STAT 534: Homework 3 Solution

Fall 2021

Due: Friday, October 1

1. In this problem, we will generate simulated data, and will then use this data to perform variable selection.

   (a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of the same length. Then generate a response vector Y according to the model

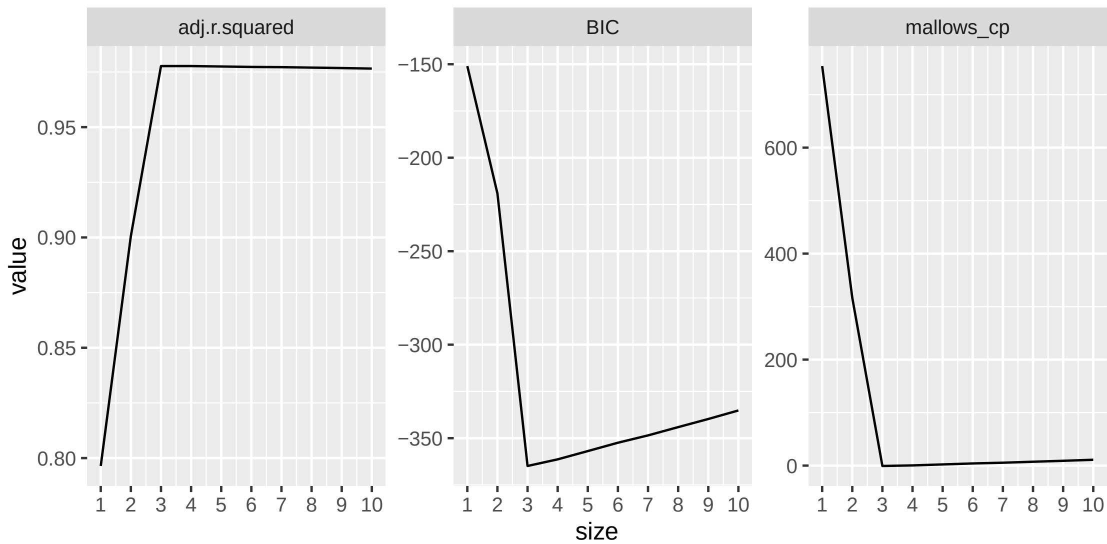
   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

   where $\beta_0, ..., \beta_3$ are constants of your choice.

```
set.seed(321)
x <- rnorm(100)
e <- rnorm(100)
y <- 10 + 3*x + 1.5*x^2 + x^3 + e
```

   (b) Given the predictors $X, X^2, ..., X^{10}$, perform best subset selection in order to choose the best model. What is the best model obtained according to C_p, AIC, BIC and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

```
### generate x, x^2,..., x^10, y
library(tidyverse)
simdata <- 1:10 %>%
  map(~x^.)
simdata <- simdata %>%
  as_tibble(.name_repair = "unique") %>%
  mutate(y = y)
### best subsets
library(leaps)
subset <- regsubsets(y~., method="exhaustive", nvmax = 10, nbest=1, data=simdata)
bs_summary <- tidy(subset)
### selection criterion
bs_summary_long <- bs_summary %>%
  mutate(size = 1:10) %>%
  pivot_longer(cols = adj.r.squared:mallows_cp,
               names_to = "criteria", values_to = "value")
bs_summary_long %>%
  ggplot(aes(x = size, y = value)) +
  facet_wrap(~criteria, scales="free_y") +
  geom_line() +
  scale_x_continuous(breaks = 1:10)
```



