Chapter 12: Strategies for Variable Selection

Class 19, 4/15/09 W

Slide 1 Chapter 12: Strategies for Variable Selection

NOTES:

HW 12 due Friday 4/17/09

Submit as Myname-HW12.doc (or *.rtf)

- HW 12 10.28: El Niño and Hurricanes
  - Due Friday 4/17/09 Noon
- HW 13 Cammen's ingestion rate data. Note that this was a 2003 final exam problem
  - Read Cammen (1980) & evaluate his regression model
  - Due Friday 4/25/09 Noon This problem will count double!
- Read Chapter 12: Selection of variables
- Run my overfitting syntax: overfitting.sps
- Read Campbell & Kenney Chapters 4 & 5 on the regression artefact and gender inequities
  - Run my Campbell & Kenny syntax: RTMCK.sps
- No Class Monday 4/20: Patriot's Day

Slide 2 HW 12 due Friday 4/17/09

NOTES:

HW12: Cammen model

Cammen (1980) compiled data from the literature on the ingestion rates of 22 deposit feeders. Deposit feeders are organisms that live in mud and sand and ingest mud and sand. Deposit feeders use the organic matter in the mud and sand for growth. Table 1 shows the species from the literature, their ingestion rates, the fraction organic matter in sediment, and the body weights of individual deposit feeders. Cammen (1980) used regression to estimate the ingestion rate of deposit feeders **(ING)** (mg dry weight/day) using the fraction organic matter in the sediment (**OM**) and body weight of the deposit feeder (**WT**). He regressed $\log_{10}$ (**ING**) as the response variable with two explanatory variables $\log_{10}$ (**WT**) and $\log_{10}$ (**OM**). He deleted the three bivalves from his analyses because they appeared to be outliers, and based his regressions on the 19 non-bivalve species.

Slide 3 HW12: Cammen model

NOTES:

Table 1. Data from Cammen (1980). Loaded on Prometheus as cammen.sav, in case you wanted to examine the data (optional). The last 3 highlighted species are bivalve molluscs (indicated under Taxon) WT is the body weight of the deposit feeder (dry weight of the animal in milligrams). ING is the ingestion rate in mg dry weight/day. Cammen scaled the ingestion rates to account for temperature effects (higher ingestion at higher temperatures). OM is the organic matter content ((weight organic matter)/(total sediment dry weight), expressed as %).

	Species	Taxon	WT	ING	OM
1	*Hydrobia neglecta*	Gastropod mollusc	0.2	0.57	18
2	*Hydrobia ventrosa*	Gastropod mollusc	0.2	0.86	17
3	*Tubifex tubifex*	Oligochaete (annelid)	0.27	0.45	29.7
4	*Hyalella azteca*	Crustacean	0.32	0.45	50
5	*Potamopyrgus jenkinsi*	Gastropod mollusc	0.46	2.7	14.4
6	*Hydrobia ulvae*	Gastropod mollusc	0.9	0.67	13
7	*Nereis succinea*	Polychaete (annelid)	5.8	20.2	6.8
8	*Pteronarcys scotti*	Crustacean (stonefly)	8.4	1.49	93
9	*Orchestia grillus*	Crustacean	12.4	4.4	88
10	*Arenicola grubii*	Polychaete (annelid)	20.4	240	2.2
11	*Thoracophelia mucronata*	Polychaete (annelid)	40	230	1
12	*Ilyoplax pusilla*	Crustacean	53	300	4.2
13	*Uca pugnax*	Crustacean	63.3	19.9	51
14	*Scopimera globosa*	Crustacean	65	50	23.6
15	*Pectinaria gouldii*	Polychaete (annelid)	80	1667	0.7
16	*Abarenicola pacifica*	Polychaete (annelid)	380	3400	1.2
17	*Abarenicola claparedi*	Polychaete (annelid)	380	9400	0.4
18	*Arenicola marina*	Polychaete (annelid)	980	4700	0.64
19	*Macrophthalmus japonicus*	Crustacean	2000	4680	2.1
20	***Macoma balthica***	Bivalve mollusc	5.1	4.49	20
21	***Portlandia arctica***	Bivalve mollusc	19.9	3.58	8.8
22	***Scrobicularia plana***	Bivalve mollusc	380	43	3.4

Slide 4

NOTES:

- Was Cammen (1980) justified in dropping the three bivalve molluscs from his regression equation?
  - Consider both the case-wise diagnostic tests (residuals vs. predicted values, Cook's D, studentized residuals, and leverage values), and the results of fitting bivalves as a dummy variable.
  - Discuss the problems in using Cook's D, leverage, and studentized residuals in detecting outliers when more than one datum may be an outlier.
  - There is no strictly right or wrong answer to this question, but you must justify your choice with evidence from the regression analyses.
- There were 5 groups of animals in Cammen's data. Is there evidence that the ingestion rates as a function of weight and organic matter differ among these 5 groups?
- Based on your analyses, produce a graph showing the relationship between ingestion rate, body weight and organic matter.
- Write the regression equation expressing the relationship between ingestion rate, organic matter, and body weight. Pay attention to significant figures, and include an estimate of the standard error of the coefficients.
- If you found that the animal groups differed in ingestion rate, your final graphs and model should reflect this full model

Slide 5

NOTES:

Homework Presentations

- William Walker for HW 8,
- Steven Kichefski for HW 9 and
- Lisa Greber for HW10

Slide 6 Homework Presentations

NOTES:

Chapter 12: Strategies for variable selection

Slide 7 Chapter 12: Strategies for variable selection

NOTES:

Using multiple regression to test causal models

Being in politics is like being a football coach. You have to be smart enough to understand the game and dumb enough to think it's important.-- Eugene McCarthy

Application to Regression & Chapter 12
To use multiple regression to test causal models, you have to know enough statistics to run the analysis, but you have to be dumb enough to think the approach is valid

Slide 8 Using multiple regression to test causal models

NOTES:

Regression errors & artifacts

- A) Covariates are often necessary
  - Fluoride & cancer (Manly 1992)
  - Storks & babies
- B) Multicollinearity:
  - Interpreting Beta signs as effects when the magnitude and sign of Beta is a function of other variables in the equation
  - Handguns & Crime rates (Lott & Mustard vs. Ayers & Donahue)
  - Peterson on school vouchers & test scores
- C) The regression artifact and improper interpretation of the effects of covariates
  - Math ability & gender
  - The Bell Curve

Slide 9 Regression errors & artifacts

NOTES:

Does fluoride cause cancer?

Manly (1992) The design & analysis of research studies

- **Yiamouyiannis & Burk 1977**
  - Fluoridation began in 1952-1956
  - Fluoridated and non-fluoridated cities matched by population size
    - 10 largest non-fluoridated cities
    - Fluoridated cities of comparable size

Table 1.2. *Cancer deaths per 100 000 population in fluoridated and non-fluoridated cities in the United States (Yiamouyiannis and Burk, 1977)*

	Fluoridated cities	Non-fluoridated cities
1950	181	179
1970	217	197
Change	+36	+18

Why ?

Slide 10 Does fluoride cause cancer?

NOTES:

Cancer & Fluoride

Manly (1992) The design & analysis of research studies

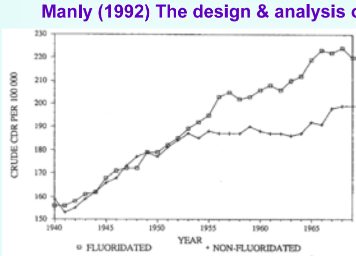


Figure 1.2. Crude cancer death rates per 100 000 of population for ten fluoridated and ten non-fluoridated cities in the United States, 1940–69. Fluoridation of cities took place over the period 1952–56.

Why ?

Yiamouyiannis & Burk 1977: 10 largest non-fl. Cities and **matched** fluorodated cities

Fluoridation took place from 1952-1956

Rebuttal: Oldham & Newell (1977) Applied Statistics

Slide 11 Cancer & Fluoride

NOTES:

Guidelines for predictive modeling

From Holmes' Causal modeling (Sage)

- Theorize before analyzing data or validate theory with additional data
- Formulate explicitly ordered hypotheses
- Measure covariation with an appropriate technique
- Examine measures of association to see if they are significant
- Reject competing models that are more complex or less based on theory
- Reject models that have “bad fit”

Slide 12 Guidelines for predictive modeling

NOTES:

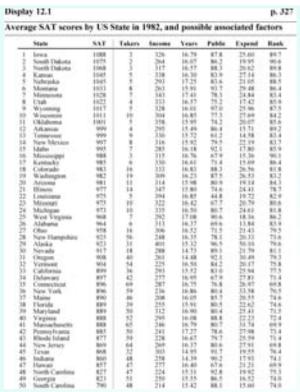
Gallagher's addenda

From Harrell & Campbell & Kenney

- Don't use multiple regression to infer causation. When more than one variable is in the model, the sign and magnitude of the coefficients for an explanatory variable often depend on the value of other variables in the equation
- Don't use stepwise or other automated selection procedures
- Beware the regression artifact and control for it
  - Use repeated measures designs, structural equation models or corrections for the regression artifact.
  - Or, design a controlled experiment to properly assess the effect

Slide 13 Gallagher's addenda

NOTES:



Case Study 12.1
SAT Scores

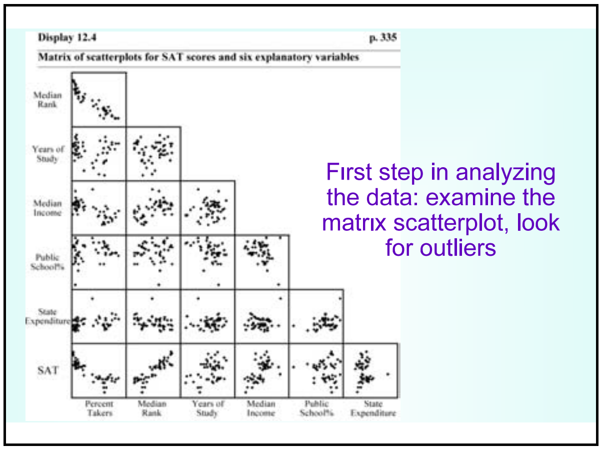
Slide 14

NOTES:

