Exploring Influences on Sleep Quality

Presenter: Paul Previde

Motivation

- The Center for Disease Control considers insufficient or poor-quality sleep a public health concern [1,2]
- An estimated 50-70 million Americans report trouble sleeping [1,2]
- The total costs associated with medical involvement and lost productivity due to sleep problems exceeds $100 billion [3]

[1] https://www.cdc.gov/features/dssleep (accessed May 6, 2017).
[2] *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. Institute of Medicine, National Academy of Sciences, Washington, D.C., 2006.
[3] E. Wickwire, F. Shaya, S. Scharf, “Health economics of insomnia treatments: The return on investment for a good night’s sleep”, *Sleep Medicine Reviews*, vol. 30, Dec. 2016, 72-82, 2016.

This Project

Sleep deficits can manifest as subjective (poor sleep quality) or quantitative (fewer hours)

Explore how various health factors influence sleep quantity and quality:

- demographic (age, gender)
- physiological (heart rate, BMI, blood pressure)
- behavioral (alcohol use, smoking)

NHANES Dataset

- Center for Disease Control and Prevention (“CDC”)
- National Health and Nutrition Examination Survey (“NHANES”)
- People nationwide are randomly chosen, and paid, to participate
- Early forms of this data started in 1959
- NHANES is cited in 1,470 Google Scholar references*
- Data source for this work: Kaggle NHANES 2013-2014 data set**

* As of May 6, 2017

** URL: https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey

Tables in the NHANES Dataset

data table	number of predictors*	number of patients
Questionnaire	952	10,175
Examinations	223	9,813
Demographics	46	10,175
Dietary	167	9,813
Laboratory	423	9,813
Medications	13	20,194
Total	1,824	10,175 distinct patients

* excludes unique patient identifier, present in each table

Missing Values

complete.cases(full_patient_df) → 0
total number of cells: 12,433,850
missing values: 8,593,662 → 69%
mean # of missing values per row → 844 (st.dev. 126) out of 1,824

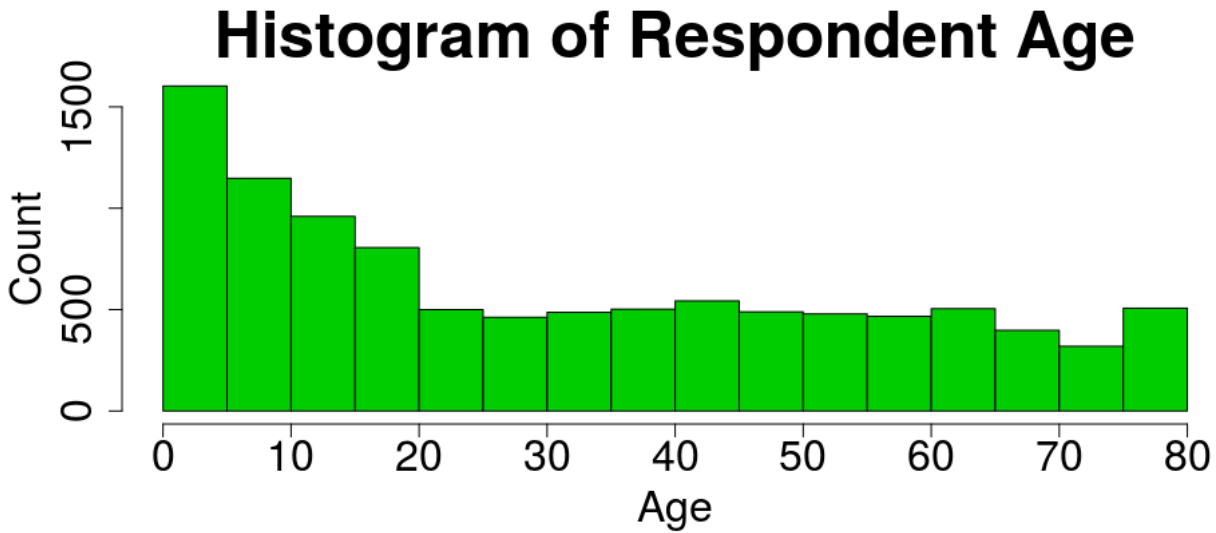
Strategies to address missing values:

- Use subsets of the available predictors
- Wait as long as possible to remove observations
- No imputation of missing values
- Use na.rm = True as an option to numerical functions