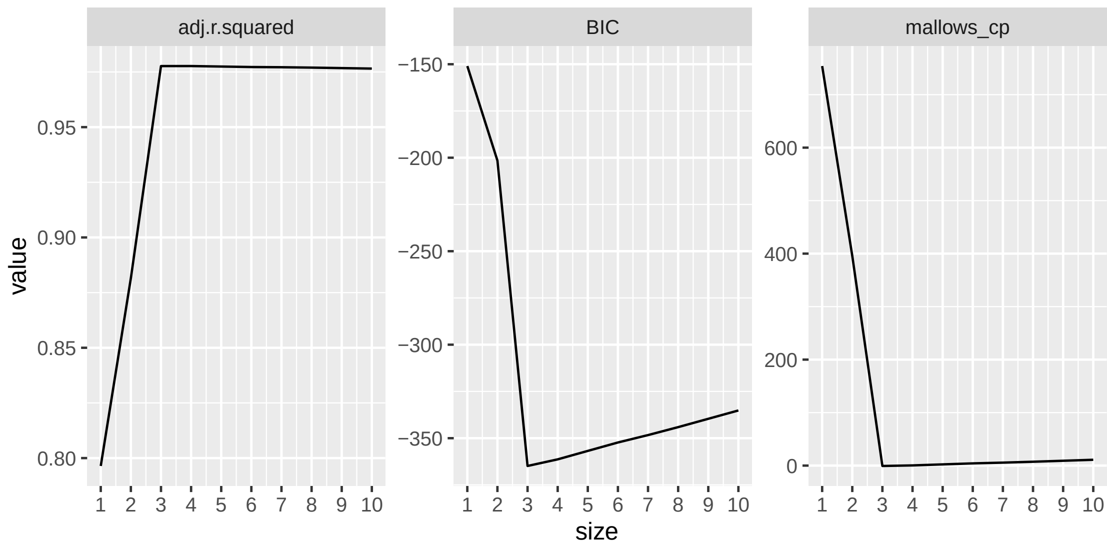

```
### final model
bs_final <- lm(y~., simdata[,c(1:3,11)])
tidy(bs_final)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   10.0      0.112       89.4 2.97e-94
## 2 ...1           3.10     0.169       18.3 4.19e-33
## 3 ...2           1.48     0.0701      21.1 9.08e-38
## 4 ...3           0.996    0.0487      20.5 8.87e-37
```

Best subsets correctly selects X, X^2, X^3 and the resulting coefficients are close to those in the true model.

(c) Repeat (b), using forward selection. How does your answer compare to the results in (b)?

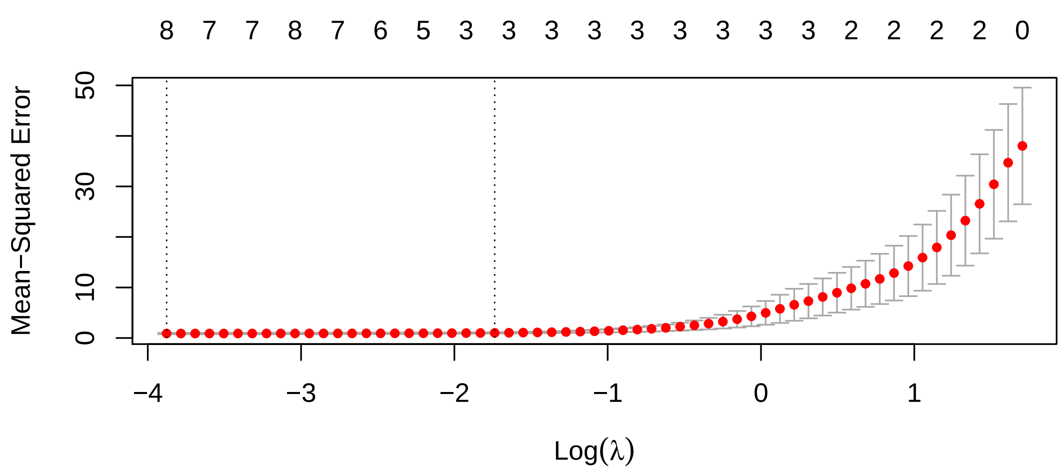
```
### forward selection
fwd <- regsubsets(y~., method = "forward", nvmax = 10, nbest=1, data = simdata)
fwd_summary <- tidy(fwd)
### selection criterion
fwd_summary_long <- fwd_summary %>%
  mutate(size = 1:10) %>%
  pivot_longer(cols = adj.r.squared:mallows_cp,
               names_to = "criteria", values_to = "value")
fwd_summary_long %>%
  ggplot(aes(x = size, y = value)) +
  facet_wrap(~criteria, scales="free_y") +
  geom_line() +
  scale_x_continuous(breaks = 1:10)
```



Forward selection leads to the same result.

