Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Ranking States by SAT Performance: Exploratory Analysis

Georgia Tech

1

About This Lesson

Georgia Tech

2

Linear Regression: Example 2

Controlling Factors:

X_1 = % of total eligible students in the state who took the exam

X_6 = Median percentile of ranking of test takers within their secondary school classes

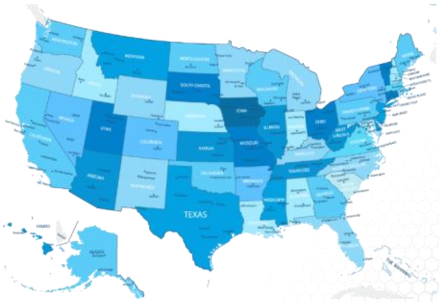
Explanatory Factors:

X_2 = Median income of families of test takers, in hundreds of dollars

X_3 = Average number of years that test takers had in social sciences, natural sciences, and humanities

X_4 = % of test takers who attended public schools

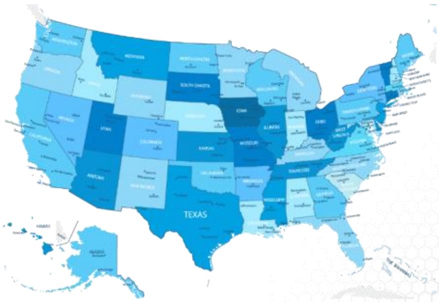
X_5 = State expenditure on secondary schools, in hundreds of dollars per student



