

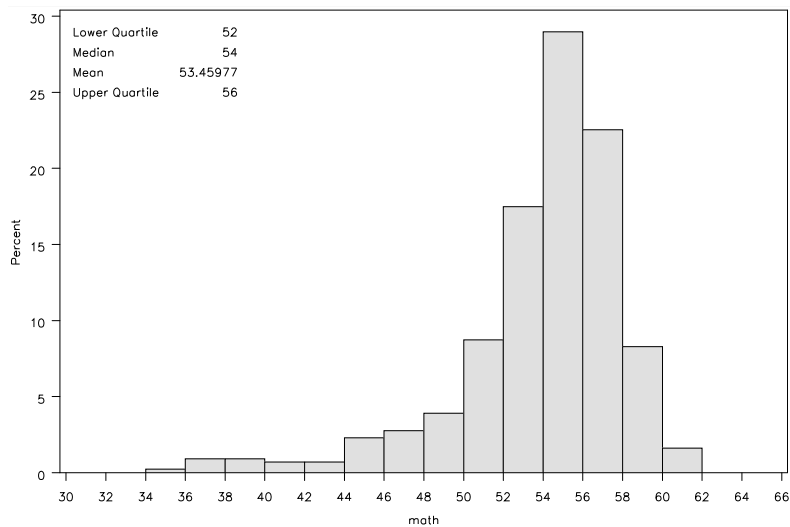
Lecture 7

1. Regression example: Minnesota 8th grade math tests
2. Log transformations
3. Horizontal outliers

1

Linear Regression Example: Minnesota 8th-grade math scores, 2000

Each year, eighth grade students in Minnesota take reading and math tests. Department of Children, Families, and Learning posts **school average scores** with characteristics at the school level. In 2000, passing score for math was 50 correct of 68 questions.



2

In 2000, Department of Children, Families, and Learning released these school-level characteristics:

- district school district number
- school ID number
- LEP_pct percent of students with Limited English Proficiency
- Special_Ed_pct percent of students in Special Education
- free_lunch_pct percent of students receiving free or reduced-price lunch
- mobility_pct mobility index (percent)
- drop_out_pct percent of students dropping out of school
- total_8th_graders total eighth grade enrollment
- total_students kindergarten through 12th grade enrollment
- operating_budget district operating expenditure per student (*district level*)
- total_budget district total expenditure per student (*district level*)

3

Aim: model math score on school and district characteristics.

- Identify characteristics of schools with high (or low) average math score
- Estimate “effect” of changing a characteristic on average math score
Which characteristics could be changed intentionally?

4

Strategy

1. Check the data. Plot the response against each continuous predictors.

Look for outliers in the horizontal direction and errors.

Set aside (temporarily omit) horizontal outliers.

Identify positive-valued variables (predictors and response) with maximum/minimum > 10 , and take logs. Use the logged variables in the analysis.

Identify cases that are extreme outliers in the horizontal direction and make a dataset without them; call this the *trimmed subset*.

If a predictor appears to have a curved association with the response, make a new variable of the predictor squared.

5

2. Fit a full model for school average math score. Check residual plot(s).

3. Try to reduce the model by dropping non-significant predictors.

4. Examine results from the final model.

Restore omitted points and re-fit—what changes?

5. Summarize.

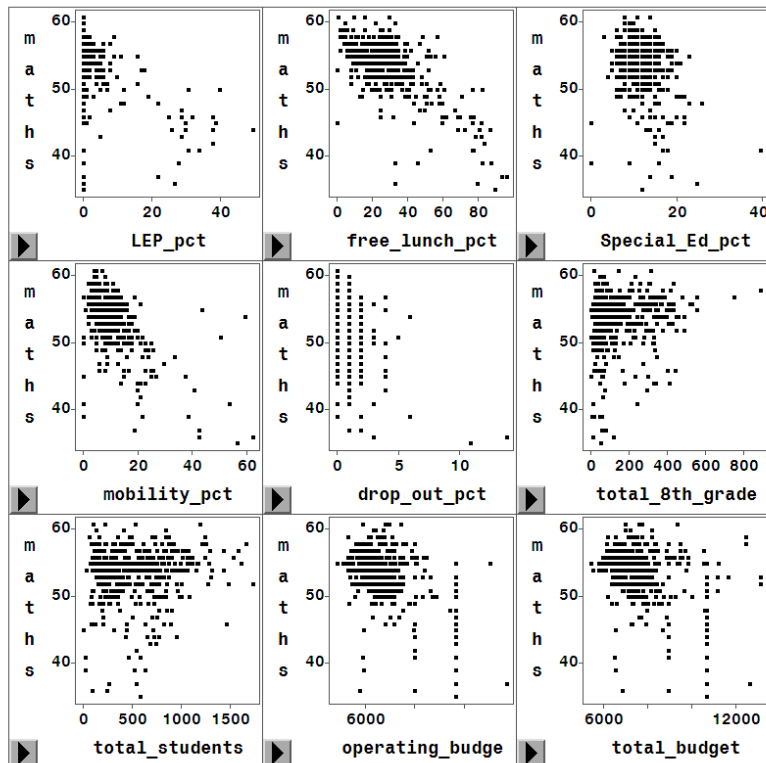
6

```
Proc Insight data=ph6470.grade8_mathscores;
```

```
scatter mathscore * LEP_pct free_lunch_pct  
Special_Ed_pct mobility_pct drop_out_pct  
total_8th_graders total_students operating_budget  
total_budget;
```

