

Stat 412/512

VARIABLE SELECTION

Feb 18 2015

Announcements

DA #2 released today



Quiz #3, material up to Monday,
postponed until next weekend.

To replicate or not?

If interactions are of interest, then replicate!

When experimental units are expensive, you can sometimes gain more by reducing variability, than increasing your replicates.

$$\text{SE of cell average} = \sigma / \sqrt{\text{number in cell}}$$

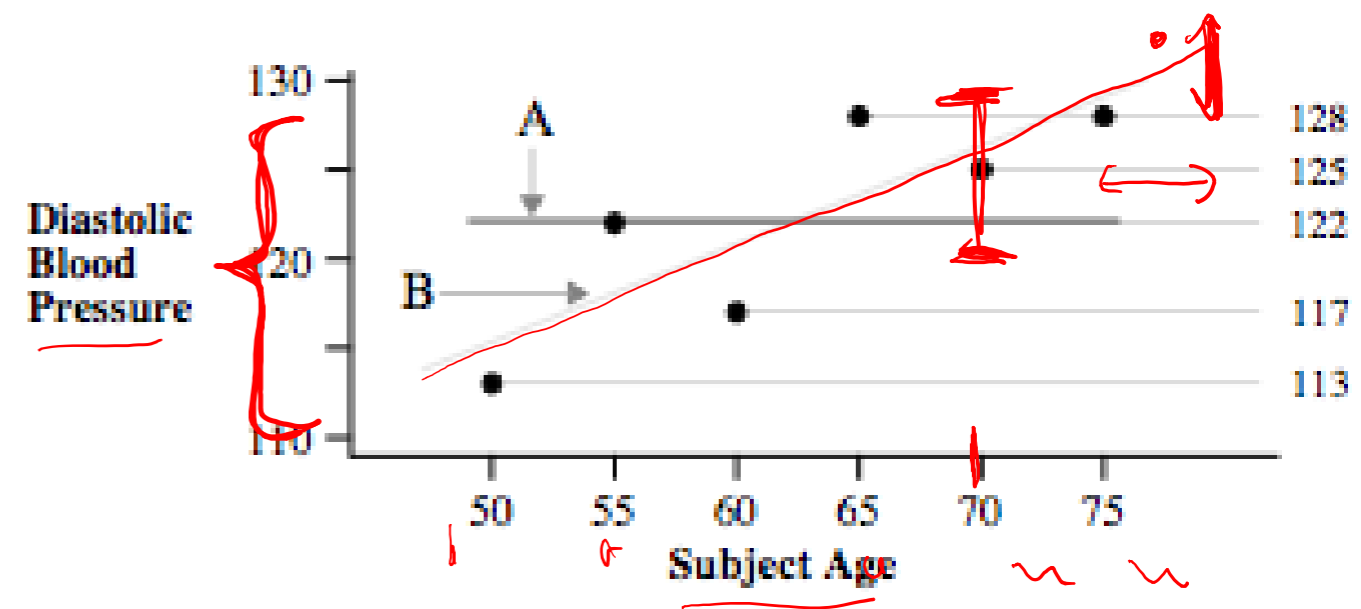
reduce this  or increase this 

Hypothetical example

New method for reducing high blood pressure.

Blood pressure tends to depend on age.

With no treatment the researcher expects something like ->



Option 1:

Ignore age, randomly assign treatment to six people aged 50 to 75. Six replicates, can make causal inferences. Expected variability = 6.1

Option 2:

Pick 6 people of the same age, randomly assign to treatment. Six replicates, can still make causal inferences (to a much reduced population). Expected variability = 3.8

Option 3:

Pick 6 people from 50-75 but pair them by similar ages, within each pair randomly assign to treatment (i.e. block by age). No replicates, can still make causal inferences. Expected variability ~ closer to 3.8

Lesson: Include important sources of variation in the design.

Identifying false replicates

a.k.a pseudo replication

The replication needs to be at the level of experimental unit (the items that are randomly assigned to treatment).

The replicates need to be **independent** applications of the same treatment.

Examples

Pygmalion study: platoon was randomized to treatment. It would be inappropriate to treat individual soldiers scores on the test as replicates.