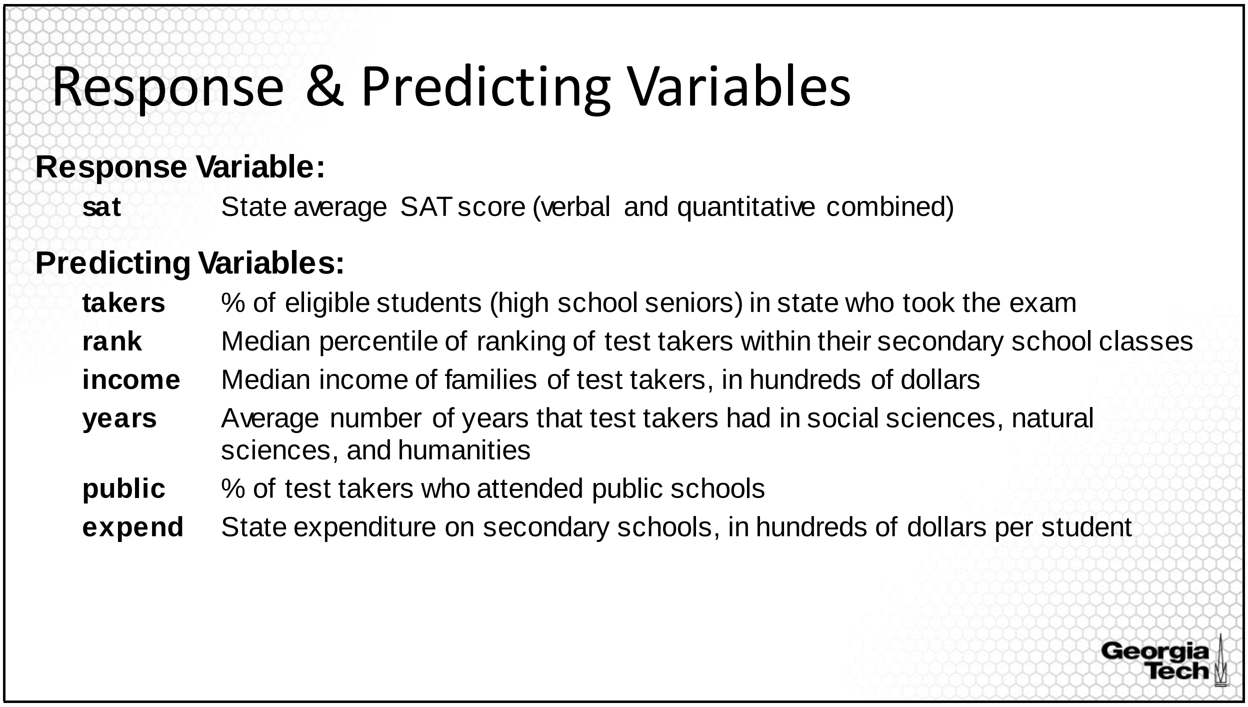
3

Ranking States by SAT Performance

- ***Which variables are associated with state average SAT scores?***
- ***After accounting for selection biases, how do the states rank?***
- ***Which states perform best for the amount of money they spend?***



4



Response & Predicting Variables

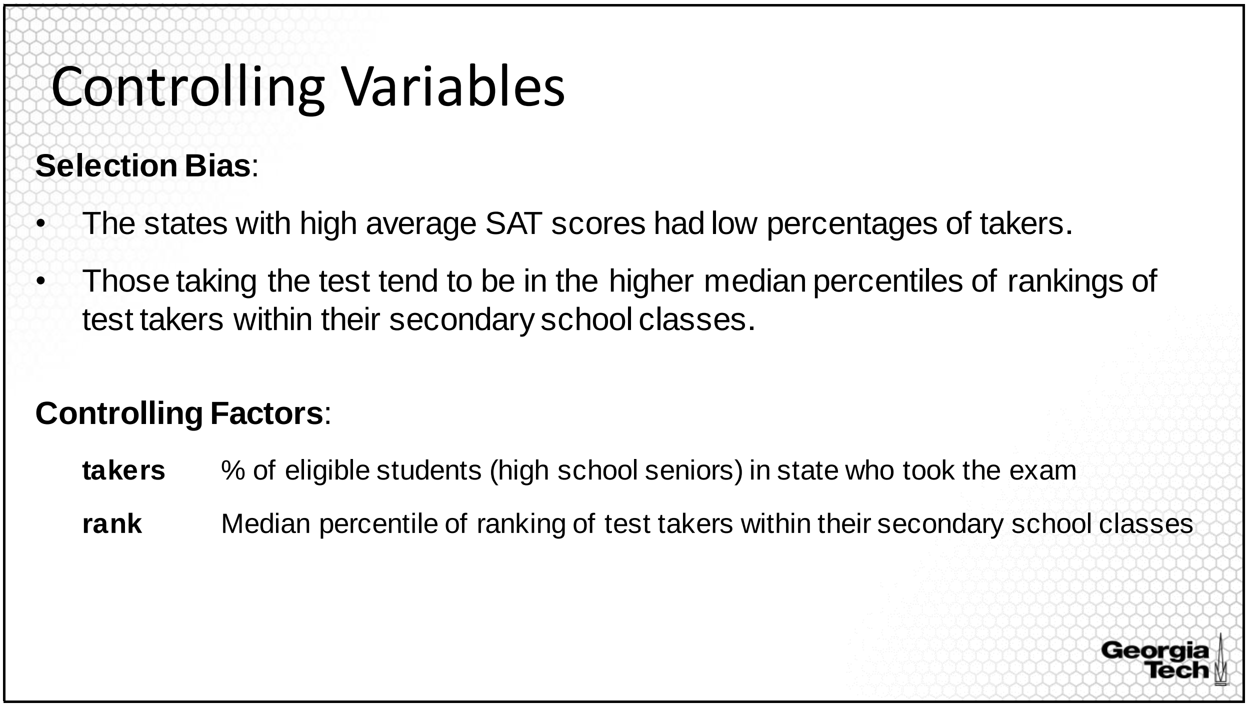
Response Variable:

sat	State average SAT score (verbal and quantitative combined)