Case Study 12.1
Final model
SAT Scores = f(% Taking exam, median class rank) or (% taking exam, rank and expenditure)
Result: NH is #1, Massachusetts is 11 or 32 (expenditure)

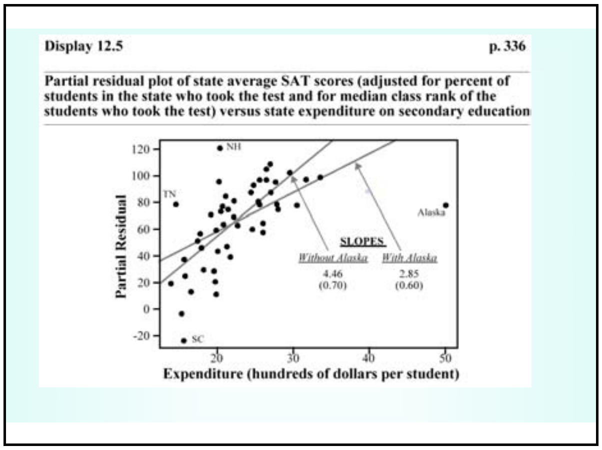
Slide 15

NOTES:



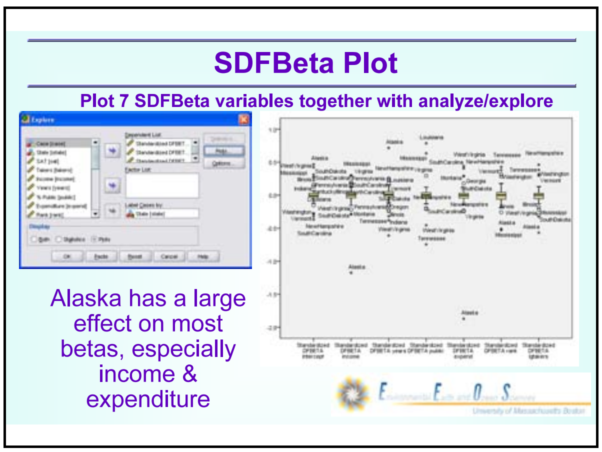
Slide 16

NOTES:



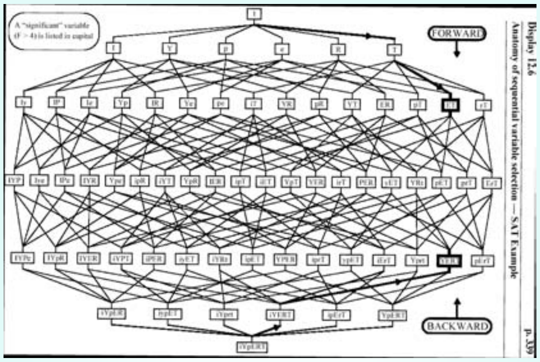
Slide 17

NOTES:



Slide 18 SDFBeta Plot

NOTES:



Slide 19

NOTES:

All possible regression models

R^2 and Adjusted R^2 choose models with TOO many parameters

- Mallow's Cp
  - Cp=p+(n-p) (σ^2-σ^2_full)/σ^2_full
- Akaike information content
- Schwarz Bayesian Information Content (BIC)
  - Ln(log(σ^2)+p log(n).
  - Can be used to calculate posterior probabilities
- Neither available in SPSS without syntax
  - All are available in SPSS syntax
    - \Statistics SELECTION
  - The BIC in SPSS is different from Sleuth

Slide 20 All possible regression models

NOTES:

SPSS selection criteria

In syntax only: \Statistics SELECTION

Model Summary[d]

					Change Statistics					Selection Criteria			
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	.385[a]	.148	.056	1.32344	.148	1.610	4	37	.192	28.216	1.082	111.619	36.905
2	.729[b]	.532	.452	1.00879	.384	14.341	2	35	.000	7.077	.655	50.935	19.241
3	.894[c]	.798	.757	.67174	.266	44.935	1	34	.000	-26.298	.297	8.000	-12.396

a. Predictors: (Constant), ln (Precipitation), ln (Area), ln (Runoff), ln (Discharge)

b. Predictors: (Constant), ln (Precipitation), ln (Area), ln (Runoff), ln (Discharge), ln (Deposition), ln (NO_3 precipitation)

c. Predictors: (Constant), ln (Precipitation), ln (Area), ln (Runoff), ln (Discharge), ln (Deposition), ln (NO_3 precipitation), ln (Density)

d. Dependent Variable: ln (NO_3)

Slide 21 SPSS selection criteria

NOTES:

Bayes' Theorem

Larsen & Marx 2nd Edition (2001)

Bayes' Theorem (Theorem 2.6.2 p. 65)

Let $\{A_i\}_{i=1}^{n}$ *be a set of n events, each with positive probability that partitions S in such a way that*

$$\bigcup_{i=1}^{n} A_i = S$$

and $A_i \cap A_j = \varnothing$ *for* $i \neq j$.

For any event B (also defined on S), where $P(B) > 0$,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

for any $1 \le j \le n$.

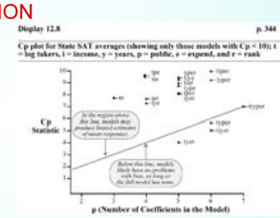
Slide 22 Bayes' Theorem

NOTES:

All possible regressions

All regression models in SAS, R & Matlab, not SPSS

- SAS procedure
- SPSS
  - /STATISTICS COEFF OUTS CI R ANOVA **COLLIN** TOL CHANGE SELECTION
- Matlab
  - Stixbox



Slide 23 All possible regressions

NOTES:

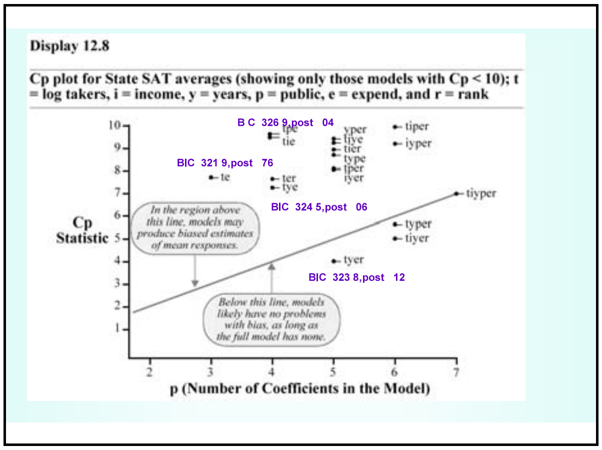
SPSS regression syntax

/STATISTICS ALL or /STATISTICS SELECTION

```
* Case 1201- note the /STATISTICS=SELECTION.
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/SELECT= istate NE 2
/MISSING LISTWISE
/STATISTICS ALL
/CRITERIA=PIN(.05) POUT(.10) CIN(95)
/NOORIGIN
/DEPENDENT sat
/METHOD=BACKWARD lgtakers income years public expend rank
/PARTIALPLOT ALL
/SCATTERPLOT=(*ZRESID ,*ZPRED )
/RESIDUALS ID( state )
/SAVE PRED COOK MCIN ICIN RESID .
```

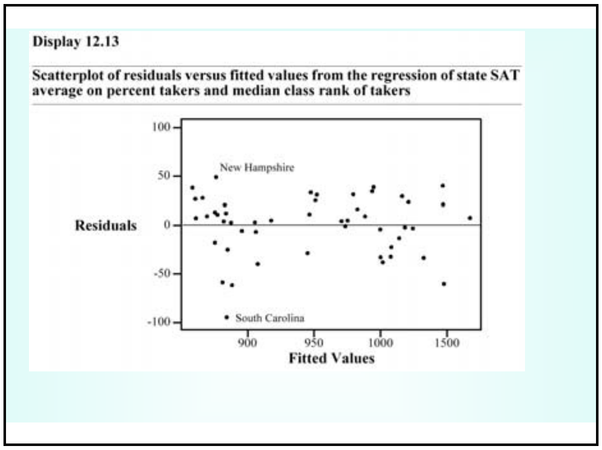
Slide 24 SPSS regression syntax

NOTES:



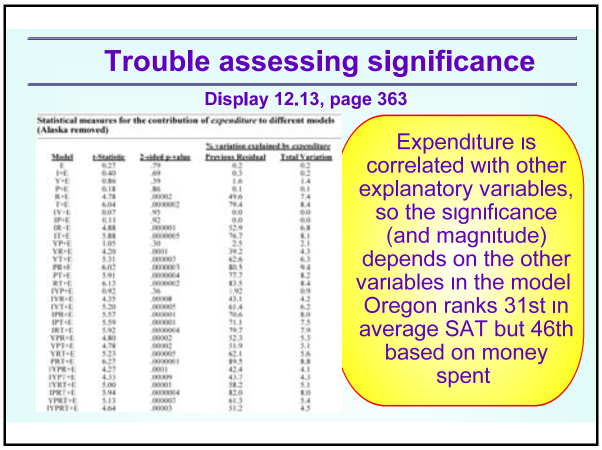
Slide 25

NOTES:



Slide 26

NOTES:



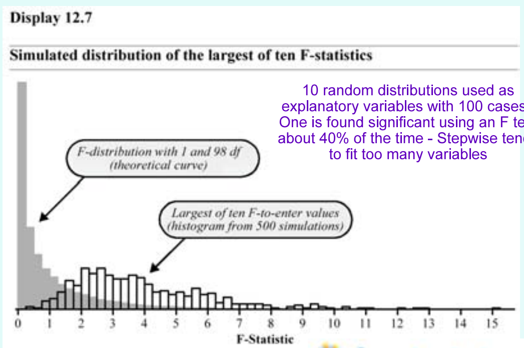
Slide 27 Trouble assessing significance

NOTES:

Overfitting: why stepwise procedures should not be used to estimate p values.