(d) Now fit a lasso model and use cross-validation to select the optimal values of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

```
### cross validation
library(glmnet)
set.seed(123)
lasso_cv_out <- cv.glmnet(as.matrix(simdata[,1:10]), as.matrix(simdata[,11]),
                          alpha = 1)
plot(lasso_cv_out)
```

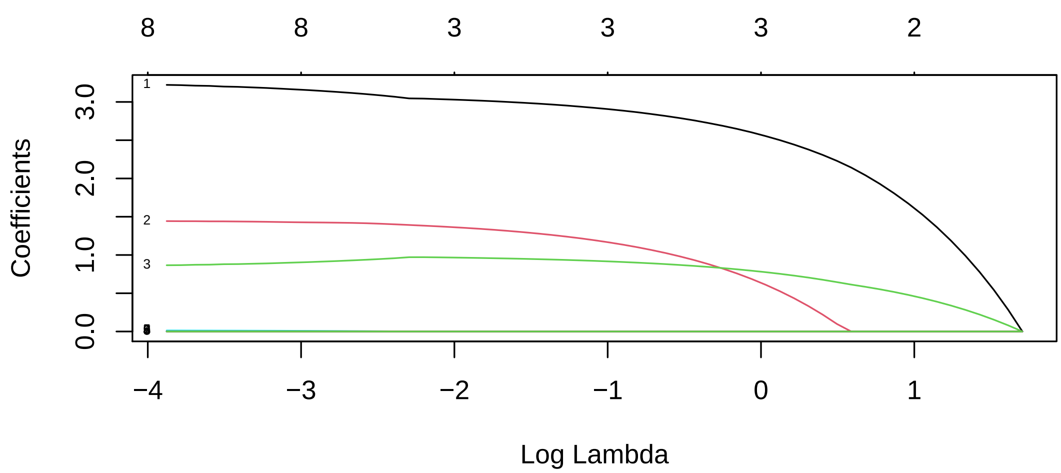



```
### final model
lasso_out <- glmnet(as.matrix(simdata[,1:10]), as.matrix(simdata[,11]), alpha = 1)
coef(lasso_out, s = lasso_cv_out$lambda.1se)

## 11 x 1 sparse Matrix of class "dgCMatrix"
##                      1
```

```
## (Intercept) 10.1499010
## ...1         3.0089680
## ...2         1.3283686
## ...3         0.9574082
## ...4         .
## ...5         .
## ...6         .
## ...7         .
## ...8         .
## ...9         .
## ...10        .

plot(lasso_out, label = TRUE, xvar = "lambda")
```



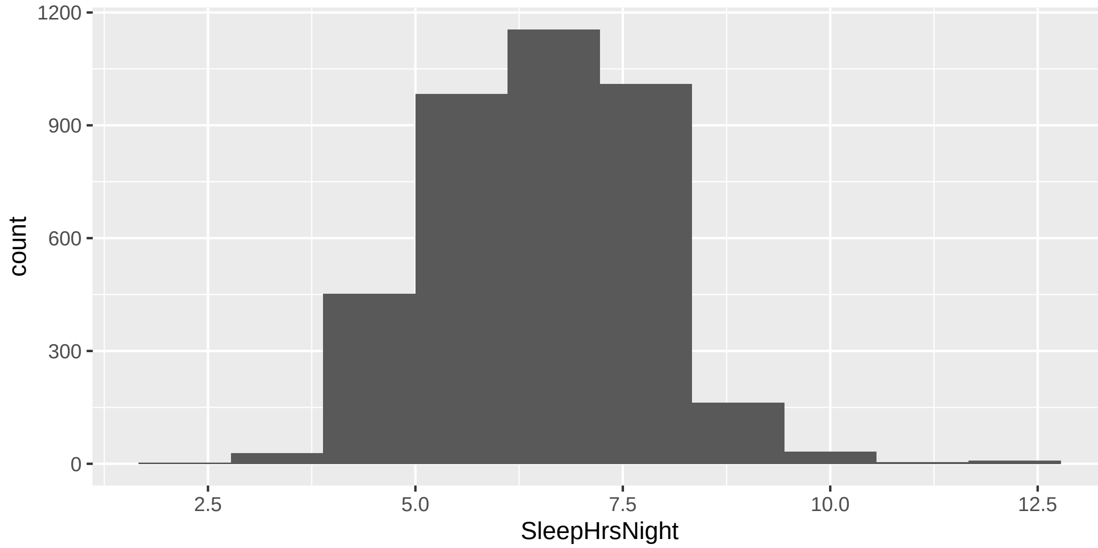
Lasso also correctly identifies the important predictors and the resulting coefficients are close to their respective true values.