Predicting Variables:

takers	% of eligible students (high school seniors) in state who took the exam
rank	Median percentile of ranking of test takers within their secondary school classes
income	Median income of families of test takers, in hundreds of dollars
years	Average number of years that test takers had in social sciences, natural sciences, and humanities
public	% of test takers who attended public schools
expend	State expenditure on secondary schools, in hundreds of dollars per student

5



Controlling Variables

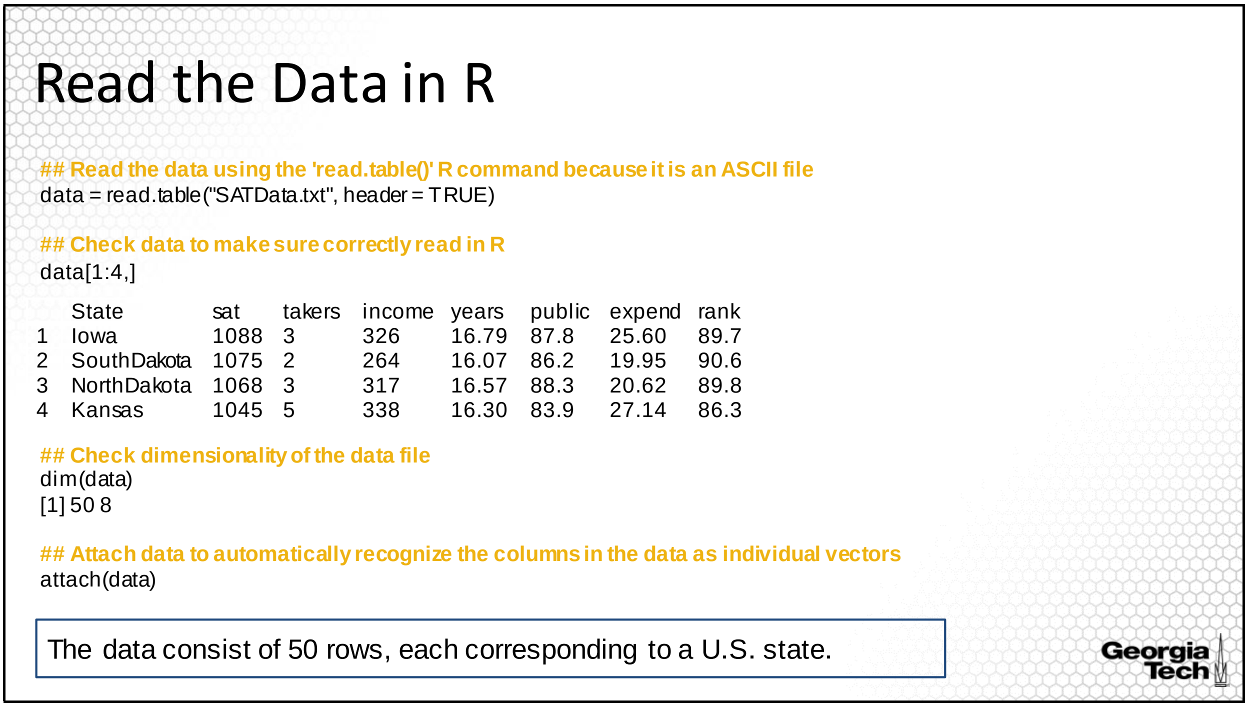
Selection Bias:

- The states with high average SAT scores had low percentages of takers.
- Those taking the test tend to be in the higher median percentiles of rankings of test takers within their secondary school classes.