Soybean study: chambers were randomized to treatment. It would be inappropriate to treat individual soybean plants as replicates.

in both cases we used the average within the experimental unit

our estimate of σ^2 tells us about the variability expected between experimental units

Experimental Design

ST513

ST513

read chapters 23 & 24

A good read:

Hurlbert, Stuart H. (1984). ["Pseudoreplication and the design of ecological field experiments"](#). *Ecological Monographs* (Ecological Society of America) **54** (2): 187–211. [doi:10.2307/1942661](https://doi.org/10.2307/1942661)

8

~~Variable selection~~

Model

We'll keep this at the high concept level.

Variable selection is the process of taking a large number of **explanatory variables** and **selecting** only a few to be in the regression model.

Red box = very important

Big concepts

There are different approaches.

We compare models with model selection criteria.

Generally, we consider a few good models, not just one “best model”.

Big problems

You can't trust inference after variable selection. **Why?**

*how to do inference
after model selection*

Model selection criteria are subject to variability too!

Unsolved

Legitimate uses of ~~variable selection~~

regression model

Adjusting for a large set of explanatory variables

You have a large number of variables you want to account for, but they are not of direct interest (you will not look at their p-values, or estimate their effects). Do variable selection on just these variables.

Prediction

You want a simple model purely to predict mean response, you will not interpret p-values or estimates (or makes statement like “x is an important predictor”).

Illegitimate uses of variable selection

Fishing for explanation

Which variables are important?

Variable selection will not uncover some "true" model. The best model in one sample, won't often be the best in another. Often there are many useful models.

Interpretation of included variables is dangerous:

inclusion depends on what other variables are being considered (particularly if they are correlated)

p-values are biased low, and estimates are biased high
(in magnitude)

Two examples

case1201 SAT: useful for illustrating methods (i.e. not too many variables).

Average SAT score for all states. Not a great measure because not every one takes the SATs (in some states only the best students take them).

case1202 Discrimination: interesting case.

After accounting for qualifications and experience, do women start on lower salaries?

Average SAT scores by US State in 1982, and possible associated factors

	State	SAT	Takers	Income	Years	Public	Expend	Rank
1	Iowa	1088	3%	326	16.79	87.8	25.60	89.7
2	South Dakota	1075	2	264	16.07	86.2	19.95	90.6
3	North Dakota	1068	3	317	16.57	88.3	20.62	89.8
4	Kansas	1045	5	338	16.30	83.9	27.14	86.3
5	Nebraska	1045	5	293	17.25	83.6	21.05	88.5
6	Montana	1033	8	263	15.91	93.7	29.48	86.4
7	Minnesota	1028	7	343	17.41	78.3	24.84	83.4
8	Utah	1022	4	333	16.57	75.2	17.42	85.9
9	Wyoming	1017	5	328	16.01	97.0	25.96	87.5