2. (Excerpted from Problem 11.6 in textbook) The ability to get a good night's sleep is correlated with many positive health outcomes. Use the *NHANES* data set from the *NHANES* package to predict *SleepHrsNight*. Check the R document for detailed information about the data set.

   (a) First separate the data set at random into 75% training and 25% testing sets.

   (b) Select your own predictors, and create plots or summary tables to explore the variables.

```
library(NHANES)
data(NHANES)
sleep <- NHANES %>%
  filter(Age>=18) %>%
  dplyr::select(c(50,3,4,7,9,10,13,16,21,25,26,34,35,40,46,52,60,62,65,69,72)) %>%
  na.omit
### exploratory study
# response
ggplot(sleep, mapping = aes(x=SleepHrsNight)) +
  geom_histogram(bins = 10) # no transformation needed
```



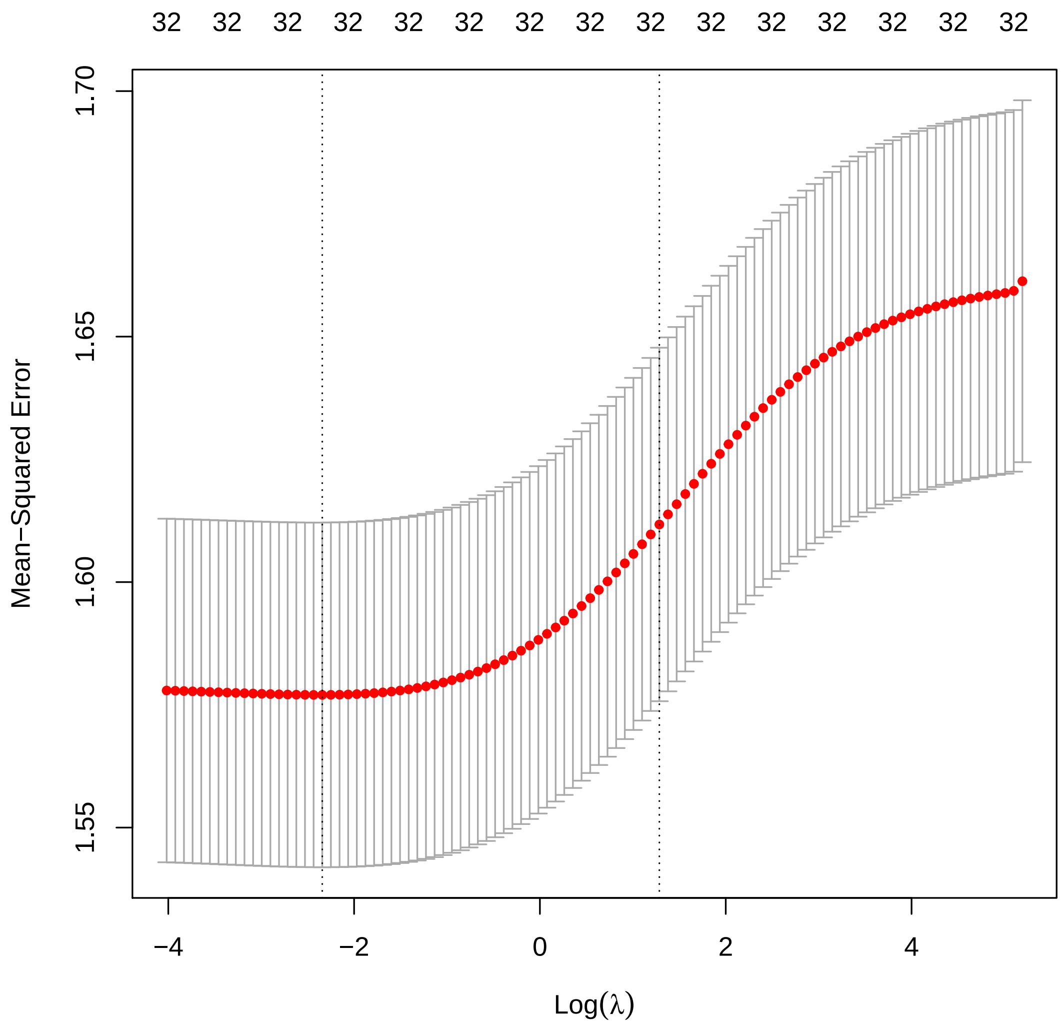

```r
# omit the exploration of the predictors...
### training vs. testing
n <- nrow(sleep)
set.seed(1)
train <- sample(1:n,round(0.75*n))
xtest <- sleep[-train,-1]
ytest <- sleep$SleepHrsNight[-train]
```