Slide 28 Overfitting: why stepwise procedures should not be used to estimate p values.

NOTES:



Slide 29

NOTES:

Gallagher's overfitting.sps

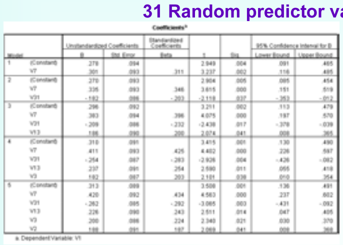
```
* Overfitting simulation, inspired by
* Nontechnical Introduction to Overfitting in Regression-Type Models, Babyak (2004).
* Michael A Babyak  What You See May Not Be What You Get: A Brief, Nontechnical
*  Introduction to Overfitting in Regression-Type Models.
* Psychosom Med 2004 66: 411-421.
* Written by E Gallagher, revised 4/12/05.
* Generate 100 cases, with 32 normally distributed variates.
new file.
input program.
loop #i = 1 to 100.
COMPUTE V1 = RV.normal (0,1) .
COMPUTE V2 = RV.normal (0,1) .
...
COMPUTE V32 = RV.normal (0,1) .
end case.
end loop.
end file.
end input program.
formats V1 to V32 (f4.2).
exe.
```

Slide 30 Gallagher's overfitting.sps

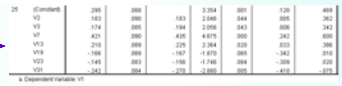
NOTES:

Results of Stepwise Selection

31 Random predictor variables



Backward (added V23, V19)



Slide 31 Results of Stepwise Selection

NOTES:

Harrell (2002, p. 56-57) on stepwise

Harrell's conclusion: Don't use stepwise!

- It yields R^2 values that are biased high
- F and χ^2 distributions don't have their claimed distributions
- SE of regression coefficients are biased low and CI's and predicted values that are falsely narrow
- P-values too small
- Regression coefficients biased high in absolute value and need shrinkage.
- Rather than solving the problem of collinearity, variable selection is made arbitrary by collinearity
- It allows us not to think about the problem

Slide 32 Harrell (2002, p. 56-57) on stepwise

NOTES:

Multicollinearity, collinearity

- If the explanatory variables are strongly correlated
  - The regression coefficient estimates have a huge variance
  - They can change in sign and significance with a slight change in the data, bouncing betas
- Diagnostics (Variance Inflation Factors (VIF's) - see next page)
- Solutions for multicolinearity for OLS regression
  - Reduce the number of explanatory variables using theory & insight into the field
  - Ridge regression
  - Principal components regression

Slide 33 Multicollinearity, collinearity

NOTES:

Collinearity [multicollinearity]

- When one or more predictors can be predicted by other predictors, the standard error of the regression coefficients can be inflated and the corresponding tests have reduced power
- Assessed with Variance inflation factors (VIF) or tolerance
  - $VIF_i = 1\ (1 - R^2_i)$, where R^2_i is the squared multiple correlation coefficient between explanatory variable 'i' and the other explanatory variables
  - Neter et al. (1996): VIF's > 10 are cause for concern (but smaller VIF's can also be a problem)

Slide 34 Collinearity [multicollinearity]

NOTES:

Ways of detecting multicollinearity

Marayuma (1998, p. 64)

- When the variance (standard errors) of beta weights is large
- When signs on beta weights are inappropriate [e.g., larger classes ⇨ higher test scores]
- When regression weights and signs change radically upon the addition or removal of single variables
- When the Variance Inflation Factor is high (VIF> 6 or 7 as a very rough rule)
- When simple correlations are > 0.8-0.9
- When correlations among predictor variables > R^2 for response with all predictor variables

Slide 35 Ways of detecting multicollinearity

NOTES:

Shooting Down the "More Guns, Less Crime" Hypothesis

Ian Ayres* & John J. Donohue III

crime from the Table 3 regressions, one finds utterly bizarre results. For example, an increase of one percentage point in the percentage of black males aged 30-39 would be expected to almost double the violent crime rate, while a similar increase in the percentage of black males aged 40-49 would lead to a drop in violent crime of 60%. Similarly, increasing the percentage of black males aged 50-64 would cause violent crime to jump by 145%, but increasing the percentage of black males over age 65 would lead to a 78% decline in violent crime. These nonsense results prevent us from understanding why the demographic controls can influence the estimates of shall-issue adoption so strongly.

Adding too many covariates can destroy a regression

Slide 36

NOTES:

Solutions to multicollinearity

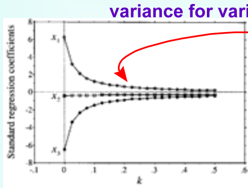
- If the goal of the model is to produce predicted values for one analysis, then multicollinearity is **not** a problem. All variables can be included.
  - However, if the equation is to be used for new data, then the model will be badly overfitted, the predicted values will be biased
  - Significant coefficients could be spurious or nonsense
- Solutions
  - Variable selection procedures (cluster analysis of variables)
  - Principal components regression
    - Use principal component scores as explanatory variables
    - Principal component scores are orthogonal (uncorrelated)
  - Ridge regression
  - Structural equation modeling

Slide 37 Solutions to multicollinearity

NOTES:

Ridge regression

Available as a macro in SPSS, LISREL (not AMOS); increase variance for variables not covariance



k	R^2
0.0	0.346
0.1	0.275
0.2	0.257
0.3	0.249
0.4	0.245
0.5	0.243

Figure 10.8 'Ridge trace' diagram showing the estimates of the standardized regression coefficients fo explanatory variables x_1 to x_3 as a function of k. Table: decrease of R^2 as a function of k.

A ridge regression parameter, k, is chosen using the ridge trace diagram(k=0.2 in the above example [the base of the horn] from Draper & Smith) that 'shrinks' the regression coefficients, especially those coefficients (Beta's) that are strongly correlated. This offers a partial solution to the problem of collinearity.

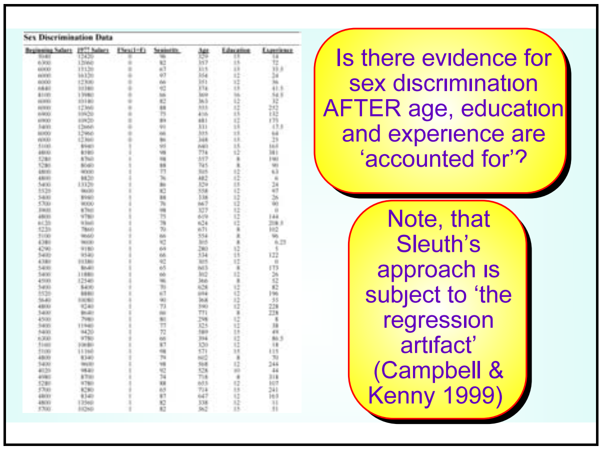
Slide 38 Ridge regression

NOTES:

Case 11.2 Gender discrimination

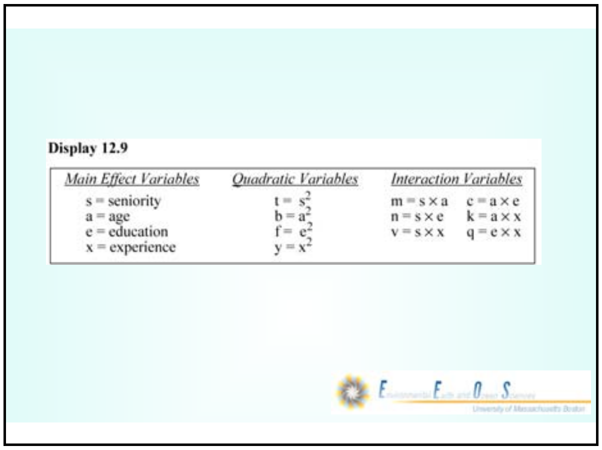
Slide 39 Case 11.2 Gender discrimination

NOTES:



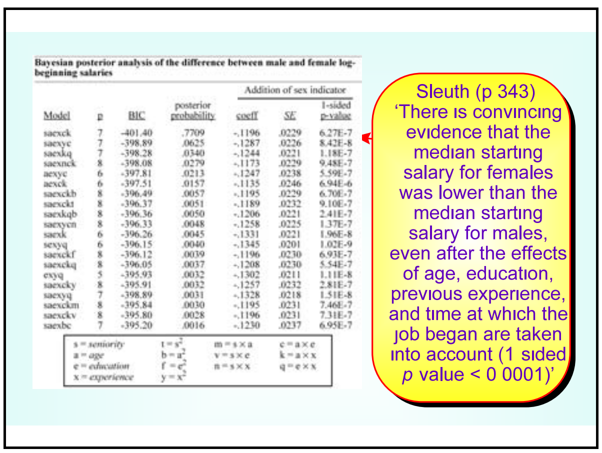
Slide 40

NOTES:



Slide 41

NOTES:



Bayesian posterior analysis of the difference between male and female log-beginning salaries