% of eligible students that take the SATs

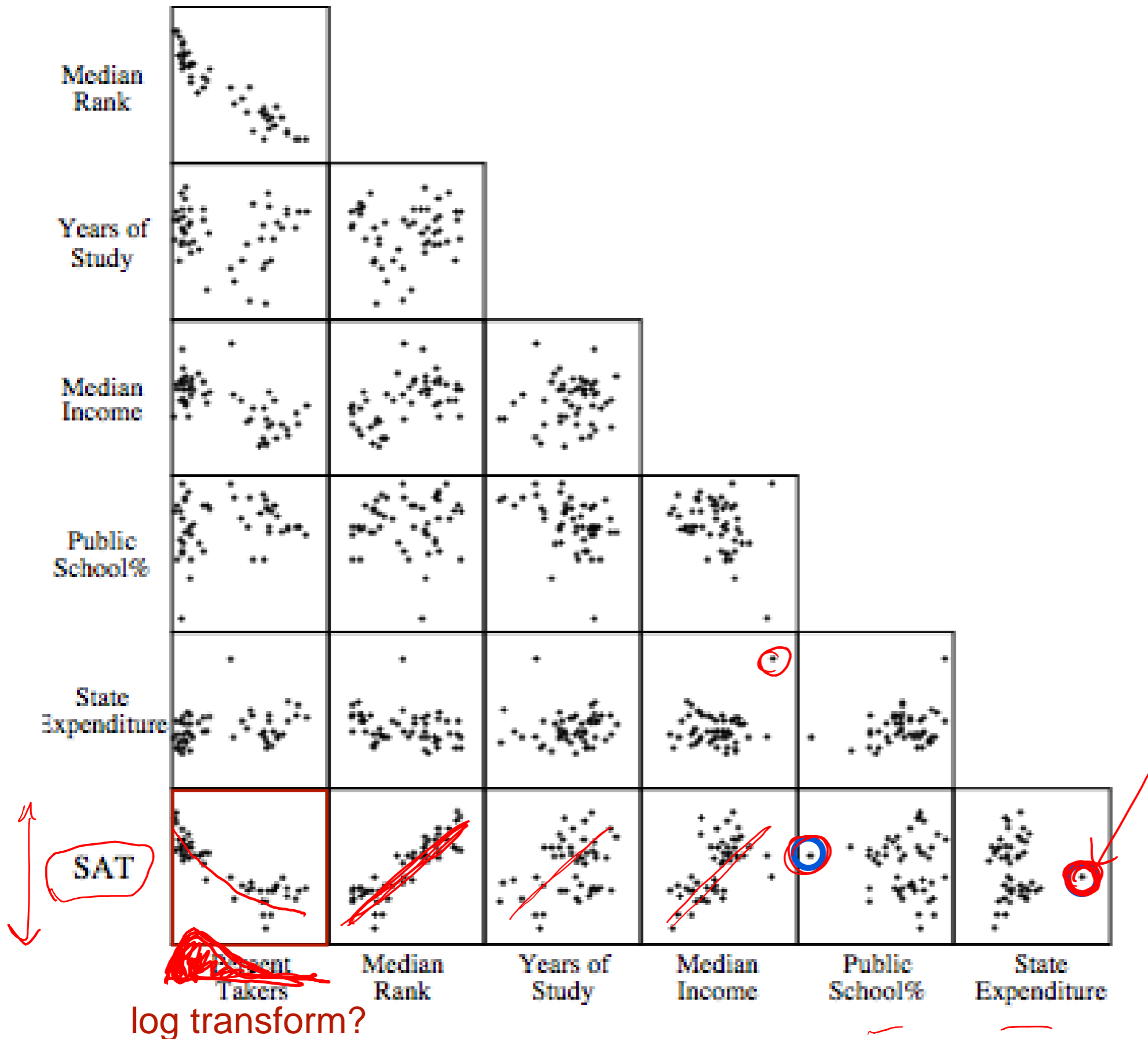
average class rank of students that take the SATs

45	Texas	868	32	303	14.95	91.7	19.55	76.4
46	Indiana	860	48	258	14.39	90.2	17.93	74.1
47	Hawaii	857	47	277	16.40	67.6	21.21	69.9
48	North Carolina	827	47	224	15.31	92.8	19.92	75.3
49	Georgia	823	51	250	15.55	86.5	16.52	74.0
50	South Carolina	790	48%	214	15.42	88.1	15.60	74.0



