Covariates

INTRODUCTION TO STATISTICAL MODELING IN R

Danny Kaplan
Instructor

Some uses for models

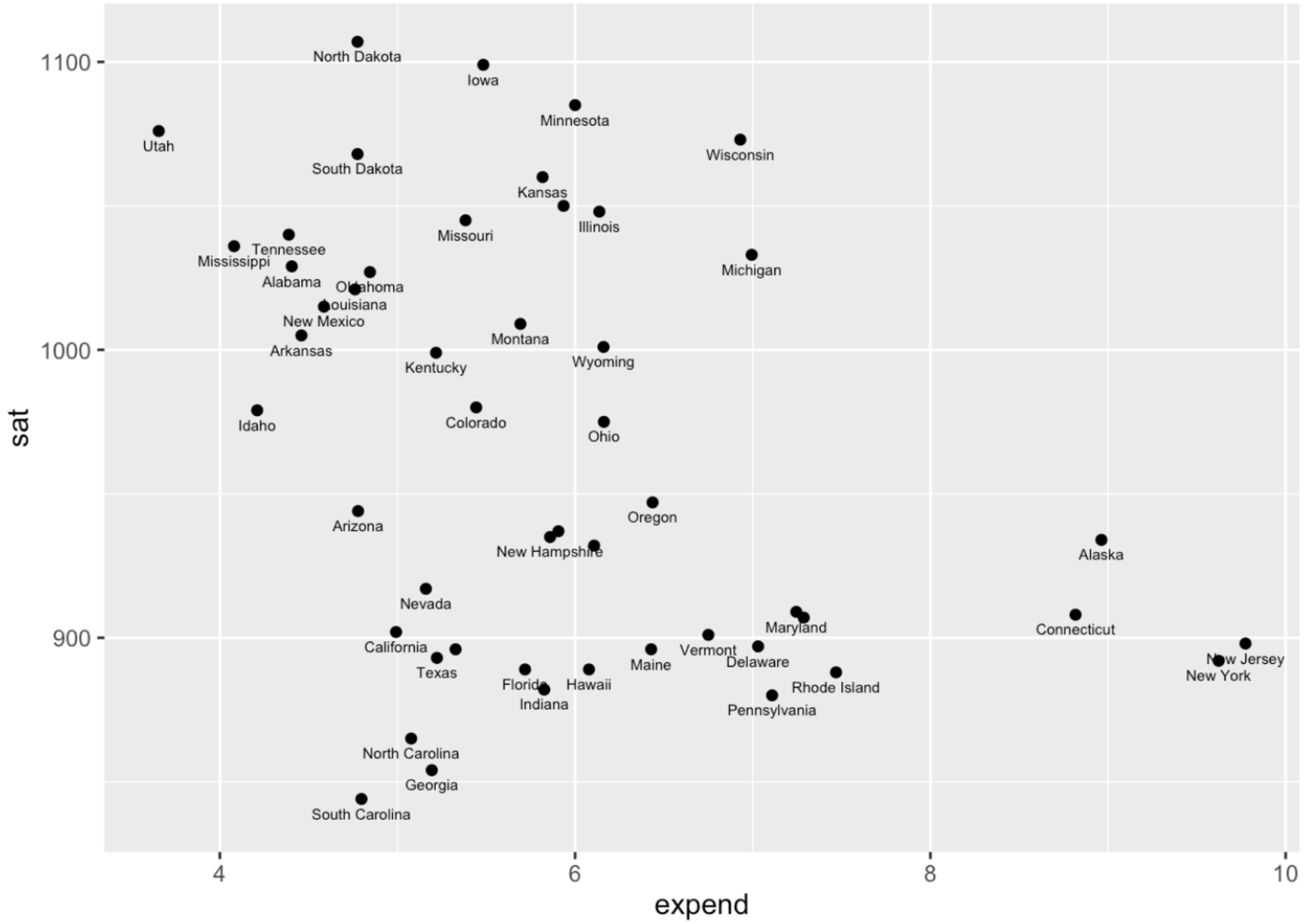
- Making predictions with available data
- Exploring a large, complex data set
- Anticipate outcome of intervention in system

Modeling educational outcomes

```
head(SAT)
```

```
       state expend ratio salary frac verbal math  sat
1    Alabama  4.405  17.2 31.144    8    491  538 1029
2     Alaska  8.963  17.6 47.951   47    445  489  934
3    Arizona  4.778  19.3 32.175   27    448  496  944
4   Arkansas  4.459  17.1 28.934    6    482  523 1005
5 California  4.992  24.0 41.078   45    417  485  902
6   Colorado  5.443  18.4 34.571   29    462  518  980
```