Produces “scatterplot matrix.”

7



8

Which variables should be log transformed?

9

To take logs of non-negative data in variable X:

IF $(X > 0)$ then $\log_X = \log(X)$; *log = natural log, ln*

ELSE IF $(X=0)$ then $\log_X = \log(X + \text{constant})$;

The constant depends on the smallest non-zero value of X, and range of X.

When the constant is small, almost the same results from

$\log_X = \log(X + \text{constant})$;

What happens when X is missing, for each method?

10

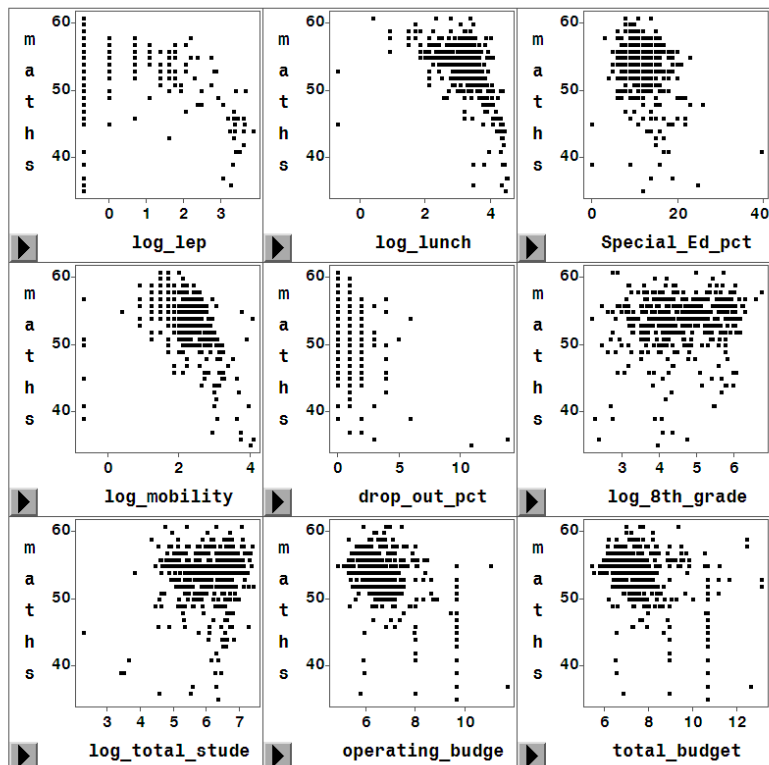
```

data schools;
  set ph6470.grade8_mathscores;

**** log transformations*****;
  log_8th_grade =log(total_8th_graders);
  if (LEP_pct > 0) then log_lep = log(LEP_pct);
    else log_lep = log(LEP_pct + .5);
  log_lunch = log(free_lunch_pct + .5);
  log_mobility = log(mobility_pct + .5);
  log_total_students = log(total_students + 10.0);

```

11



12

Next: check for horizontal outliers

Set aside horizontal outliers because:

- they may have large impact on regression coefficients
- regression model is almost extrapolating

13

```
data B; ***** trimmed data without outliers *****;
  set A;
  if (3 < Special_Ed_pct < 35);
  if (log_lunch > 0);
  if (log_mobility > 0);
  if (drop_out_pct < 10);
  if (log_total_students > 3);
  if (operating_budget < 10);
  if (total_budget < 12);
```

14