case1201 SAT

Matrix of scatterplots for SAT scores and six explanatory variables

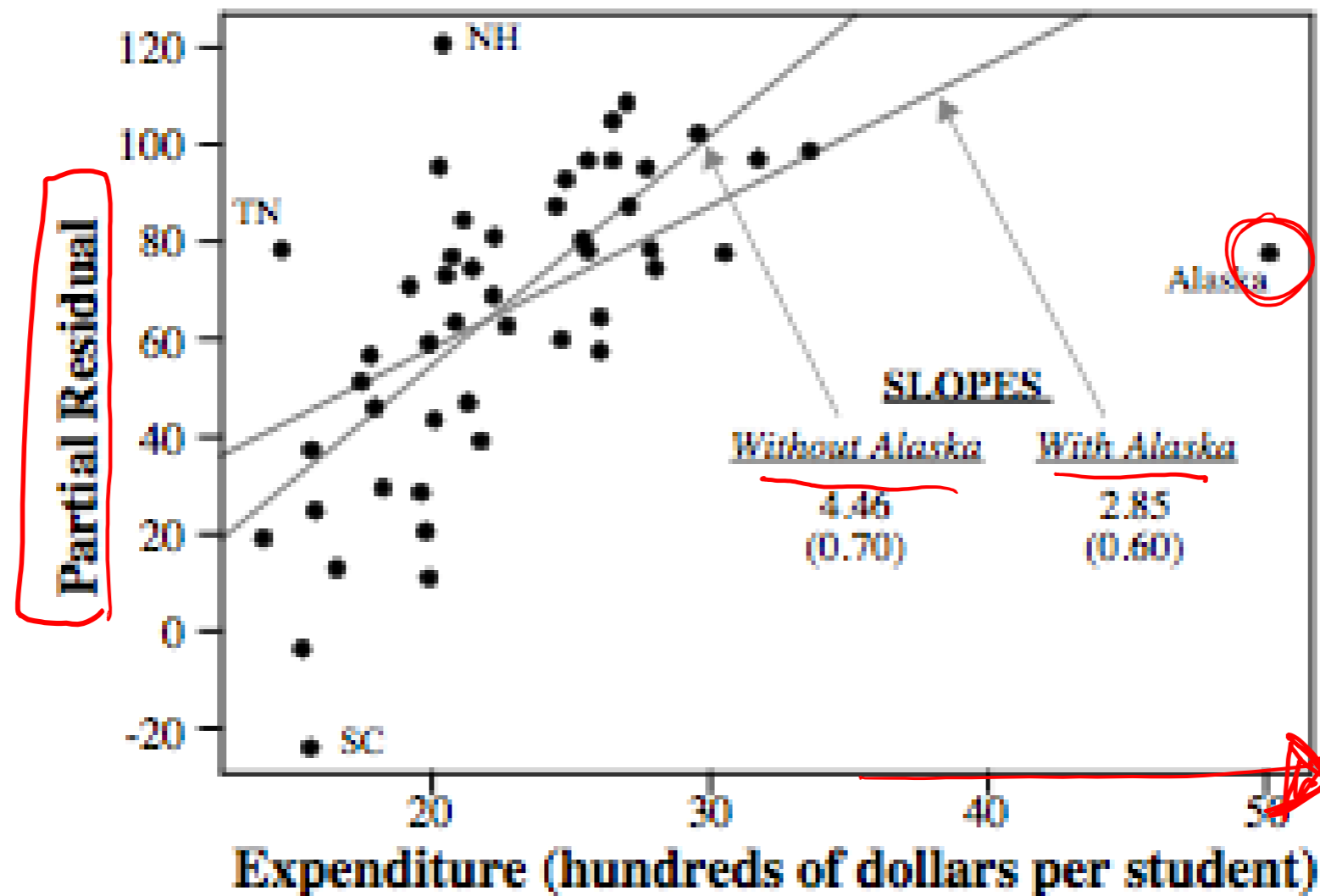


tentative model: $\mu\{ \text{SAT} \mid \dots \} = \text{log takers} + \text{rank}$

Display 12.5

p. 349

Partial residual plot of state average SAT scores (adjusted for percent of students in the state who took the test and for median class rank of the students who took the test) versus state expenditure on secondary education



Historically popular

Stepwise methods

add or remove a variable one at a time

Only looks at a subset of all possible models.

Things that can be controlled:

- starting point, path through the models
- choice of next "best" step, and stopping point
we'll use F-tests

computationally quick, but no guarantee any two approaches will arrive at the same model

Stepwise methods

Forward selection

Start with an intercept term. Test each term for inclusion, include the "best" one. Repeat until no term passes our threshold.

smallest p-value from F-test

Backward elimination

Start with a full model. Test each term for deletion, delete the "worst" one. Repeat until no term fails our threshold.

Stepwise selection

biggest p-value from F-test

Start with a constant mean model. One step of forward, then one step of backward and repeat.