Controlling Factors:

takers	% of eligible students (high school seniors) in state who took the exam
rank	Median percentile of ranking of test takers within their secondary school classes

6



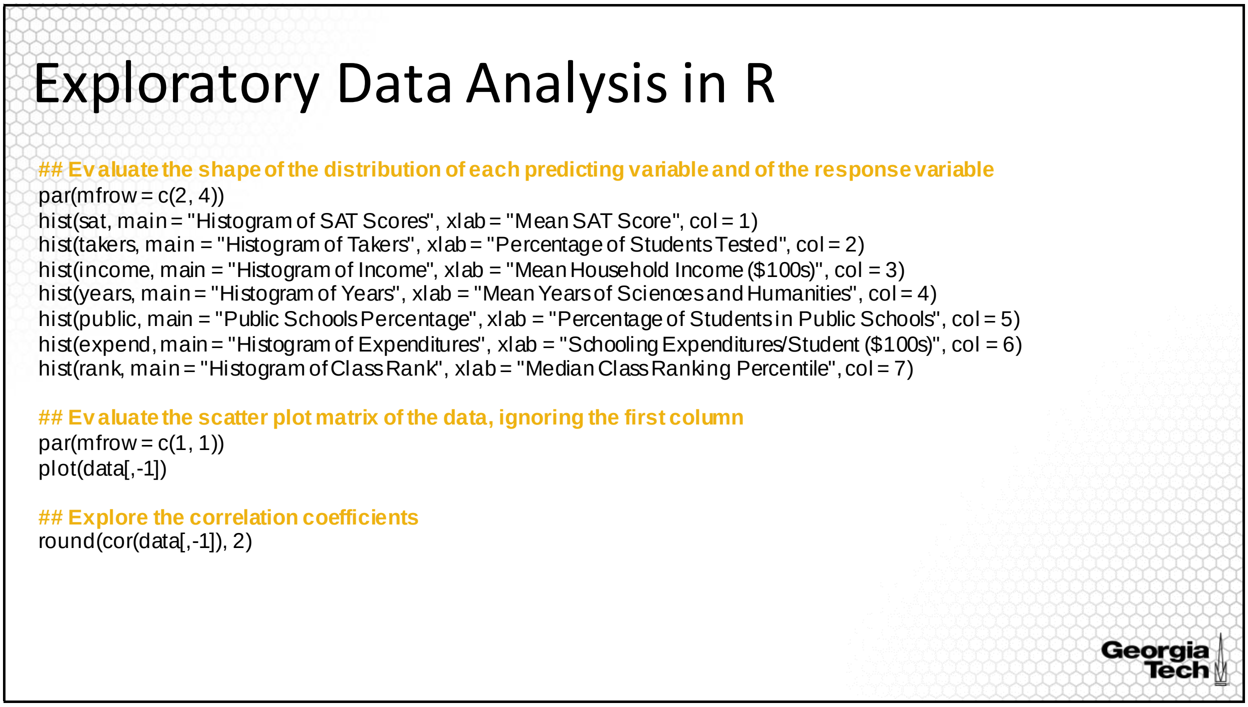
Read the Data in R

```
## Read the data using the 'read.table()' R command because it is an ASCII file
data = read.table("SATData.txt", header = TRUE)

## Check data to make sure correctly read in R
data[1:4,]

  State        sat   takers  income  years  public  expend  rank
1 Iowa         1088  3       326     16.79  87.8    25.60   89.7
2 SouthDakota  1075  2       264     16.07  86.2    19.95   90.6
3 NorthDakota  1068  3       317     16.57  88.3    20.62   89.8
4 Kansas       1045  5       338     16.30  83.9    27.14   86.3

## Check dimensionality of the data file
dim(data)
[1] 50 8

## Attach data to automatically recognize the columns in the data as individual vectors
attach(data)
```

The data consist of 50 rows, each corresponding to a U.S. state.