Subset of the Data Studied

Response variables		Predictor variables	
Continuous	Categorical	Continuous	Categorical
hours of sleep per night	sleep problem reported or diagnosed	Examination: • blood pressure • heart rate • BMI Questionnaire: • alcoholic drinks • cigarettes • drug use	Demographic: • age • gender

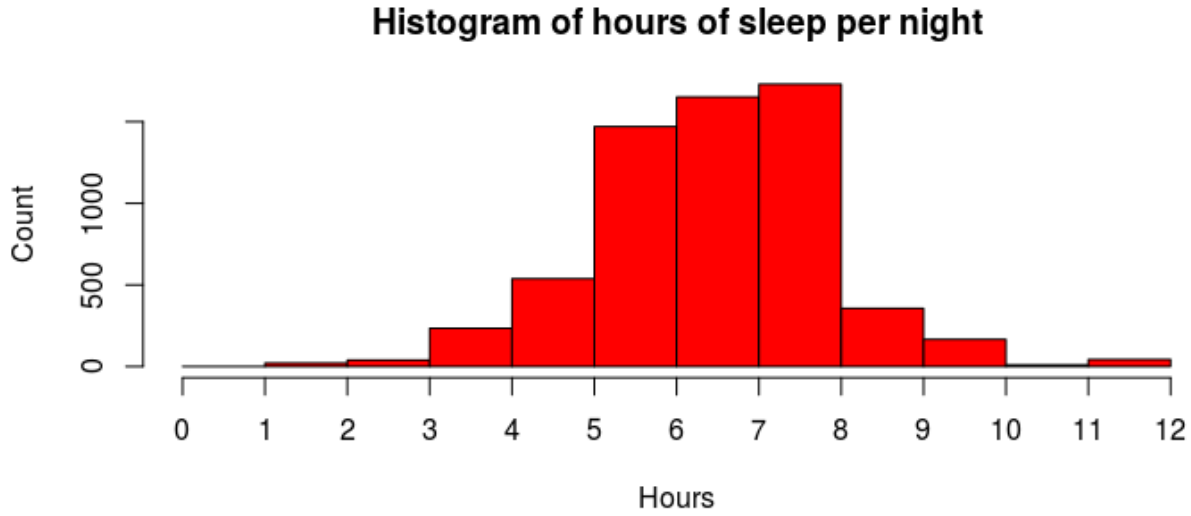
Overview of Respondent Pool



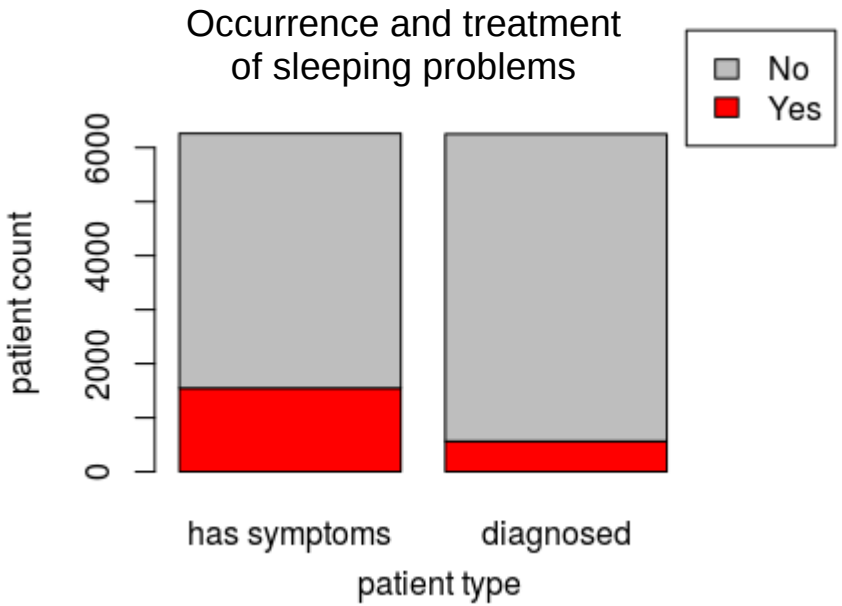
- Many of the responses are for children
- Adults are evenly distributed
- Gender:
  5,003 males
  5,172 females