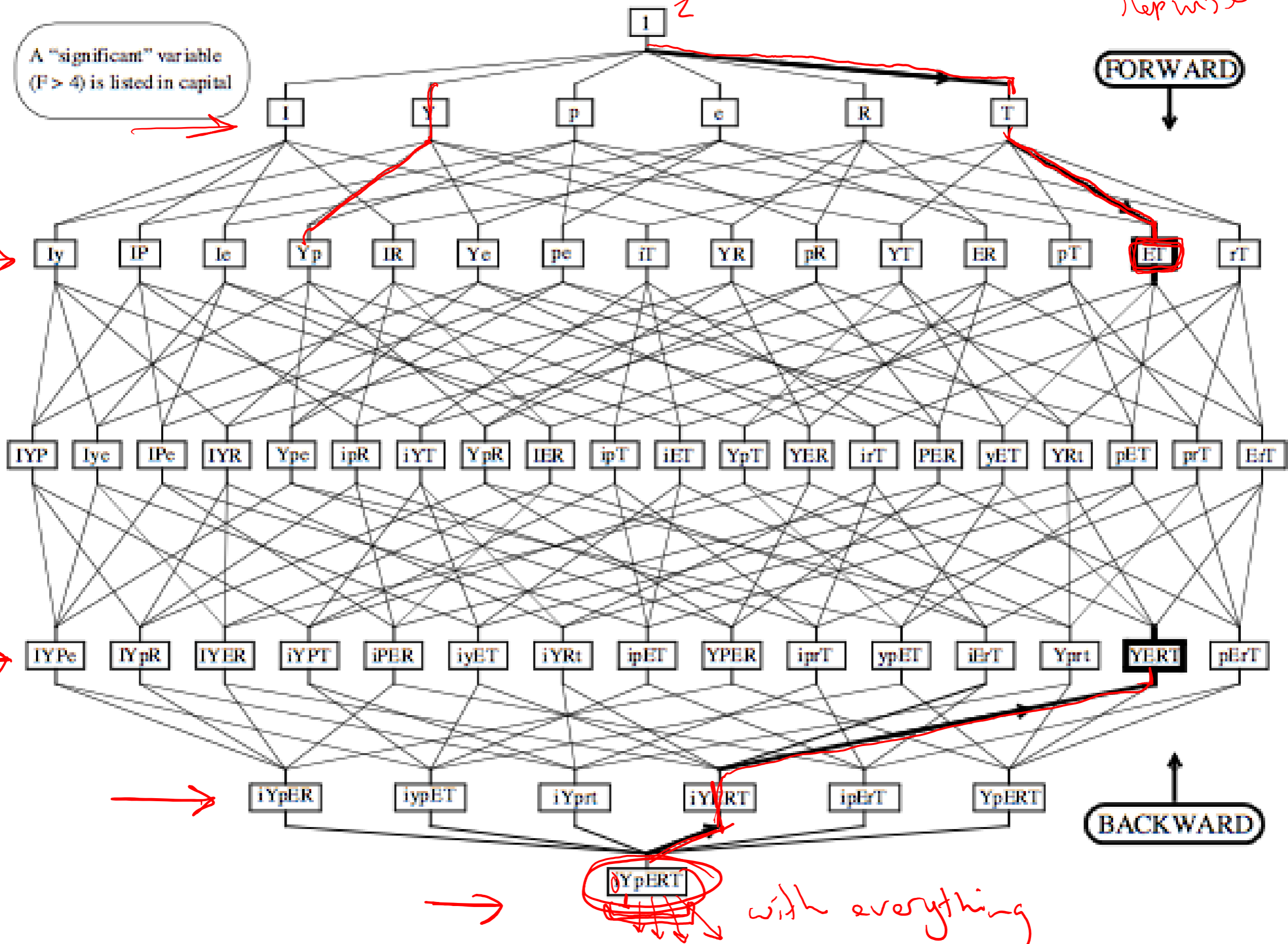
A "significant" variable (F > 4) is listed in capital

constant

stepwise

FORWARD

BACKWARD



with everything

All subsets

Look at all possible models.

Then judge them on some measure of fit.

Generally learn the most by looking at a few good models.

Measures of fit

If the number of parameters are the same, we prefer the model with smaller residual sum of squares (RSS).

If the number of parameters are different, we want to balance smaller RSS with fewer parameters.

remember RSS always gets smaller if you add another parameter

big is bad!