Model	p	BIC	posterior probability	Addition of sex indicator: coeff	SE	1-sided p-value
saexck	7	-401.40	.7709	-.1196	.0229	6.27E-7
saexyc	7	-398.89	.0625	-.1287	.0226	8.42E-8
saexkq	7	-398.28	.0340	-.1244	.0221	1.18E-7
saexnck	8	-398.08	.0279	-.1173	.0229	9.48E-7
aexyc	6	-397.81	.0213	-.1247	.0238	5.59E-7
aexck	6	-397.51	.0157	-.1135	.0246	6.94E-6
saexckb	8	-396.49	.0057	-.1195	.0229	6.70E-7
saexckt	8	-396.37	.0051	-.1189	.0232	9.10E-7
saexkqb	8	-396.36	.0050	-.1206	.0221	2.41E-7
saexycn	8	-396.33	.0048	-.1258	.0225	1.37E-7
saexk	6	-396.26	.0045	-.1331	.0221	1.96E-8
sexyq	6	-396.15	.0040	-.1345	.0201	1.02E-9
saexckf	8	-396.12	.0039	-.1196	.0230	6.93E-7
saexckq	8	-396.05	.0037	-.1208	.0230	5.54E-7
exyq	5	-395.93	.0032	-.1302	.0211	1.11E-8
saexcky	8	-395.91	.0032	-.1257	.0232	2.81E-7
saexyq	7	-398.89	.0031	-.1328	.0218	1.51E-8
saexckm	8	-395.84	.0030	-.1195	.0231	7.46E-7
saexckv	8	-395.80	.0028	-.1196	.0231	7.31E-7
saexbc	7	-395.20	.0016	-.1230	.0237	6.95E-7

Slide 42

NOTES:

SPSS output using forward, backward or stepwise

Model Summary[h]

Model	Selection Criteria: Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	-395.813[a]	.858	38.600	-390.747
2	-407.042[b]	.761	23.681	-399.444
3	-410.713[c]	.731	19.134	-400.582
4	-415.957[d]	.691	13.330	-403.294
5	-419.552[e]	.665	9.706	-404.356
6	-421.539[f]	.651	7.718	-408.876
7	-427.248[g]	.612	2.501	-412.053

a. Predictors: (Constant), f (e^2)
b. Predictors: (Constant), f (e^2), n (s * e)
c. Predictors: (Constant), f (e^2), n (s * e), v (s * x)
d. Predictors: (Constant), f (e^2), n (s * e), v (s * x), k (a*x)
e. Predictors: (Constant), f (e^2), n (s * e), v (s * x), k (a*x), x (Experience)
f. Predictors: (Constant), f (e^2), n (s * e), k (a*x), x (Experience)
g. Predictors: (Constant), f (e^2), n (s * e), k (a*x), x (Experience), q (e*x)
h. Dependent Variable: ln (Salary)

Slide 43 SPSS output using forward, backward or stepwise

NOTES:

Has gender equity really been rejected?

Campbell & Kenny: statistical equating often produces gender discrimination when there is none, and racial differences when there are none

Slide 44 Has gender equity really been rejected?

NOTES:

Statistical Equating & RTM

Campbell & Kenny: The regression artifact

- The sophomore jinx
- Spontaneous remission of depression
- Misclassification of individuals using standardized tests
- Perhaps:
  - Ashland cancer study
  - Washington D.C. vouchers
  - Sander's analysis of African-American failure on the bar exam
- Statistical equating
  - Regression to the mean leads to a bias in estimating gender differences using "equating"
  - Page 84: Ethnic differences in intellectual ability:
    - "We believe that the bias in statistical equating for ethnic differences in achievement and intelligence testing is underadjustment"

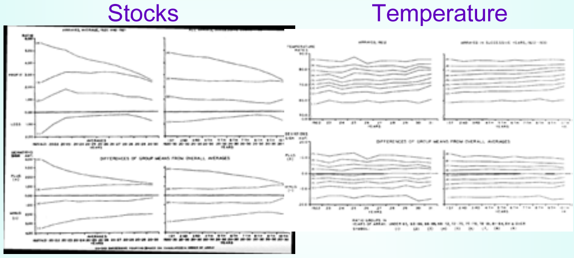
Slide 45 Statistical Equating & RTM

NOTES:

Poor Horace Secrist (1933)

Identify companies that had lower than average profits and invest in them; he was aware of RTM
Profits (left), temperature (right)

Stocks | Temperature



Slide 46 Poor Horace Secrist (1933)

NOTES:

Hotelling's (1933) JASA review

- Business varies, but average temperatures don't vary nearly as much
  - Secrist chose cities spread out throughout the country and looked at interannual variability
  - Small year-to-year variations compared to the big city-to-city variations
- Secrist rebuttal (1934)

Slide 47 Hotelling's (1933) JASA review

NOTES:

Hotelling's (1934) rejoinder

Quoted in Stigler's "Statistics on the Table"

"To 'prove' such a mathematical result [regression to the mean in annual reports] by a costly and prolonged numerical study of many kinds of business profit and expense ratios is analogous to proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals. The performance, though perhaps entertaining, and having a certain pedagogical value, is not an important contribution to either zoology or to mathematics."

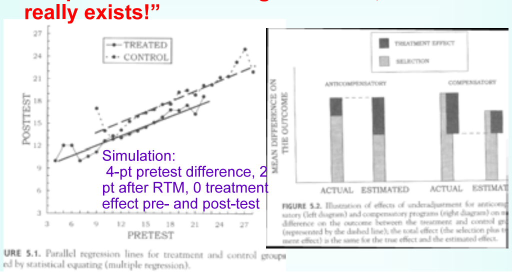
Slide 48 Hotelling's (1934) rejoinder

NOTES:

Statistical Equating

Effects on gender bias & racial differences

"Including a covariate, like socioeconomic status, can produce a racial or gender bias, when none really exists!"



Slide 49 Statistical Equating

NOTES:

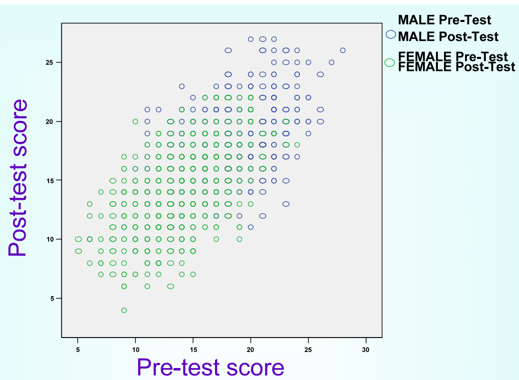
A hypothetical test of gender effects

Read Campbell & Kenny Chapters 4 & 5

- Are women inferior in mathematics?
- Randomly select 500 women & 500 men for admission to a intense workshop on advanced mathematics.
- Give both groups a pretest of mathematical ability
  - In the simulation (rtm-ck.sps) generate test scores by 4 tosses of a die. Assign males 4 units higher score in both pre & post test
    - Males: sum of 4 dice + 4
    - Females: sum of 4 dice + 0.
- Assume that the workshop does NOTHING to improve ability for either group
- Retest each student, the post-test, which is modeled to have a a correlation of 0.5 between pre- & post-test
  - 2 dice the same, 2 new dice throws for each student
- Test whether males did better than females in this advanced workshop, even after controlling for their previous math background

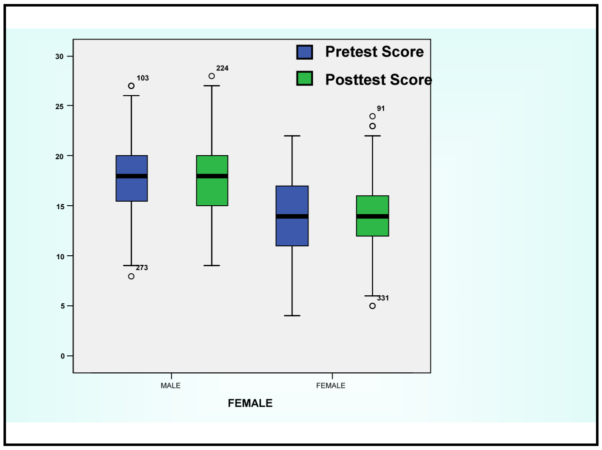
Slide 50 A hypothetical test of gender effects

NOTES:



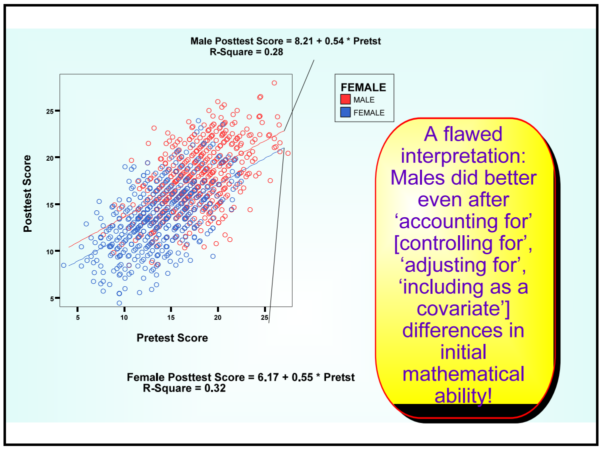
Slide 51

NOTES:



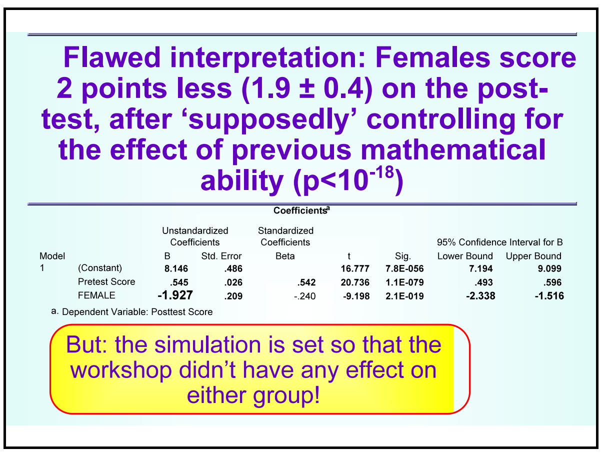
Slide 52

NOTES:



Slide 53

NOTES:



Slide 54 Flawed interpretation: Females score 2 points less (1.9 ± 0.4) on the post-test, after 'supposedly' controlling for the effect of previous mathematical ability (p<10-18)

NOTES:

Classic Analysis of covariance

Huge Male-female difference in post-workshop scores, after 'controlling' for pre-test ability

* Classic analysis of covariance (ANCOVA)
* to test for treatment effect
* with pretest as the covariate.
ANOVA postst BY treat(0,1) with pretst
 /STATISTICS=ALL.