Sleep Quality and Quantity

1,548 out of 6,264 respondents (24.7%) reported sleep problems

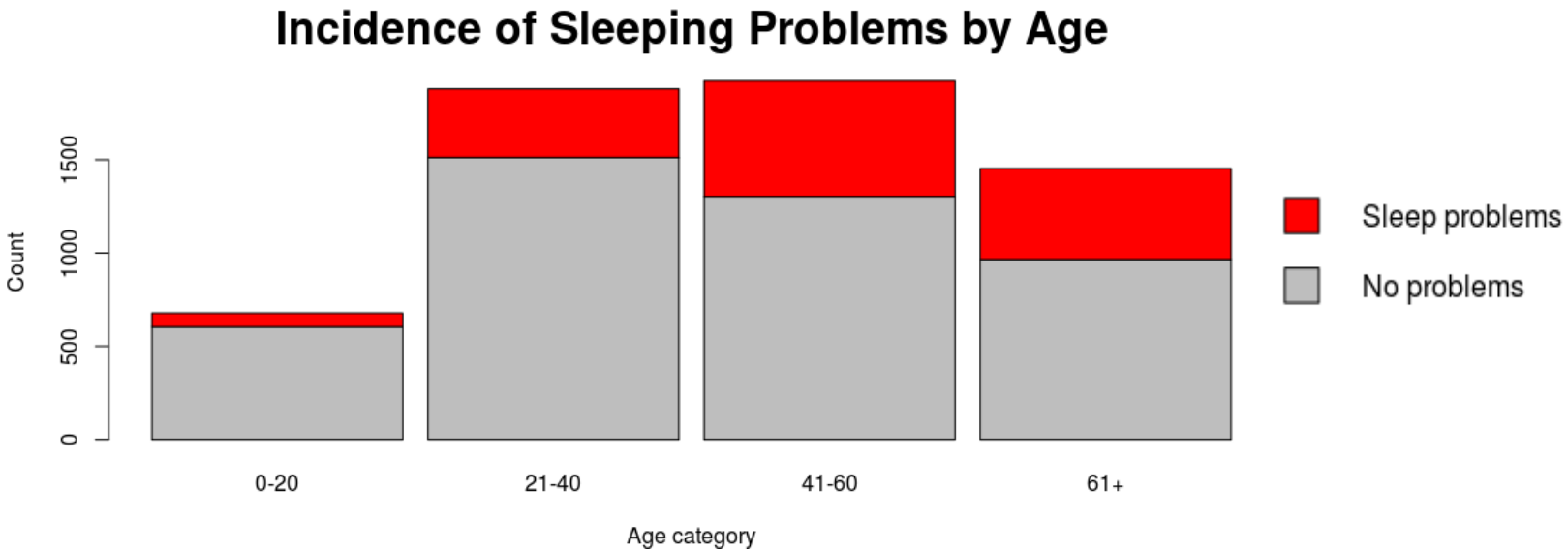


Sleep Problems Go Untreated

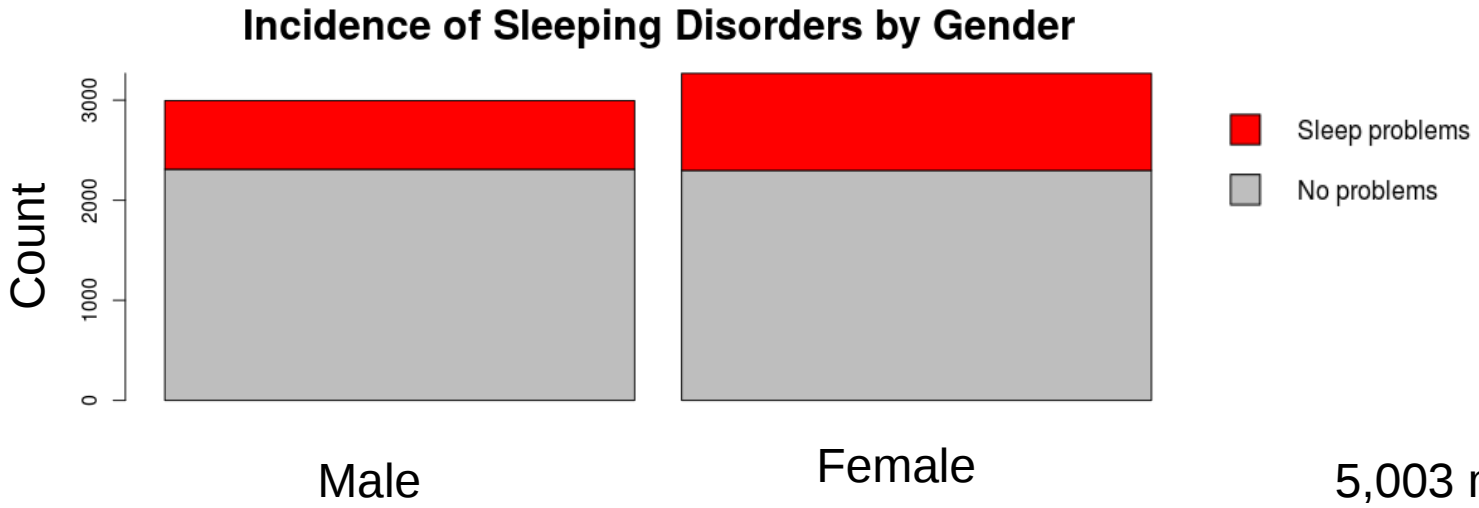


	Patient has symptoms of sleep disorder	Patient seeks diagnosis