3 common
model
selection
criteria

Cp

BIC

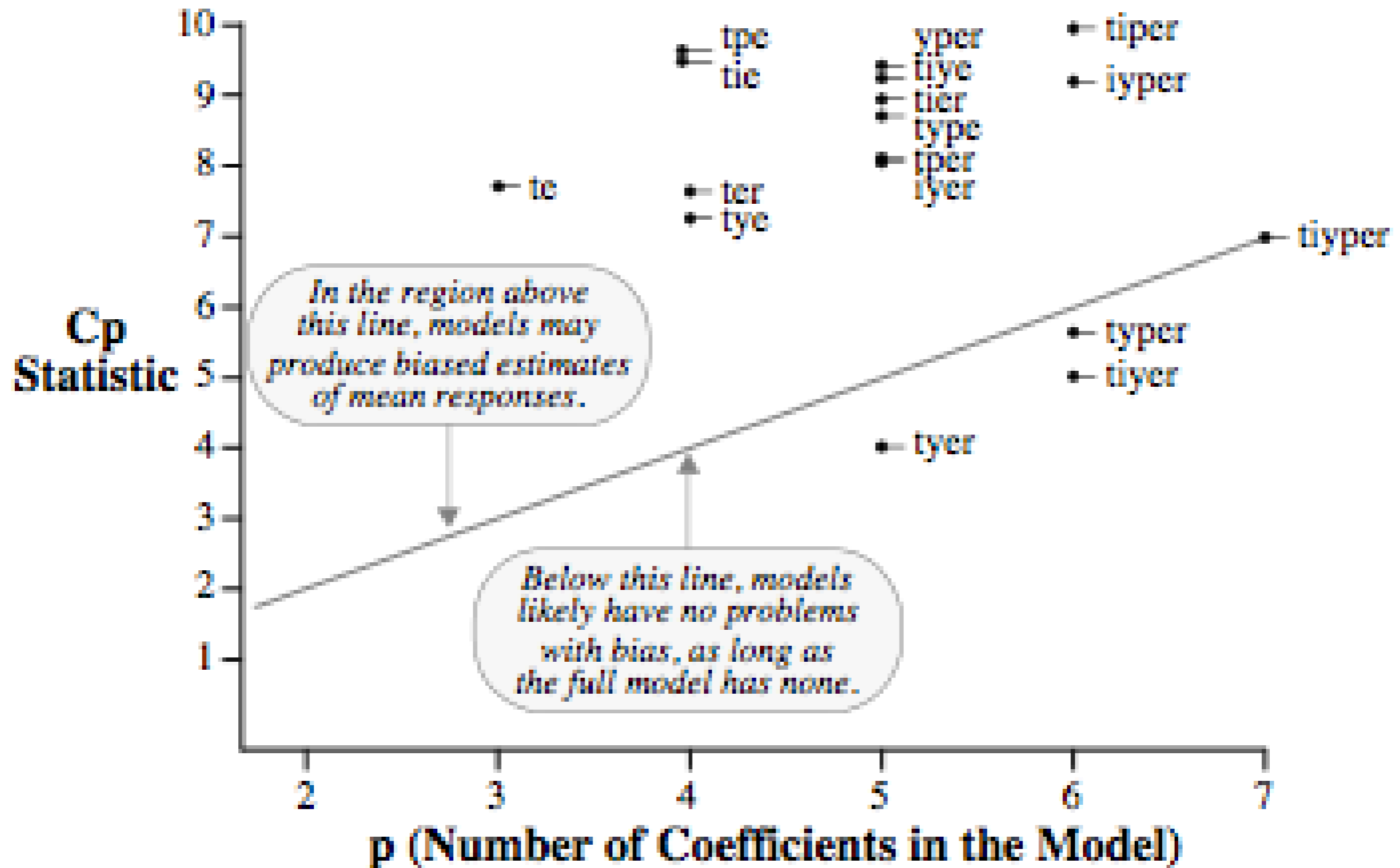
AIC

$$\begin{array}{l}
 \frac{SSRes}{\hat{\sigma}_{full}^2} - n + p \\
 n \times \log \left(\frac{SSRes}{n} \right) + \log(n) \times (p + 1) \\
 n \times \log \left(\frac{SSRes}{n} \right) + 2 \times (p + 1)
 \end{array}$$

big if there are
lots of parameters

big if RSS is big

Cp plot for State SAT averages (showing only those models with $C_p < 10$); t = log takers, i = income, y = years, p = public, e = expend, and r = rank



You can't trust inference after variable selection.

Why?

We choose variables to be in the model if, in our data, they show some power to explain the response.

If a variable appears in our final model, it has by construction, shown some power to explain the response.

It doesn't then make sense to ask if the variable is significant...we'll get a small p-value because we only selected variables that gave low p-values!

Lab: with explanatory variables generated to have **absolutely no relationship** to the response, the best model selected by model selection has very small p-values!

Methods we won't talk about

but can be useful

Principal component based methods

Penalized methods (ridge, lasso, lars)