ANOVA[a,b]

			Unique Method					
			Sum of Squares	df	Mean Square	F	Sig.	B
Posttest Score	Covariates	Pretest Score	**3630.112**	**1**	**3630.112**	**429.995**	**1.08E-079**	**.545**
	Main Effects	**FEMALE**	**714.243**	**1**	**714.243**	**84.604**	**2.08E-019**	
	Model		7654.148	2	3827.074	453.326	9.44E-141	
	Residual		8416.888	997	8.442			
	Total		16071.036	999	16.087			

a. Posttest Score by FEMALE with Pretest Score
b. All effects entered simultaneously

Slide 55 Classic Analysis of covariance

NOTES:

Repeated measures designs (Chapter 16) produce the correct solution: No effect of gender on post-test

There is no pre-test to post-test x gender interaction

Tests of Within-Subjects Effects

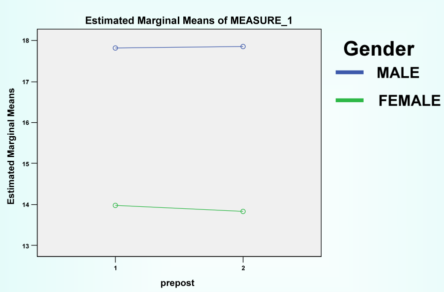
Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
prepost	Sphericity Assumed	**1.458**	**1**	**1.458**	**.266**	**.606**
	Greenhouse-Geisser	**1.458**	**1.000**	**1.458**	**.266**	**.606**
	Huynh-Feldt	**1.458**	**1.000**	**1.458**	**.266**	**.606**
	Lower-bound	**1.458**	**1.000**	**1.458**	**.266**	**.606**
prepost * treat	Sphericity Assumed	**4.232**	**1**	**4.232**	**.771**	**.380**
	Greenhouse-Geisser	4.232	1.000	4.232	.771	**.380**
	Huynh-Feldt	4.232	1.000	4.232	.771	**.380**
	Lower-bound	4.232	1.000	4.232	.771	**.380**
Error(prepost)	Sphericity Assumed	5476.310	998	5.487		
	Greenhouse-Geisser	5476.310	998.000	5.487		
	Huynh-Feldt	5476.310	998.000	5.487		
	Lower-bound	5476.310	998.000	5.487		

Slide 56 Repeated measures designs (Chapter 16) produce the correct solution: No effect of gender on post-test

NOTES:

Profiles from Repeated Measures ANOVA



Slide 57 Profiles from Repeated Measures ANOVA

NOTES:

Change score: Do paired *t* tests on males & females separately

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Pretest Score - Posttest Score	**-.038**	3.365	.150	**-.334**	**.258**	-.253	499	.801

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Pretest Score - Posttest Score	**.146**	3.260	.146	**-.140**	**.432**	1.002	499	.317

Slide 58 Change score: Do paired t tests on males & females separately

NOTES:

Why didn't regression & ANCOVA work?

See Cambell & Kenny (Ch 4-5) for full analysis

- Whenever there is less than perfect correlation between the covariate and the response, the effect of the covariate on the response is **not** removed by regression (=Analysis of covariance)
- This is due to regression to the mean
- Since the correlation between pre-test and post-test was set at r=0.5, only 50% of the pre-test effect can be 'explained' or accounted for by multiple regression
- Whenever the covariate is less than perfectly correlated with the response, multiple regression does not fully 'control for' or 'account for' or 'adjust for' the effects of the covariate.
  - Note that if the pre-test score had a correlation with the post-test score of 0.25, then only 1/4 of the pre-test difference would be accounted for by including pre-test as a covariate. There would a 3-point advantage for males after including pre-test as a covariate

Slide 59 Why didn't regression & ANCOVA work?

NOTES:

Galton's regression to the mean

Son's height 1" taller than father's, r=0.5, SD=2.5"

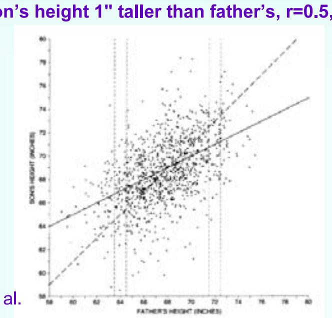
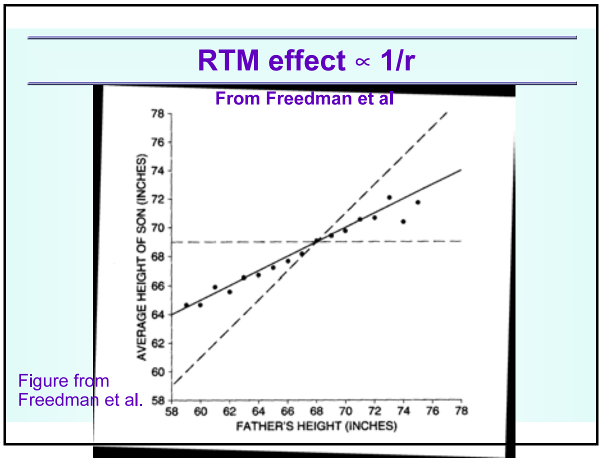


Figure from Freedman et al.

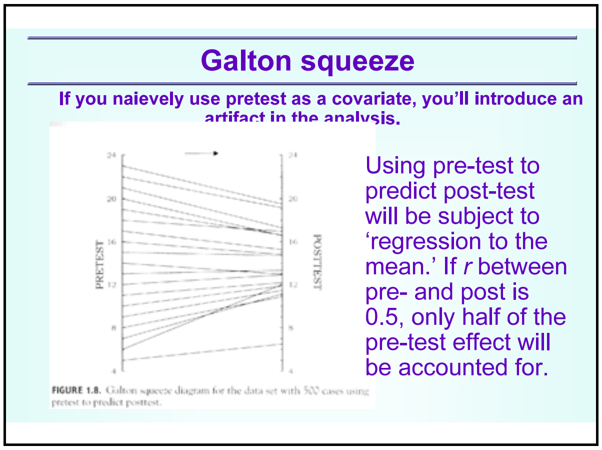
Slide 60 Galton's regression to the mean

NOTES:



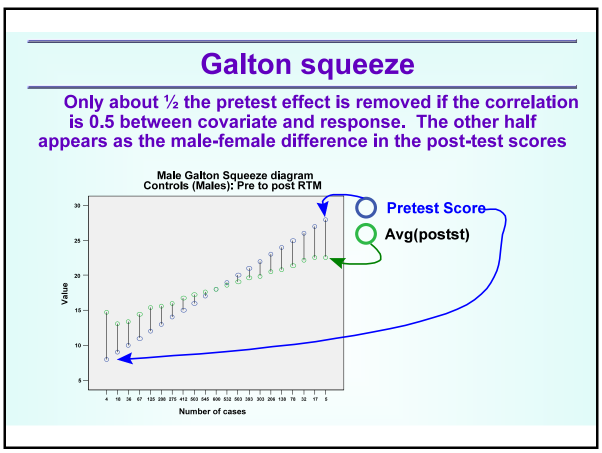
Slide 61 RTM effect ∝ 1/r

NOTES:



Slide 62 Galton squeeze

NOTES:

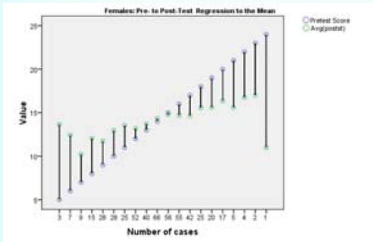


Slide 63 Galton squeeze

NOTES:

Galton squeeze, if r=0.25

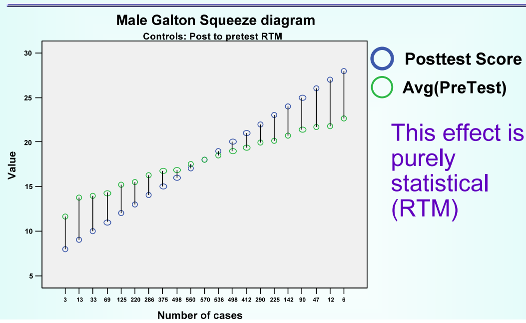
Only about ¼ of the pretest effect is removed if the correlation is 0.25 between covariate and response. The other ¾ appears as the male-female difference in the post-test scores



Slide 64 Galton squeeze, if r=0.25

NOTES:

Regression to the mean applies forward & backward



This effect is purely statistical (RTM)

Slide 65 Regression to the mean applies forward & backward

NOTES:

The Regression Fallacy

Stigler (1999) Chapter 9 Regression toward the mean

- "I suspect that **the regression fallacy** is the most common fallacy in the statistical analysis of economic data." Milton Friedman (1992) [emphasis added]
- "The recurrence of **regression fallacies** is testimony to its subtlety, deceptive simplicity, and I speculate, to the wide use of the word regression to describe least squares fitting of curves, lines, and surfaces. Researchers may err because they believe they know about regression, yet in truth have never fully appreciated how Galton's concept works. History suggests that this will not change soon. Galton's achievement remains one of the most attractive triumphs in the history of statistics, but it is one that each generation must leart to appreciate anew, on that seemingly never loses its power to surprise."