SAT scores and school expenditures



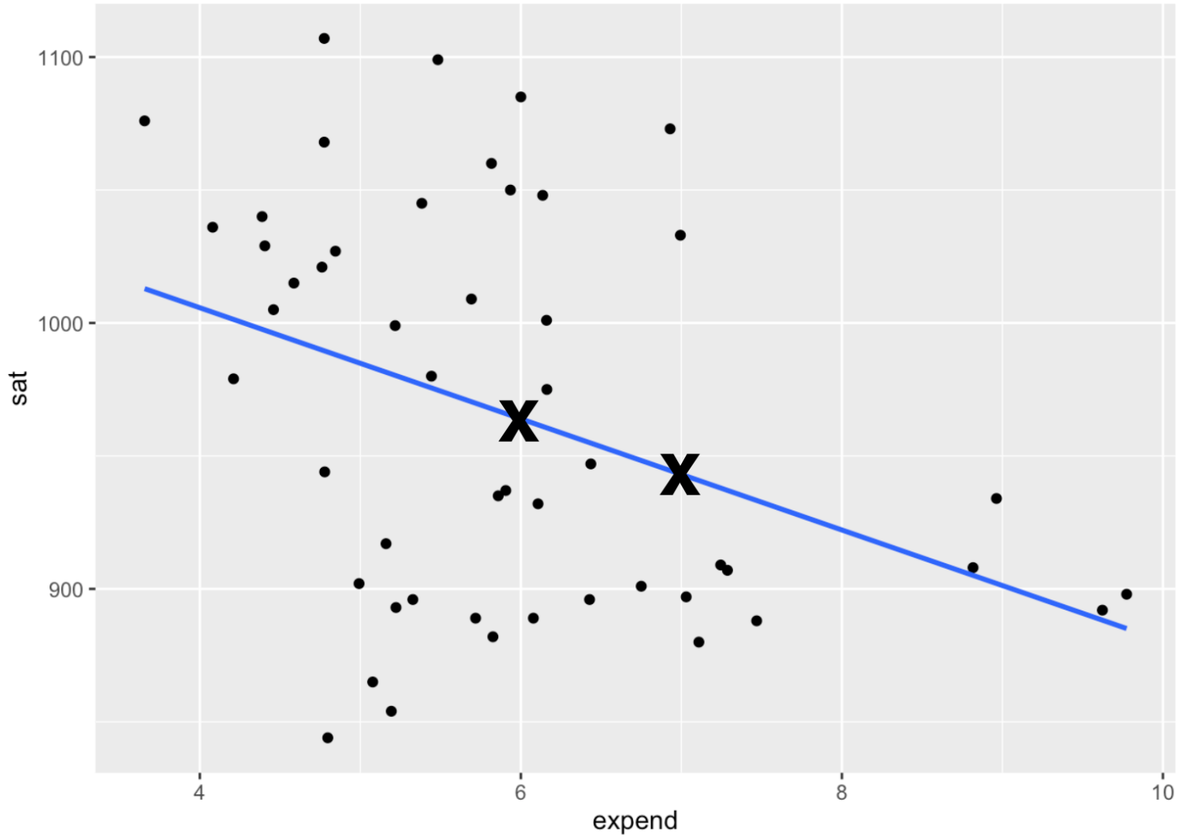
Modeling SAT as a function of expenditures

```
predict(mod_a, newdata
= data.frame(expend = 7))
```