I select all adults and the following predictors: Gender, Age, Race1, Education, MaritalStatus, Poverty, Work, BMI, BPSysAve, BPDiaAve, DirectChol, TotChol, Diabetes, Depressed, PhysActive, AlcoholYear, Smoke100, Marijuana, HardDrugs, and SexNumPartnLife.

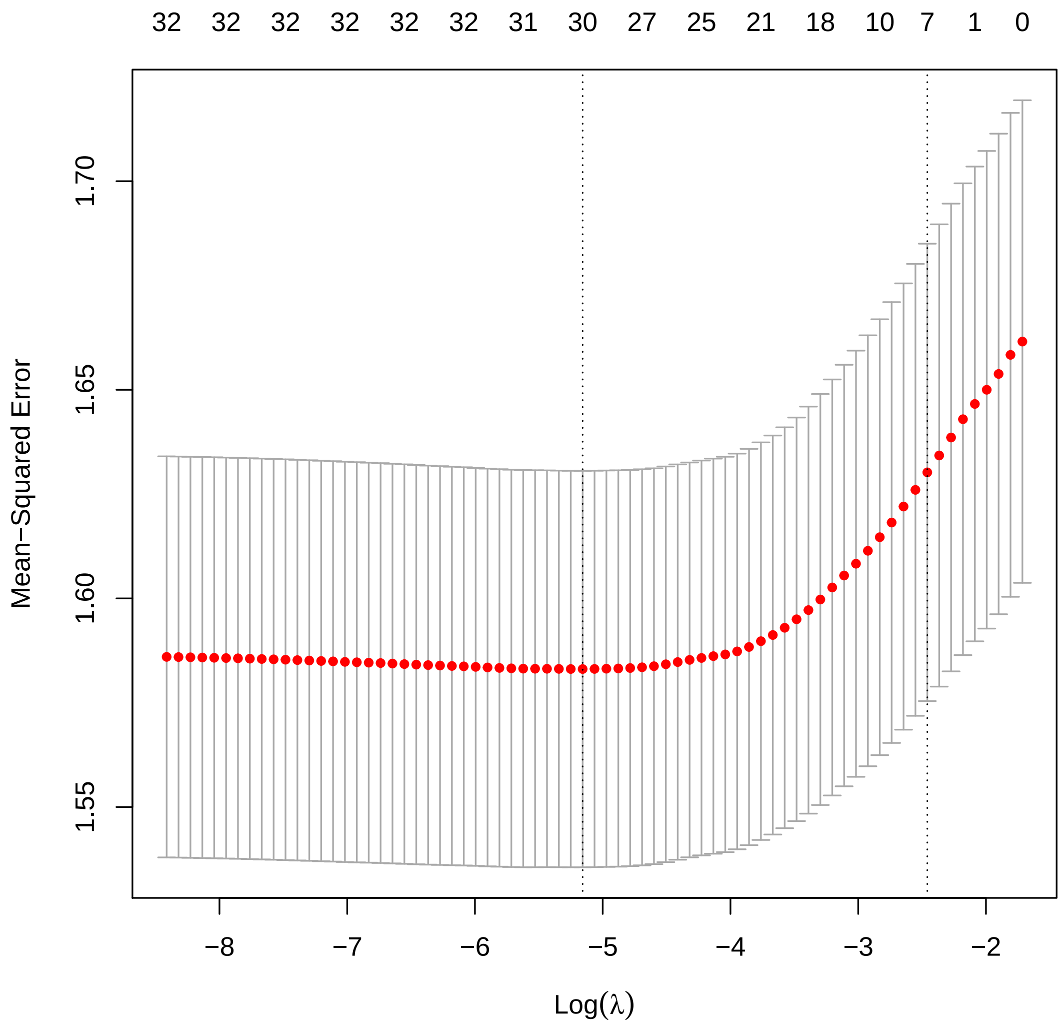
(c) Build the following models using the training set with your predictors of choice:

- Multiple linear regression
- Ridge regression
- LASSO regression

```r
### multiple linear regression
lm_fit <- lm(SleepHrsNight~., data = sleep, subset = train)
lm_tr <- summary(lm_fit)$sigma^2 #training error
lm_pred <- predict(lm_fit, xtest)
lm_te <- mean((lm_pred-ytest)^2) #testing error
### ridge
x <- model.matrix(SleepHrsNight~., sleep)[,-1]
xtrain <- x[train,]
xtest <- x[-train,]
y <- sleep$SleepHrsNight
ytrain <- y[train]
set.seed(2)
ridge_cv_out <- cv.glmnet(xtrain, ytrain, alpha = 0)
plot(ridge_cv_out)
```




```
ridge_pred <- predict(ridge_cv_out,
                      s = ridge_cv_out$lambda.1se,
                      newx = xtrain)
ridge_tr <- mean((ridge_pred-ytrain)^2) #training error
ridge_pred <- predict(ridge_cv_out,
                      s = ridge_cv_out$lambda.1se,
                      newx = xtest)
ridge_te <- mean((ridge_pred-ytest)^2) #testing error
### lasso
set.seed(3)
lasso_cv_out <- cv.glmnet(xtrain, ytrain, alpha = 1)
plot(lasso_cv_out)
```




```
lasso_pred <- predict(lasso_cv_out,
                      s = lasso_cv_out$lambda.1se,
                      newx = xtrain)
lasso_tr <- mean((lasso_pred-ytrain)^2) #training error
lasso_pred <- predict(lasso_cv_out,
                      s = lasso_cv_out$lambda.1se,
                      newx = xtest)
lasso_te <- mean((lasso_pred-ytest)^2) #testing error
```

(d) Compare the effectiveness of each model on training vs. testing data.

```
errsum <- tribble(
  ~model, ~train, ~test,
  "MLR", lm_tr, lm_te,
  "Ridge", ridge_tr, ridge_te,
  "LASSO", lasso_tr, lasso_te
)
errsum
```

```
## # A tibble: 3 x 3
##   model train  test
##   <chr> <dbl> <dbl>
## 1 MLR    1.56  1.56
## 2 Ridge  1.60  1.62
## 3 LASSO  1.62  1.64
```

(e) Choose one best model and interpret the results. What have you learned about people's sleeping quality?
Since the performances of the three models are comparable, I choose lasso considering sparsity.

```
lasso_out <- glmnet(x, y, alpha = 1)
coef(lasso_out, s = lasso_cv_out$lambda.1se)
```

```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                      1
## (Intercept)                6.872098172
## Gendermale                -0.006119974
## Age                        .
## Race1Hispanic              .
## Race1Mexican               .
## Race1White                 .
## Race1Other                 .
## Education9 - 11th Grade    .
## EducationHigh School       .
## EducationSome College      .
## EducationCollege Grad      0.025314019
## MaritalStatusLivePartner   .
## MaritalStatusMarried       .
## MaritalStatusNeverMarried  .
## MaritalStatusSeparated     .
## MaritalStatusWidowed       .
## Poverty                    0.009956879
## WorkNotWorking             .
## WorkWorking                .
## BMI                        .
## BPSysAve                   .
## BPDiaAve                   .
## DirectChol                 .
## TotChol                    .
## DiabetesYes                .
## DepressedSeveral          -0.016894651
## DepressedMost             -0.299777805
## PhysActiveYes              .
## AlcoholYear                .
## Smoke100Yes               -0.154007119
## MarijuanaYes               .
## HardDrugsYes               .
## SexNumPartnLife            .
```

Severe depression and smoking are the two major negative factors. Higher socioeconomic status is associated with longer sleeping hours.