Slide 66 The Regression Fallacy

NOTES:

Statistical matching & equating

Creates 'bias' in assessing treatment effects

- **Matching:** If a covariate (e.g., pretest scores) is used to select groups, and there is less than perfect correlation between pre-and post-test assessments, then there will be regression to the mean.
  - Each group will regress to its own mean
  - **The regression to the mean effect will produce a treatment difference due to the treatment when none may have existed.**
  - Scaling College math performance vs. Gender based on categorical variables like (high school algebra I, Algebra I & II, Algebra I, II & Calculus) is still prone to the regression artifact
- **Equating**: If the covariate is weakly correlated with the presumed factor that it is controlling for (SES), & the covariate is positively associated with the response, then differences among groups can be magnified by the addition of the covariate.

Slide 67 Statistical matching & equating

NOTES:

Structural modeling *vs.* ANCOVA

Cook & Campbell 1979. Primer on Regression artifacts

- "The usefulness of analysis of covariance is closely coupled to the assumption that each covariate be measured without error"
  - Other assumptions too
  - Violation of this assumption could be disastrous
- Using unreliable covariates can produce treatment effects that do not exist and can mask strong treatment effects.
  - Gender discrimination
  - Racial differences on standardized tests
- Really unreliable covariates can change the sign of a treatment effect

Slide 68 Structural modeling vs. ANCOVA

NOTES:

Solutions to Equating & matching problems

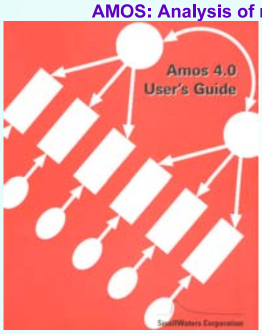
- Need a procedure that can adjust for the effect of the covariate, to correct for the 'bias' due to the regression to the mean phenomenan
- Equating & ANCOVA, may be ok when
  - Randomized assignment of subjects to cases
    - Equating not needed at all for reliability, but only for increasing 'power'
  - **If there is little correlation between the treatment groups and the covariate.**
- Alternatives to multiple regression: Structural equation modeling, change-score analysis (Campbell & Kenny 1999), Hierarchical linear models, James-Stein (empirical Bayes) estimators

Slide 69 Solutions to Equating & matching problems

NOTES:

Structural equation modeling

AMOS: Analysis of moment structures



Covered in EEOS612:
No time in EEOS611

Slide 70 Structural equation modeling

NOTES:

Path analysis and regression

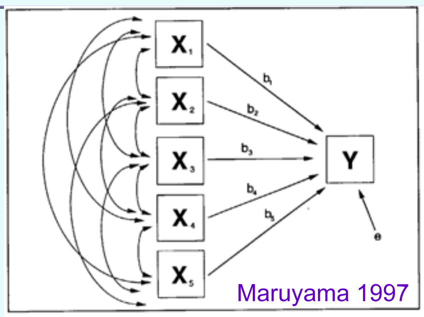


Figure 2.1. Regression Model With Five Predictor Variables

Slide 71 Path analysis and regression

NOTES:

Regression: a subset of structural equation modeling

The path diagram in **Figure 1** shows a model for these data:

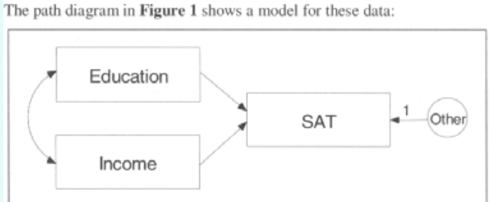
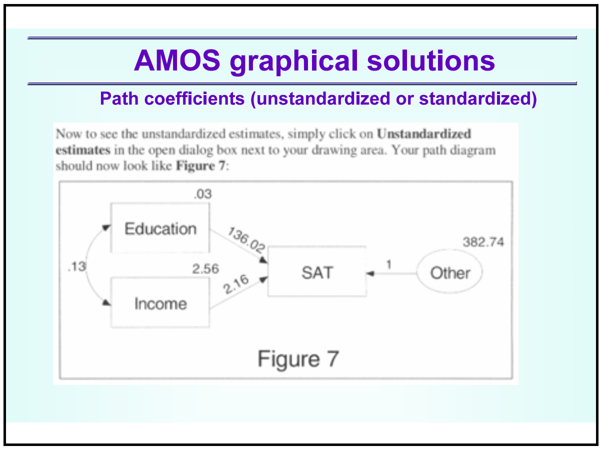


Figure 1

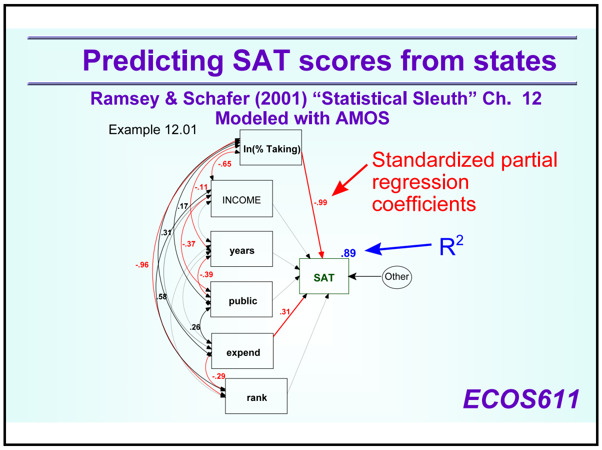
Slide 72 Regression: a subset of structural equation modeling

NOTES:



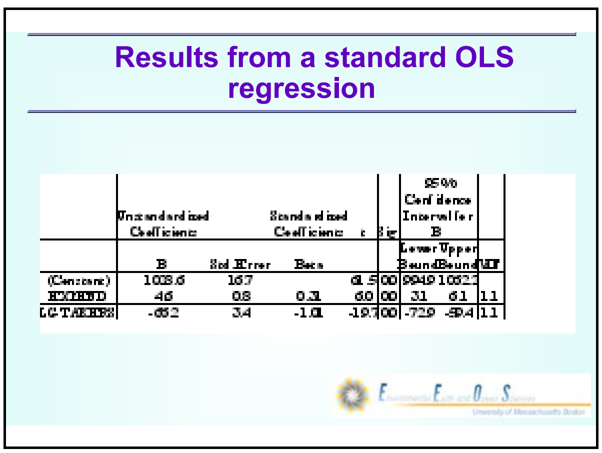
Slide 73 AMOS graphical solutions

NOTES:



Slide 74 Predicting SAT scores from states

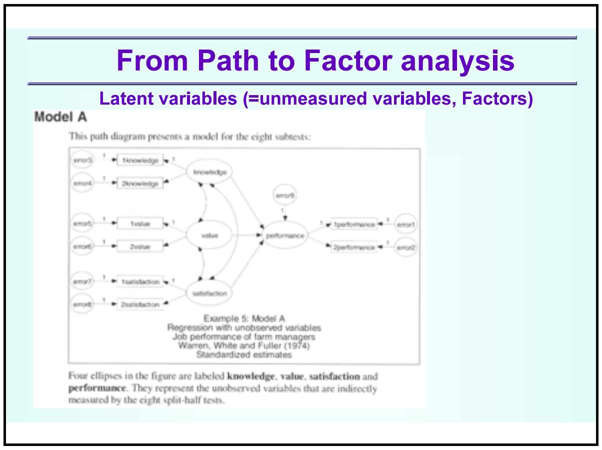
NOTES:



	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	VIF
(Constant)	1028.6	16.7		61.5	.000	994.9	1062.3	
EXPEND	4.6	0.8	0.31	6.0	.000	3.1	6.1	1.1
LGTAKERS	-65.2	3.4	-1.01	-19.1	.000	-72.9	-58.4	1.1

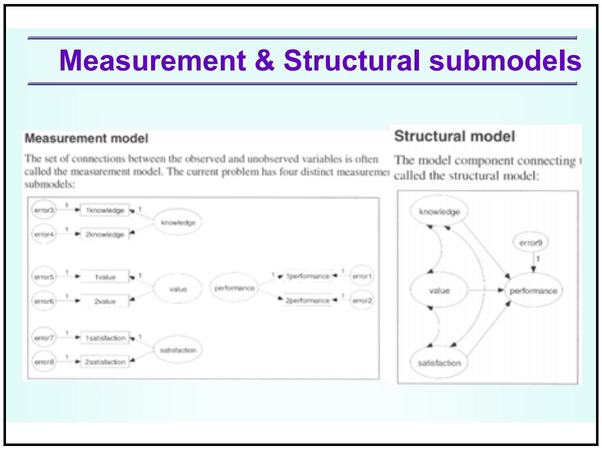
Slide 75 Results from a standard OLS regression

NOTES:



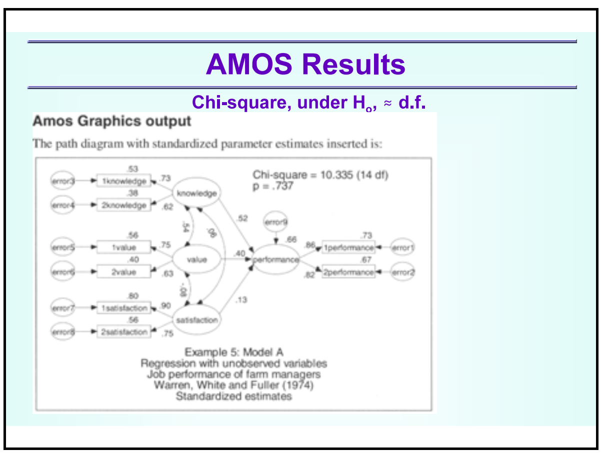
Slide 76 From Path to Factor analysis

NOTES:



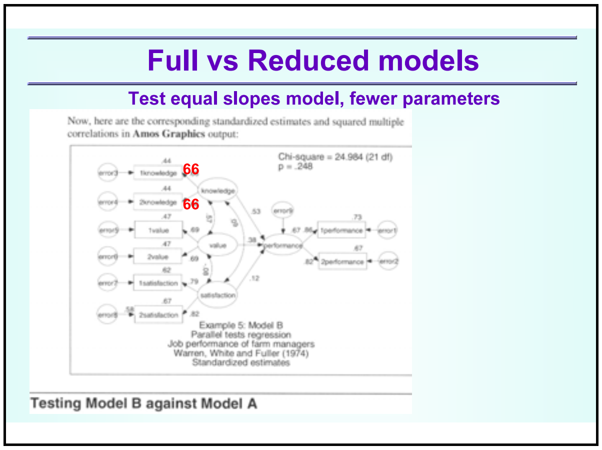
Slide 77 Measurement & Structural submodels

NOTES:



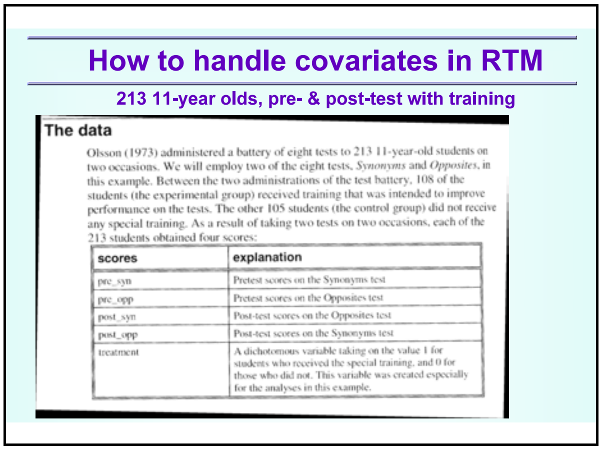
Slide 78 AMOS Results

NOTES:



Slide 79 Full vs Reduced models

NOTES:



How to handle covariates in RTM

213 11-year olds, pre- & post-test with training

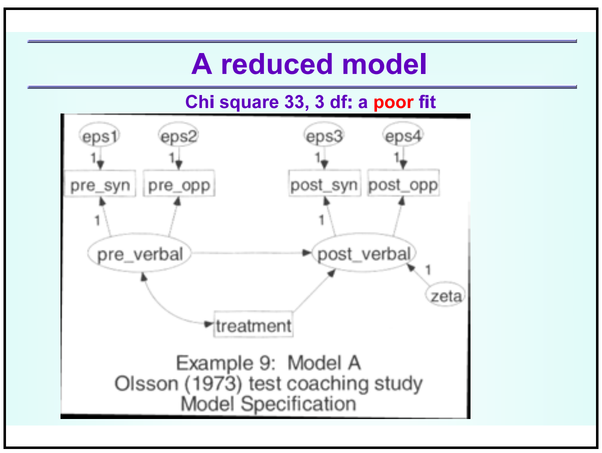
The data

Olsson (1973) administered a battery of eight tests to 213 11-year-old students on two occasions. We will employ two of the eight tests, *Synonyms* and *Opposites*, in this example. Between the two administrations of the test battery, 108 of the students (the experimental group) received training that was intended to improve performance on the tests. The other 105 students (the control group) did not receive any special training. As a result of taking two tests on two occasions, each of the 213 students obtained four scores:

scores	explanation
pre_syn	Pretest scores on the Synonyms test
pre_opp	Pretest scores on the Opposites test
post_syn	Post-test scores on the Opposites test
post_opp	Post-test scores on the Synonyms test
treatment	A dichotomous variable taking on the value 1 for students who received the special training, and 0 for those who did not. This variable was created especially for the analyses in this example.

Slide 80 How to handle covariates in RTM

NOTES:

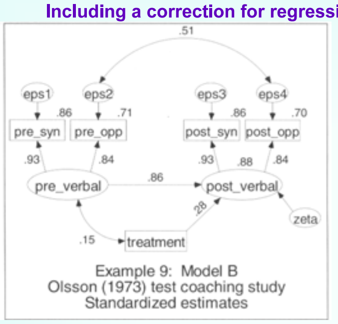


Slide 81 A reduced model

NOTES:

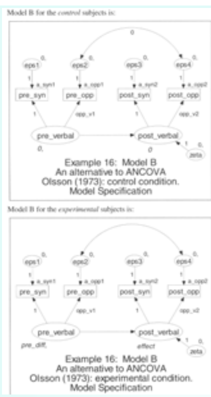
SEM, testing between groups

Including a correction for regression to the mean

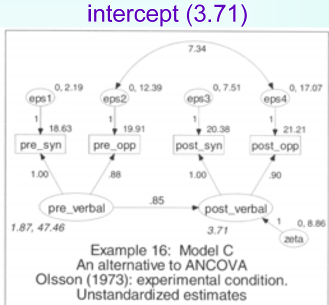


Slide 82 SEM, testing between groups

NOTES:



Testing treatment vs. Control with regression to mean; can estimate intercept (3.71)



Slide 83

NOTES:

MCAS Analyses and the thrip fallacy

Slide 84 MCAS Analyses and the thrip fallacy

NOTES:

Applications to SAT & MCAS

- SAT scores: can be analyzed using SEM
  - % Taking exams and expenditure per students are the most important variabels
- How should socioeconomic factors be included in evaluating schools with MCAS
  - Strong collinearity among socio-economic variables
  - Gaudet & UMASS Donahue Instiute
    - Socioeconomic variables are strongly correlated
    - Used principal component regression (didn't need to)
    - Could have used ridge regression
  - Tuerck, Beacon Hill Institute
    - Class size increases MCAS scores: probably an artifact, but need original data.
  - Chen & Ferguson (2002) simultaneous spatial autoregressive model (SAR)

Slide 85 Applications to SAT & MCAS

NOTES:

Gaudet's Ranking of MA Schools

1998 UMASS/Amherst Ph.D. and Donahue Institute Annual reports

- Gaudet's method for evaluating school quality
  - Socioeconomic variables from the 1990 census database, per student expenditure from MA DOE, MEAP results
  - 6 variables used in a "Major Axis" or principal components regression
    - average education level, average income, poverty rate, single-parent status, language spoken, and percentage of school-age population enrolled in private schools.
  - 86% of the variation in 1998 MCAS score is due to socioeconomic background of the students
  - Reduced to 85%, 83%, 81% and 81%MA
- Rerank 240 communities after controlling for 6 socioeconomic factors.

Slide 86 Gaudet's Ranking of MA Schools

NOTES:

The best 10th grade classes

Gaudet's ranking for President Bulger's office

District	ELA 10 Score	Overscore
Berlin Boylsto	255	10
Lenox	251	8
Stoneham	250	8
Northampton	248	8
Harvard	254	8
Nauset	250	8
Braintree	250	8
Clinton	245	7
Wareham	244	7
Shrewsbury	251	7
Pentucket Re	250	6
Norwood	248	5
Westborough	251	5

District	Math 10 Score	Overscore
Harvard	254	10
Lenox	250	9
Newburyport	251	8
Westborough	253	8
Amesbury	246	8
Northampton	245	7
Gardner	240	7
Nauset	247	7
Shrewsbury	249	7
Berlin Boylsto	250	7
Braintree	247	6
Nashoba	250	6
Tyngsboroug	245	6

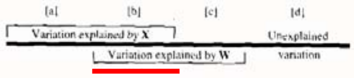
Similar to Case Study 12.1, the residual after fitting covariates (Socio-economic factors) is used to assess teaching Quality

Slide 87 The best 10th grade classes

NOTES:

The thrip/regression fallacy

Variation in Response variable(s) Y can be partititioned



From Legendre & Legendre (1998)

Figure 10.10 Partition of the variation of a response variable **y** among two sets of explanatory variables **X** and **W**. The length of the horizontal line corresponds to 100% of the variation in **y**. Fraction [b] is the intersection of the linear effects of **X** and **W** on **y**. Adapted from Legendre (1993).

Andrewartha & Birch (1954) on 'weather' vs. Biological interactions controlling thrip abundance and Smith's critique

Slide 88 The thrip/regression fallacy

NOTES:

Chen & Ferguson (2002)

Evaluating school quality

$$Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \varepsilon_i \qquad \text{(A5.1)}$$

where, $Y_i, i = 1,2,\cdots 226$ is the grand average of MCAS scores for years 1998, 1999, and 2000 for district i, and X_{ij}, $j = 1,2,3,4$ are the covariates of economic and demographic factors. They are AFRICAN-AMERICAN, PERCAP, TWOPHLD, and TAFDCPER. (LIM.ENG, which might quite reasonably be deemed a non-school related variable, is not used in this equation, since in combination with these variables alone it is not significant.) Once again, however, a Moran test indicates that the residuals of (A5.1) are spatially autocorrelated.

Slide 89 Chen & Ferguson (2002)

NOTES:

Just as in the earlier equation we employ spatial models. Here the model is:

$$Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \delta_i + \varepsilon_i \qquad \text{(A5.2)}$$

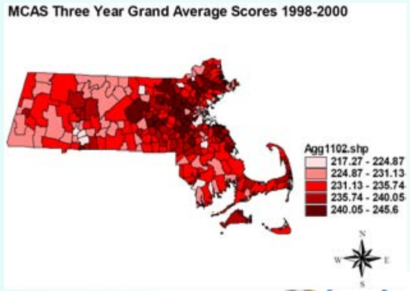
Again, as in Appendix 3, we estimate both a Conditional Spatial Autoregression (CAR) model using S-Plus and a Bayesian spatial approach estimated with WinBUGS. The estimated coefficients and p-values are listed in Table A5.3.