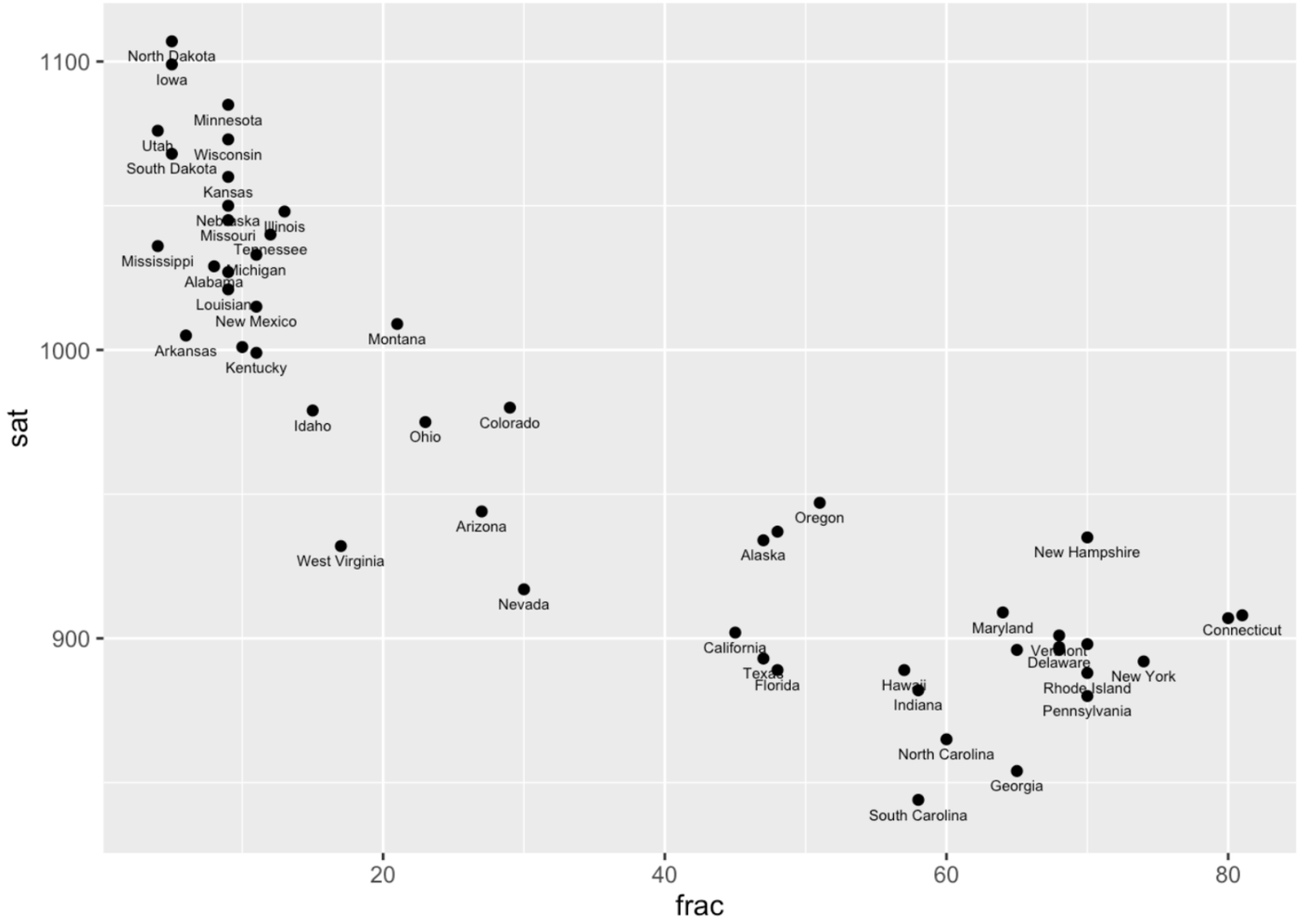
```
1    943.0485
```

```
predict(mod_a, newdata
= data.frame(expend = 6)
```

```
1    963.9407
```