Georgia Tech

7



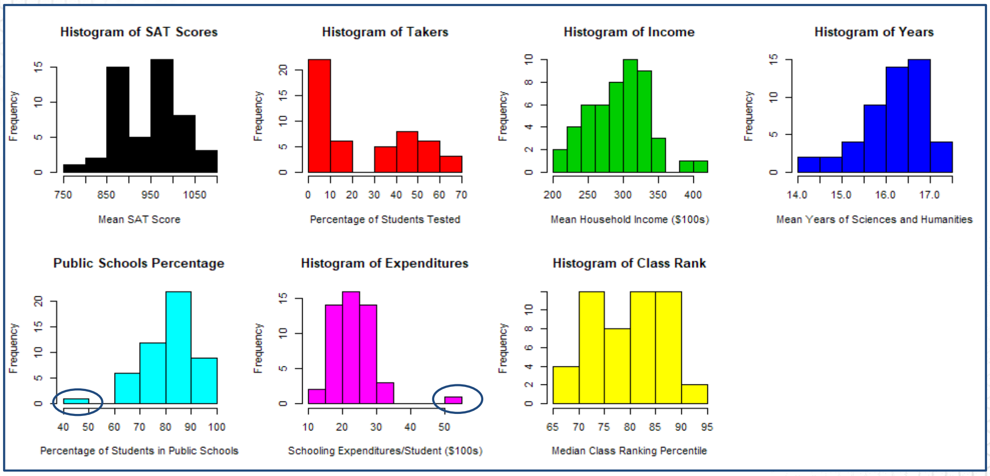
Exploratory Data Analysis in R

```
## Evaluate the shape of the distribution of each predicting variable and of the response variable
par(mfrow = c(2, 4))
hist(sat, main = "Histogram of SAT Scores", xlab = "Mean SAT Score", col = 1)
hist(takers, main = "Histogram of Takers", xlab = "Percentage of Students Tested", col = 2)
hist(income, main = "Histogram of Income", xlab = "Mean Household Income ($100s)", col = 3)
hist(years, main = "Histogram of Years", xlab = "Mean Years of Sciences and Humanities", col = 4)
hist(public, main = "Public Schools Percentage", xlab = "Percentage of Students in Public Schools", col = 5)
hist(expend, main = "Histogram of Expenditures", xlab = "Schooling Expenditures/Student ($100s)", col = 6)
hist(rank, main = "Histogram of Class Rank", xlab = "Median Class Ranking Percentile", col = 7)

## Evaluate the scatter plot matrix of the data, ignoring the first column
par(mfrow = c(1, 1))
plot(data[,-1])

## Explore the correlation coefficients
round(cor(data[,-1]), 2)
```

Georgia Tech

8

Exploratory Data Analysis in R

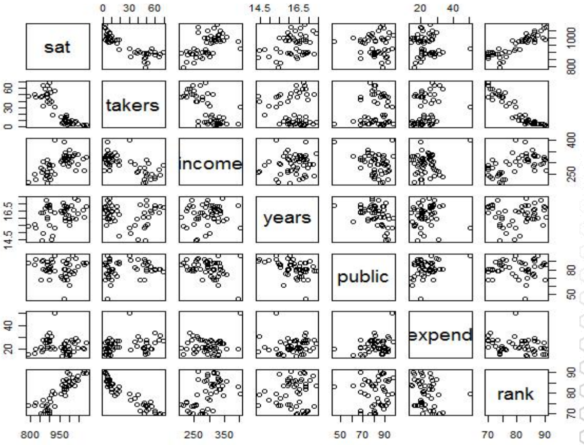


Georgia Tech

9

Exploratory Data Analysis in R (Cont'd)

```
          sat  takers  income  years public expend  rank
sat       1.00  -0.86   0.58   0.33 -0.08  -0.06  0.88
takers   -0.86   1.00  -0.66  -0.10  0.12   0.28 -0.94
income    0.58  -0.66   1.00   0.13 -0.31   0.13  0.53
years     0.33  -0.10   0.13   1.00 -0.42   0.06  0.07
public   -0.08   0.12  -0.31  -0.42  1.00   0.28  0.05
expend   -0.06   0.28   0.13   0.06  0.28   1.00 -0.26
rank      0.88  -0.94   0.53   0.07  0.05  -0.26  1.00
```



Georgia Tech

10

Summary

11