Yes	1,548	564
No	4,716	5,689

Sleeping Problems by Age

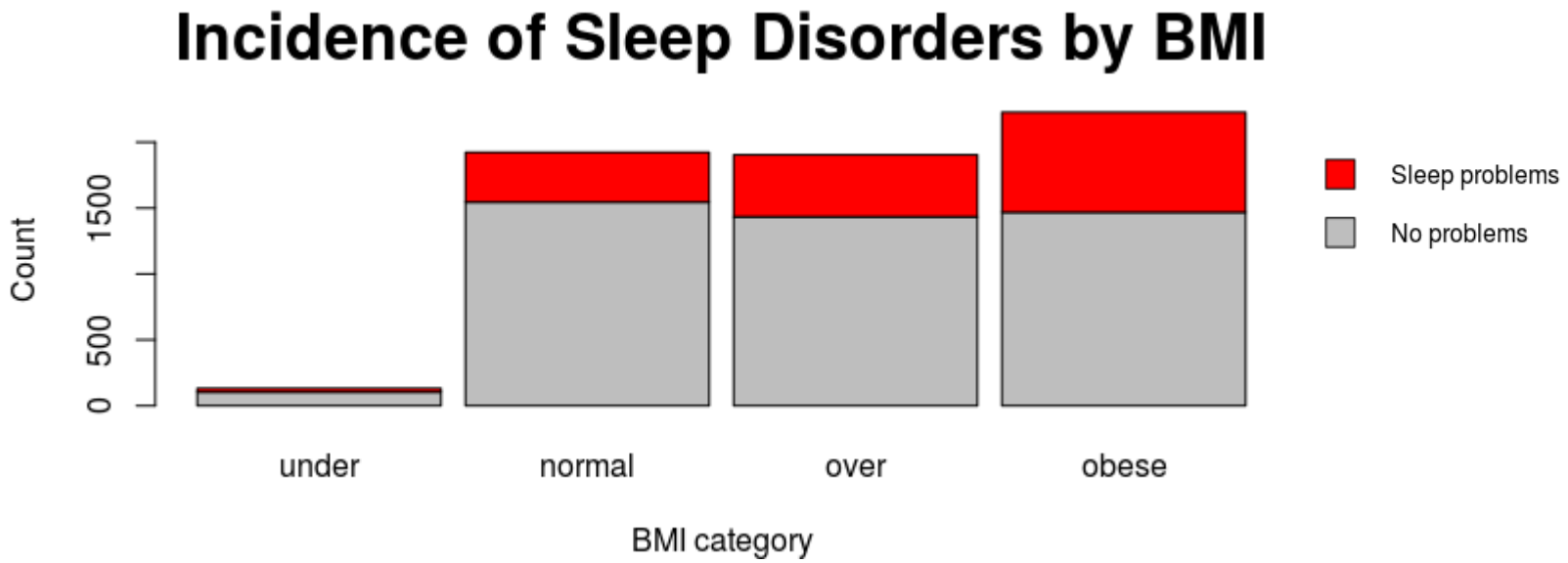


Sleeping Problems by Gender



5,003 males
5,172 females

Sleeping Disorders by BMI



Linear Regression

Evidence of association exists between sleep quantity and the following:

- body mass index
- heart rate
- age
- smoking

Not supported by evidence:

- alcohol
- gender

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.1181674  0.2719006  22.501  < 2e-16 ***
RIDAGEYR     0.0088472  0.0020084   4.405 1.11e-05 ***
RIAGENDR    -0.0169660  0.0673517  -0.252  0.80114
HR           0.0056405  0.0028182   2.001  0.04547 *
BMXBMI      -0.0108414  0.0048293  -2.245  0.02488 *
SMQ040       0.1061040  0.0373273   2.843  0.00452 **
ALQ120Q      0.0001906  0.0010575   0.180  0.85700
```

Logistic Regression

Logistic regression using all available predictors

Evidence of association exists between sleep quality/complaints and the following:

- age
- gender
- heart rate
- BMI
- smoking

Not supported by evidence:

- alcohol

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.3630813  0.4018106  -8.370  < 2e-16 ***
RIDAGEYR     0.0129772  0.0029606   4.383 1.17e-05 ***
RIAGENDR     0.4409362  0.0958158   4.602 4.19e-06 ***
HR           0.0092590  0.0040260   2.300  0.02146 *
BMXBMI       0.0384629  0.0068505   5.615 1.97e-08 ***
SMQ040      -0.1724387  0.0538261  -3.204  0.00136 **
ALQ120Q      0.0005686  0.0014466   0.393  0.69428
```

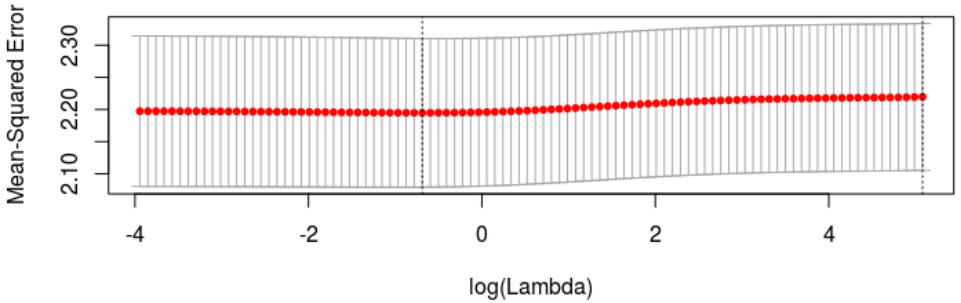
Ridge and Lasso

Regression	Test MSE
Linear	2.247
Ridge	2.343
Lasso	2.266

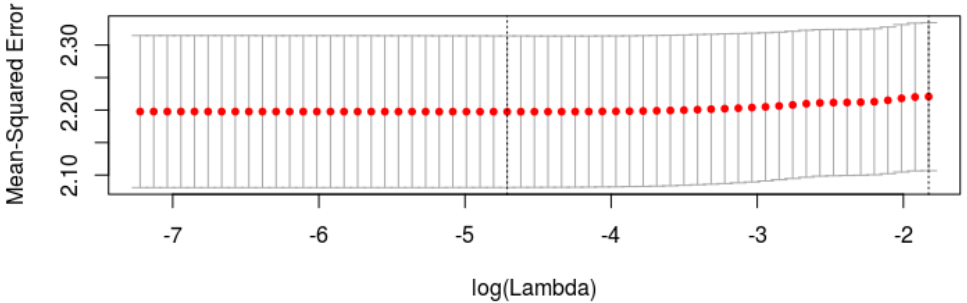
Lasso reduced the coefficients:

```
(Intercept)  6.243502960
RIDAGEYR     0.005430769
RIAGENDR    -0.019762023
HR           0.005328356
BMXBMI      -0.007877787
ALQ120Q      .
SMQ040       0.085482129
```