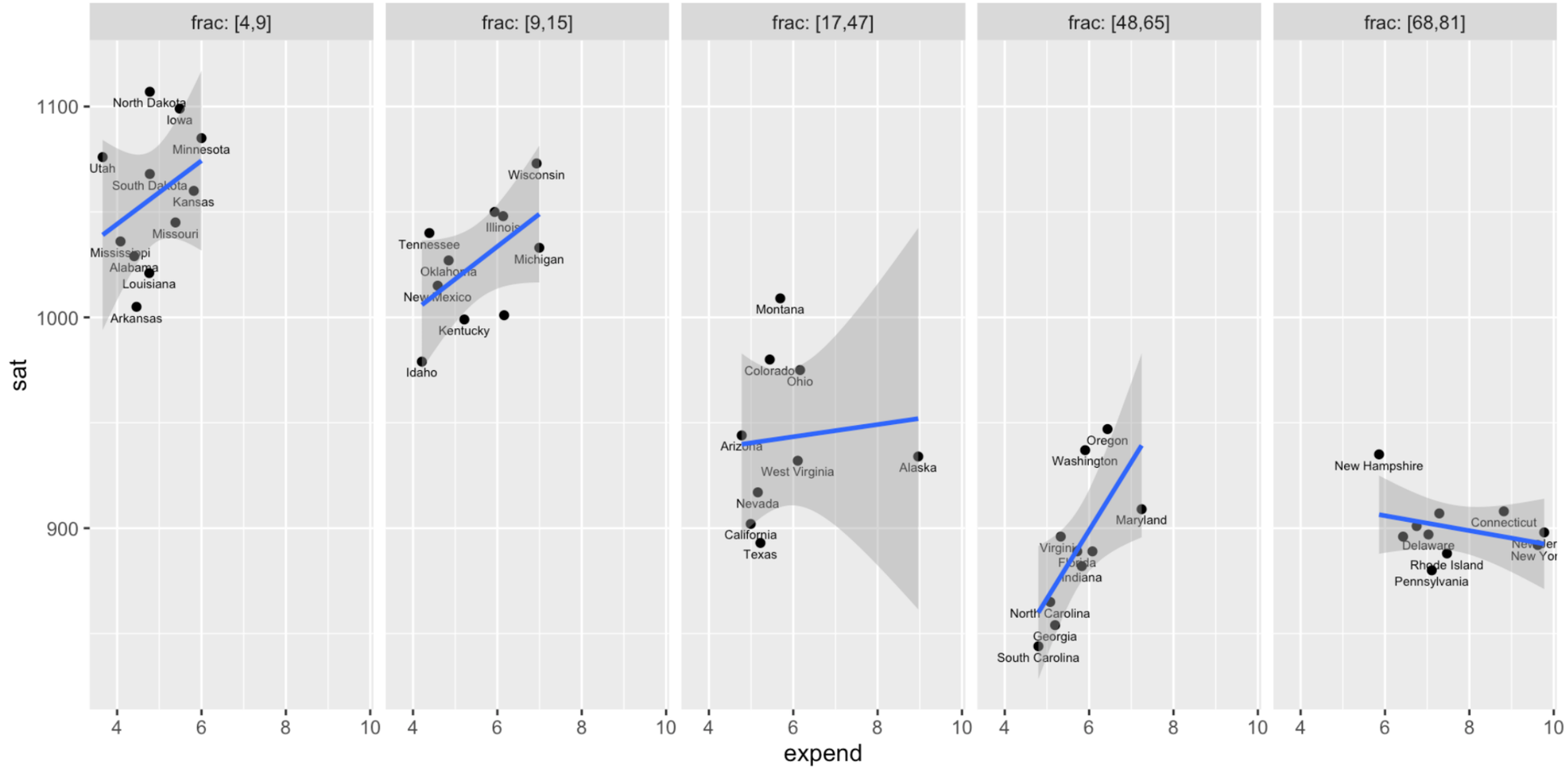
Average SAT score vs. fraction taking the test



Covariates

Explanatory variables that are not themselves of interest to the modeler, but which may shape the response variable

Stratifying by fraction taking the test



A model with expend and frac

```
# Train model
mod_b <- lm(sat ~ expend + frac, data = SAT)

# Modeling experiment with frac constant
predict(mod_b, newdata = data.frame(expend = 7, frac = 0.5))
```

```
   1
1078
```

```
predict(mod_b, newdata = data.frame(expend = 6, frac = 0.5))
```

```
   1
1066
```

Some possible models

- `sat ~ frac` : Capture state-to-state variation in SAT scores, ignoring `expend`
- `sat ~ expend` : See how expend relates to state-to-state variation in SAT scores, ignoring `frac`
- `sat ~ expend + frac` : See the role of expend in the context of what's explained by `frac`

Let's practice!

INTRODUCTION TO STATISTICAL MODELING IN R

Effect size

INTRODUCTION TO STATISTICAL MODELING IN R

Danny Kaplan
Instructor

datacamp

Measuring effect sizes

- How does changing an input to a model change the output?
  - Does the output go up or down?
- How much does the model output change for a given change in the input?
  - Effect size

Cause and effect

- In our model, the inputs *cause* the output
- Modeler's interest is often in cause and effect
- Doesn't mean the real world system works that way
- For models to give insight into cause and effect, we must build models that are faithful...

Natural units for effect sizes

- Quantitative inputs and outputs have *natural units*
  - Wages measured in $/hour
  - Education in years of schooling
- Can quantify effect size as a *rate* or a *difference*

Effect size for quantitative input

- Effect size represented as a rate
- Change in response / change in input
- For example: $/hour per year

Effect size for categorical input

- Units of effect size are those of the response variable
- Categorical variables *do not* have units
- Effect size represented as a *difference*
  - Numerical difference in output when the input is changed from one category to another

Calculating effect size

```
# Train model
wage_model <- lm(wage ~ educ + sector + sex + expend,
                 data = CPS85)
library(statisticalModeling)
evaluate_model(wage_model,
           at = list(educ = 11:12, sector = "prof",
           sex = "F", exper = 10))
```

```
  educ sector sex exper model_output
1   11   prof   F    10     7.077766
2   12   prof   F    10     7.795729
```

Let's practice!

INTRODUCTION TO STATISTICAL MODELING IN R