample of geographical units (deliberate oversampling of the elderly and minorities).

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided).

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

- in-depth physical examination

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

- in-depth physical examination (e.g. audiometry,

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

- in-depth physical examination (e.g. audiometry, all the physical measurements,

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

- in-depth physical examination (e.g. audiometry, all the physical measurements, X-ray absorption,

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

- in-depth physical examination (e.g. audiometry, all the physical measurements, X-ray absorption, retinal imaging battery,

The data (Part I)

They try to measure various indicators:

- demographic
- socioeconomic
- dietary
- health-related

The first three indicators are queried via surveys (both in-person interviews and computer-aided). The health related data collection is a combination of surveying and a sophisticated:

- in-depth physical examination (e.g. audiometry, all the physical measurements, X-ray absorption, retinal imaging battery, dental decay, etc.)

The data (Part II)

- in-depth laboratory tests

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood,

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood, pesticides in urine,

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood, pesticides in urine, testing for common STDs,

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood, pesticides in urine, testing for common STDs, thyroid profile,

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood, pesticides in urine, testing for common STDs, thyroid profile, glucose and insulin)

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood, pesticides in urine, testing for common STDs, thyroid profile, glucose and insulin)

How did they collect this data?

The data (Part II)

- in-depth laboratory tests (e.g. lead and heavy metals in blood, pesticides in urine, testing for common STDs, thyroid profile, glucose and insulin)

How did they collect this data? Via "mobile equipment centers":

Why is this useful?

Even via univariate analyses a la Stat 101, the NHANES data has been instrumental in:

Why is this useful?

Even via univariate analyses a la Stat 101, the NHANES data has been instrumental in:

- Constructing growth charts
- National mandate to add folate and Iron to cereal
- Measuring lead poisoning
- Nationwide efforts to reduce cholesterol

Why is this useful?

Even via univariate analyses a la Stat 101, the NHANES data has been instrumental in:

- Constructing growth charts
- National mandate to add folate and Iron to cereal
- Measuring lead poisoning
- Nationwide efforts to reduce cholesterol

And, using basic regression a la Stat 102, you can detect "associations" and actually publish papers.

Why is this useful?

Even via univariate analyses a la Stat 101, the NHANES data has been instrumental in:

- Constructing growth charts
- National mandate to add folate and Iron to cereal
- Measuring lead poisoning
- Nationwide efforts to reduce cholesterol

And, using basic regression a la Stat 102, you can detect "associations" and actually publish papers. And, using the methods of this class, hopefully we can find causal relationships...

What are we interested in?

What are we interested in?

What *causes* Sleep

Why?

What are we interested in?

What *causes* Sleep

Why? Because it's pretty important... and...

Previous Studies using NHANES and Sleep

We searched google scholar for "`sleep`" AND "`NHANES`" and of the first 100 results, we found 13 papers...

Previous Studies using NHANES and Sleep

We searched google scholar for "`sleep`" AND "`NHANES`" and of the first 100 results, we found 13 papers...

- 10 / 13 use sleep as covariate to predict something else

Previous Studies using NHANES and Sleep

We searched google scholar for "`sleep`" AND "`NHANES`" and of the first 100 results, we found 13 papers...

- 10 / 13 use sleep as covariate to predict something else
- NONE of the papers use any methods from this class

$\Rightarrow$

Previous Studies using NHANES and Sleep

We searched google scholar for "`sleep`" AND "`NHANES`" and of the first 100 results, we found 13 papers...

- 10 / 13 use sleep as covariate to predict something else
- NONE of the papers use any methods from this class

$\Rightarrow$ We are interested in proving certain lifestyles *cause* a good (or bad) night's sleep.

Our Dataset

We used the 2007/8 data $n = 9,762$

Our Dataset

We used the 2007/8 data $n = 9,762$ (chosen because of some covariates which later turned out to be not important, so the choice was arbitrary).

Our Dataset

We used the 2007/8 data $n = 9,762$ (chosen because of some covariates which later turned out to be not important, so the choice was arbitrary). We threw out nutritional data, so $p = 2,355$.

Our Dataset

We used the 2007/8 data $n = 9,762$ (chosen because of some covariates which later turned out to be not important, so the choice was arbitrary). We threw out nutritional data, so $p = 2,355$. Our Response variable was chosen to be "SLD010H".

> *How much sleep do you usually get at night on weekdays or workdays? ENTER HOURS.*

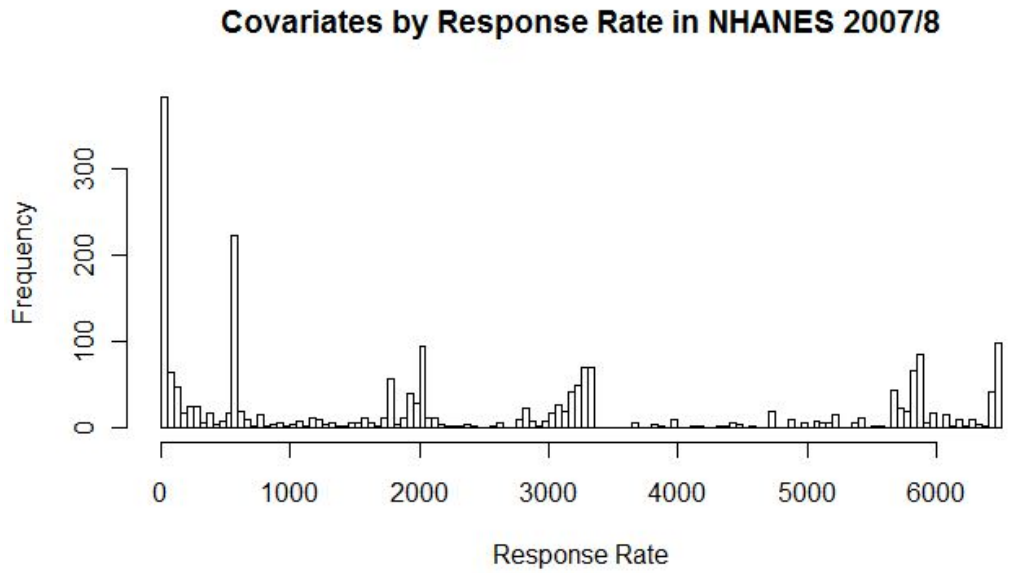
$n_0 = 6,498 \approx \frac{2}{3}n$ answered this question.

Our Dataset

We used the 2007/8 data $n = 9,762$ (chosen because of some covariates which later turned out to be not important, so the choice was arbitrary). We threw out nutritional data, so $p = 2,355$. Our Response variable was chosen to be "SLD010H".

> *How much sleep do you usually get at night on weekdays or workdays? ENTER HOURS.*

$n_0 = 6,498 \approx \frac{2}{3}n$ answered this question. What about other covariates?

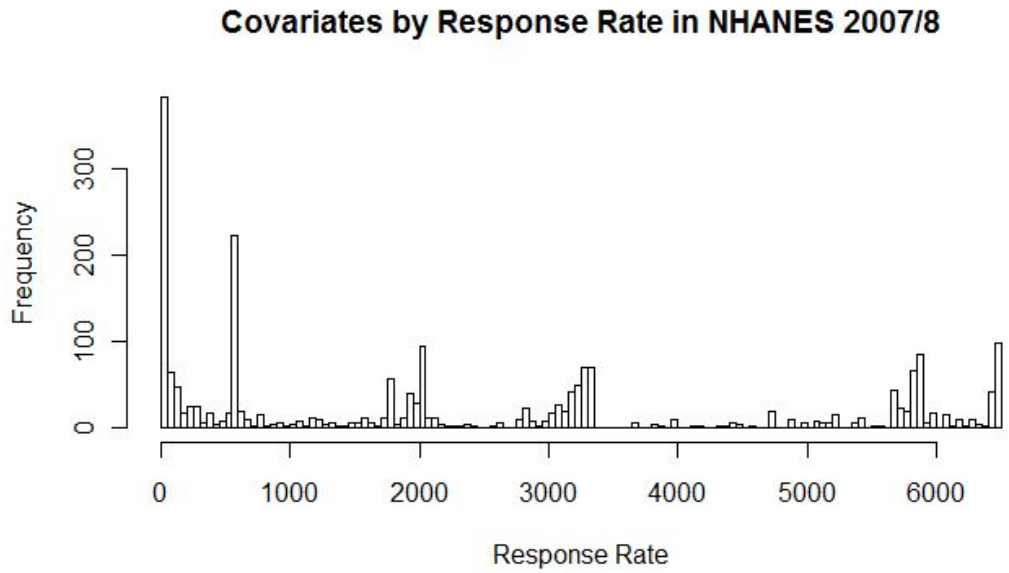


Our Dataset

We used the 2007/8 data $n = 9,762$ (chosen because of some covariates which later turned out to be not important, so the choice was arbitrary). We threw out nutritional data, so $p = 2,355$. Our Response variable was chosen to be "SLD010H".

> *How much sleep do you usually get at night on weekdays or workdays? ENTER HOURS.*

$n_0 = 6,498 \approx \frac{2}{3}n$ answered this question. What about other covariates?



Preprocessing

- Download and label all data
- Merge all data tables

Preprocessing

- Download and label all data
- Merge all data tables
- Pick covariates that are not missing for records that have sleep variables

Preprocessing

- Download and label all data
- Merge all data tables
- Pick covariates that are not missing for records that have sleep variables
- Pick a treatment variable "PAQ635"

Preprocessing

- Download and label all data
- Merge all data tables
- Pick covariates that are not missing for records that have sleep variables
- Pick a treatment variable "`PAQ635`" - Walk or bicycle

> *... I would like to ask you about the usual way you travel to and from places. For example to work, for shopping, to school. Do you walk or use a bicycle for at least 10 minutes continuously to get to and from places?*

Preprocessing

- Download and label all data
- Merge all data tables
- Pick covariates that are not missing for records that have sleep variables
- Pick a treatment variable "PAQ635" - Walk or bicycle

> *... I would like to ask you about the usual way you travel to and from places. For example to work, for shopping, to school. Do you walk or use a bicycle for at least 10 minutes continuously to get to and from places?*

- Set intersect all records available for all covariates

What are our control variables?

cov code	cov description	num_obs_non_null
RIAGENDR	Gender	6498
RIDAGEYR	Age at Screening Adjudicated - Recode	6498
DMDMARTL	Marital Status	5893
BMXWT	Weight (kg)	6157
BMXHT	Standing Height (cm)	6157
OHAEXSTS	Overall Oral Health Exam Status	6262
AUQ131	General condition of hearing	6498
BPQ020	Ever told you had high blood pressure	6498
CBQ020	Fruits available at home	6418
CBQ060	Soft drinks available at home	6418
DIQ010	Doctor told you have diabetes	6498
HIQ011	Covered by health insurance	6498
HUQ090	Seen mental health professional/past yr	6498
INQ020	Income from wages/salaries	6436
MCQ010	Ever been told you have asthma	6498
MCQ160M	Ever told you had a thyroid problem	5893
PFQ090	Require special healthcare equipment	5893
RDQ070	Wheezing or whistling in chest - past yr	6498
SMQ020	Smoked at least 100 cigarettes in life	5893
SMD410	Does anyone smoke inside home?	6446
VIQ031	General condition of eyesight	6490