Ridge regression: selection of λ



Lasso regression: selection of λ



Polynomial Regression

The squares of the quantitative predictors were used in polynomial regression

Regression	Test MSE
Linear	2.247
Ridge	2.343
Lasso	2.266
Polynomial	2.194

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.66460559	0.13489343	49.4064495	0.000000e+00
poly(RIDAGEYR, 2)1	4.36677718	1.55952268	2.8000729	5.158590e-03
poly(RIDAGEYR, 2)2	8.58393934	1.50145045	5.7170980	1.247950e-08
poly(BMXBMI, 2)1	-1.05325729	1.55474309	-0.6774478	4.982009e-01
poly(BMXBMI, 2)2	0.71743657	1.50153032	0.4778036	6.328426e-01
RIAGENDR	-0.00544986	0.06915777	-0.0788033	9.371970e-01
poly(HR, 2)1	3.39127384	1.51437519	2.2393881	2.524108e-02
poly(HR, 2)2	0.74153092	1.48845309	0.4981890	6.184061e-01
SMQ040	0.07659417	0.03739205	2.0484077	4.065141e-02
poly(ALQ120Q, 2)1	0.60010791	1.48116347	0.4051598	6.854037e-01
poly(ALQ120Q, 2)2	-5.40571724	1.51488519	-3.5684006	3.676953e-04

Polynomial Regression: Analysis of Variance

For regression of sleep quantity vs. age:

Quadratic polynomial is justified

```
Analysis of Variance Table

Model 1: SLD010H ~ RIDAGEYR
Model 2: SLD010H ~ poly(RIDAGEYR, 2)
Model 3: SLD010H ~ poly(RIDAGEYR, 3)
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1   1993 4459.2
2   1992 4385.5  1    73.717 33.4766 8.358e-09 ***
3   1991 4384.3  1     1.220  0.5539    0.4568
```

For regression of sleep quantity vs. BMI:

No justification for quadratic or cubic polynomial

```
Analysis of Variance Table

Model 1: SLD010H ~ BMXBMI
Model 2: SLD010H ~ poly(BMXBMI, 2)
Model 3: SLD010H ~ poly(BMXBMI, 3)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1993 4479.7
2   1992 4478.9  1   0.77281 0.3435 0.5579
3   1991 4478.9  1   0.00855 0.0038 0.9508
```

LDA and QDA

LDA and QDA were performed to predict whether a patient had sleep problems using the following predictors:

- age
- gender
- blood pressure
- heart rate
- BMI
- smoking frequency
- drinking frequency

Technique	Success rate
LDA	64.2%
QDA	64.4%

LDA:

lda_results	0	1
0	606	329
1	28	35

QDA:

qda_results	0	1
0	588	309
1	46	55

Bagging and Random Forest

Prediction of quantitative response: number of hours of sleep

Technique	Test MSE	Most important variables
Bagging	2.399	age, alcohol
RF	2.343	age, alcohol

Bagging

```
> importance(rf_fit_subtree)
            %IncMSE IncNodePurity
RIDAGEYR 10.6435291     419.83257
BMXBMI    3.0202502     485.39702
BPS      -0.6433664     451.36870
HR        2.8223723     329.83665
SMQ040    5.4368586      88.43805
ALQ120Q   7.4194133     262.40408
```

Random Forest

```
> importance(rf_fit)
            %IncMSE IncNodePurity
RIDAGEYR 10.8060935     399.48224
BMXBMI    2.1201911     430.20637
BPS       0.5049536     417.00986
HR        1.4308605     317.13585
SMQ040    6.0040835      88.37476
ALQ120Q   6.6540039     262.37245
```

Bagging and Random Forest

Prediction of categorical response: low-quality sleep reported

Technique	Success rate	Most important variables
Bagging	63.3%	BMI, blood pressure, age
RF	63.2%	BMI, age, blood pressure

Conclusions

1. The techniques learned in this course provided distinct lines of evidence for the significance of certain factors in understanding sleep quality and quantity:
   - age
   - body mass index

2. There is no “one size fits all” solution for regression or classification problems

Future work:
- formal subset selection to identify best model(s)
- support vector classifier and regression

References

[1] Centers for Disease Control and Prevention web site <https://www.cdc.gov/features/dssleep> (accessed May 6, 2017).
[2] “Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem,” Institute of Medicine, National Academy of Sciences, Washington, D.C., 2006.
[3] E. Wickwire, F. Shaya, S. Scharf, “Health economics of insomnia treatments: The return on investment for a good night’s sleep”, *Sleep Medicine Reviews*, vol. 30, Dec. 2016, 72-82, 2016.
[4] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, ISBN 978-1-4614-7138-7, Springer Science and Business Media, 2013.

Questions?