	S-PLUS	WinBuGS
INTERCEPT	221.54(.00)	224.20
AFRICAN	-0.160(.00)	-0.162
PERCAP	0.594(.00)	0.602
TWOPHLD	0.122(.00)	0.125
TAFDCPER	-2.124(.00)	-2.213

Slide 90 Chen & Ferguson (2002)

NOTES:

Spatially correlated residuals



Slide 91 Spatially correlated residuals

NOTES:

(See Text – Details of Economic/Demographic Equation Below)

RANK	SCHOOL	GRADV890	RES890	BAYRE890	3 YR TOT STUDS
1	AMHERST PELHAM	237.35	6.73	7.29	8709
2	LENOX	239.78	6.03	5.37	1701
3	HARVARD	245.60	5.57	6.10	2451
4	WESTBOROUGH	242.76	5.01	5.16	6251
5	BELMONT	243.95	4.61	4.02	7083
6	NAUSET	238.97	4.46	3.13	6837
7	NORTH READING	241.52	4.42	4.32	4468
8	NORTHAMPTON	234.23	4.20	4.21	6101
9	ACTON BOXBOROUGH	245.32	4.20	4.49	10307
10	HAMILTON WENHAM	241.37	3.89	3.62	5026
11	SANDWICH	240.02	3.79	3.04	7946
12	ARLINGTON	239.12	3.45	3.26	8153
13	NEWTON	243.60	3.34	2.59	22600
14	HADLEY	238.29	3.33	4.00	1274

Slide 92 Chen & Ferguson (2002)

NOTES:

208	MARBLEHEAD	238.98	-2.94	-2.82	5612
209	BELLINGHAM	232.54	-3.00	-2.65	5359
210	SOUTH HADLEY	231.13	-3.01	-2.47	4797
211	SAUGUS	231.55	-3.04	-3.39	6638
212	WINCHENDON	227.46	-3.09	-3.43	3995
213	TAUNTON	226.04	-3.33	-3.20	15658
214	EASTHAMPTON	228.51	-3.49	-2.80	3731
215	MARLBOROUGH	231.44	-3.55	-3.57	8028
216	CAMBRIDGE	226.07	-3.67	-3.25	13788
217	LAWRENCE	217.50	-3.68	-2.76	21674
218	HAVERHILL	226.92	-4.24	-3.86	16712
219	MAYNARD	231.77	-4.45	-4.00	2656
220	AVON	227.79	-4.45	-4.12	1660
221	LOWELL	222.05	-4.60	-4.56	29854
222	WESTPORT COMMUNITY	229.16	-5.02	-4.19	3871
223	NARRAGANSETT	227.87	-5.17	-5.27	2941
224	SOUTHERN BERKSHIRE	228.34	-6.08	-5.65	2157
225	DOVER SHERBORN	244.12	-6.11	-5.96	3760
226	WESTON	245.15	-6.40	-6.83	4087

Slide 93 Chen & Ferguson (2002)

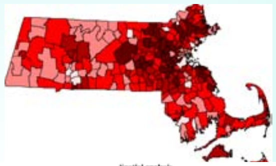
NOTES:

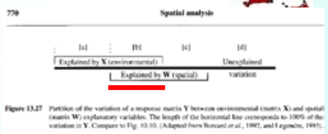
What factors affect test scores?

Table A4.2

Name of Variable	Definition	S-PLUS	WinBUGS
TPERCAP	Per Capita Income (Unit: $1,000)	.4505 (.000)	0.4409
TWOPHLD	Percent of Households With 2 Parents	.1348 (.000)	0.1241
TAFDCPER	Percentage of Pop Receiving TAFDC	-1.3860 (.000)	-1.435
SAL985	Superintendent's Salary – 1998-99 (Unit: $1,000)	.0469 (.000)	0.04162
AFRICAN-AMERICAN	Percentage of Students of African American Descent	-0.2112 (.000)	-0.2015
TSMX989	Teachers Maximum Salary 1998-99 (Unit: $1,000)	0.1356 (.004)	0.1297
INTERNTET	Percent of Classrooms With Internet Access	0.0111 (.009)	0.01184
LTOTAL	Log of total spending as % of Foundation Budget '97	3.3421 (.001)	4.468
LIM.ENG	Percent of Students Classified as Limited English Speakers	-0.1071 (.034)	-0.1334
ATHLETIC	Percent of Foundation Budget Spent on Athletics	-0.0081 (.015)	-0.007242
BOOKEQUL	Percent of Foundation Budget Devoted to Books and Equipment	-0.0099 (.032)	-0.00974
SENATE O	Senator Faced Opponent in 2000 (1= Yes, 0= No)	0.0604 (0.08)	0.8792

Note that: Column 3 is the estimated coefficients and P-values based on using S-Plus and Column 4 is the posterior mean based on on Bayesian approach using WinBUGS.





Slide 94 What factors affect test scores?

NOTES:

Beacon Hill Institute Study

Goal to rank schools & to evaluate educational policy

- Use 2000 MCAS scores as response variables
- Variables in Multiple regression:
  - Policy: % change in per pupil spending, percentage change in student-teacher ratios, number of students per computer
  - Socioeconomic: crime rates, % of workers that are professionals, % households headed by single females, Urban or non-urban
  - Choice variables: % students in charter schools, % students in METCO
  - Previous performance: 1994 MEAP scores

Slide 95 Beacon Hill Institute Study

NOTES:

Beacon Hill Results

Increase class sizes for "good schools"

- SES
  - School performance rises with % professionals or managers
  - School performance drops as the crime rate increases
  - School performance drops with higher % single parent households
  - Urbanized school districts have poorer perfomrance
- Choice
  - Charter schools 'spur schools to do better'
  - METCO has no effect
  - % of students attending public schools positively associated with scores
- Policy implications
  - Spending doesn't improve performance
  - Increased class size for "good districts" improves perfomrance
  - "Win-win situation" Increase class size in good districts by decreasing their funding and shift to poorer districts

Slide 96 Beacon Hill Results

NOTES:

The 15 best schools?

The 15 Best-Performing Massachusetts School Districts

	Achieving Good Performance (G Rating)			*Reducing Poor Performance (P Rating)*		
DISTRICT (number of ratings for which district fell in the top 10)	*4th*	*8th*	*10th*	*4th*	*8th*	*10th*
Hadley (5)	X	X	X		X	X
Clinton (3)	X	X		X		
Methuen (3)	X			X	X	
Stoneham (3)		X	X			X
Tyngsborough (3)	X		X			X
Nantucket (2)		X			X	
Chelsea (2)				X		X
Dighton-Rehoboth (2)		X			X	
Eastham (2)	X			X		
Everett (2)	X			X		
Hanover (2)		X			X	
Oxford (2)	X			X		
Provincetown (2)			X			X
Shrewsbury (2)			X			X
Sutton (2)	X			X		

Slide 97 The 15 best schools?

NOTES:

The 12 worst schools?

Beacon Hill Inst: Weighted average of 4th, 8th & 10th grades

The 12 Worst-Performing Massachusetts School Districts

	Achieving Good Performance (G Rating)			*Reducing Poor Performance (P Rating)*		
DISTRICT (number of ratings for which district fell in the bottom 10)	*4th*	*8th*	*10th*	*4th*	*8th*	*10th*
Narragansett (4)	X		X	X		X
Gateway (3)		X	X			X
Somerset (3)			X		X	X
Chesterfield-Goshen (2)	X			X		
Adams Cheshire (2)	X			X		
Hudson (2)		X			X	
Leicester (2)		X			X	
Millis (2)	X			X		X
Mount Greylock (2)		X			X	
Randolph (2)			X			X
Swampscott (2)			X			X
Watertown (2)		X	X			

Slide 98 The 12 worst schools?

NOTES:

The Worst 10th grade schools

Beacon Hill Institute

Foxborough	86
Weston	22
Quabbin	128
North Attleborough	171
Berkshire Hills	133
Uxbridge	170
Quaboag Regional	168
Harvard	17
Peabody	193
Longmeadow	46
Southwick Tolland	199
North Middlesex	88
Sutton	152
Hopedale	135
Mount Greylock	60
Douglas	172
Saugus	197
Taunton	210

Taunton	210
Winchendon	192
Wareham	186
Melrose	113
Carver	187
Leicester	142
Winthrop	188
Westford	63
Lunenburg	104
Randolph	200
Littleton	67
Lincoln-Sudbury	36
Watertown	132
Bellingham	174
Somerset	196
Narragansett	191
Swampscott	141
Gateway	207

Slide 99 The Worst 10^{th} grade schools

NOTES:

The Beacon Hill Institute Report

Would increasing class size improve performance?

- Beacon Hill study
  - No attempt was made to assess colinearity among the many strongly correlated explanatory variables
  - Multicollinearity would invalidate many of their interpretations of betas, especially class size
    - The authors should have calculated VIF's
    - Solutions
      - Do ridge regression or principal components regression
      - Create a structural equation model for the hypotheses
  - A major conclusion from the study that increased class size improves MCAS performance runs counter to controlled experiments
- Experiments or quasi-experiments peformed on class size indicate a negative correlation between class size and performance
  - STAR
  - SAGE

Slide 100 The Beacon Hill Institute Report

NOTES:

Class size and test scores

Inference: reduced class size causes improved performance

- The Tennessee Star Study
  - A controlled experiment
  - Students randomly assiged to class sizes of 15 or 24
  - Long-lasting effects
- The Wisconsin SAGE study
  - Students randomly assigned to small and large classes.
- Analysis of covariance (i.e, multiple regression) IS NOT a valid alternative to a randomized experiment

Slide 101 Class size and test scores

NOTES:

Conclusions

- Regression to the mean will be present whenever an explanatory variable (covariate) exhibits less than perfect correlation with the response variable. The higher the variability in the covariate, the more the regression to the mean effect
- For pre-test vs. Post-test analyses, regressing with pretest score as an explanatory variable DOES NOT remove the effects of pre-test differences.
- Better approaches: Repeated measures designs, hierarchical linear longitudinal models, or subtract pretest from posttest (called change score analysis)

Slide 102 Conclusions

NOTES: