

<div> <div>Chapter 12: Strategies for Variable Selection (Class 1 of 2)</div> <div>Class 19, 4/15/09 W</div> </div>	<div>Slide 1 Chapter 12: Strategies for Variable Selection (Class 1 of 2)</div> <div>NOTES:</div>
<div> <div>HW 12 due Friday 4/24/09</div> <div>Submit as Myname-HW12.doc (or *.rtf)</div> <ul style="list-style-type: none"> WIMBA sessions: <ul style="list-style-type: none"> Tonight & every Weds 10-11 pm Thursday Noon - 1 pm (log on from anywhere) New Homework due dates <ul style="list-style-type: none"> HW 12 10:28: El Niño and Hurricanes Due Friday 4/24/09 Noon Note: There will 2 WIMBA sessions available on this topic HW 13 Cammen's ingestion rate data. Note that this was a 2003 final exam problem <ul style="list-style-type: none"> Read Cammen (1980) & evaluate his regression model Due Weds 4/29/09 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 </div>	<div>Slide 2 HW 12 due Friday 4/24/09</div> <div>NOTES:</div>
<div> <div>HW13: Cammen model</div> <p>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 $\log_{10}(\text{ING})$ as the response variable with two explanatory variables $\log_{10}(\text{WT})$ and $\log_{10}(\text{OM})$. He deleted the three bivalves from his analyses because they appeared to be outliers, and based his regressions on the 19 non-bivalve species.</p> </div>	<div>Slide 3 HW13: Cammen model</div> <div>NOTES:</div>

Table 1. Data from **Cammen (1980)**. Loaded on **Prometheus** as **cammen.csv**, in case you wanted to examine the data (optional). The last 5 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 rate to account for temperature effects (higher ingestion at higher temperatures). **CPI** is the organic matter content (% wet plus organic matter) of total sediment dry weight, expressed as %.

Species	Taxon	WT	ING	CPI
1. <i>Nereis virens</i>	Gastropod mollusc	0.2	0.57	16
2. <i>Nereis virens</i>	Gastropod mollusc	0.2	0.66	17
3. <i>Tubificoides</i>	Oligochaete (annelid)	0.27	0.46	29.7
4. <i>Nereis virens</i>	Crustacean	0.32	0.48	50
5. <i>Phoronopsis viridis</i>	Gastropod mollusc	0.46	2.7	14.4
6. <i>Nereis virens</i>	Gastropod mollusc	0.9	0.67	13
7. <i>Nereis virens</i>	Polychaete (annelid)	5.8	20.2	6.8
8. <i>Phoronopsis viridis</i>	Crustacean (annelid)	8.4	1.49	93
9. <i>Orchestoidea</i>	Crustacean	12.4	4.4	88
10. <i>Alvinicella</i>	Polychaete (annelid)	20.4	24.0	2.2
11. <i>Monoporeia</i>	Polychaete (annelid)	46	250	1
12. <i>Ampelisca</i>	Crustacean	53	300	4.2
13. <i>Uca</i>	Crustacean	63.3	19.9	51
14. <i>Scapharca</i>	Crustacean	65	59	23.6
15. <i>Phoronopsis</i>	Polychaete (annelid)	80	1007	0.7
16. <i>Alvinicella</i>	Polychaete (annelid)	280	3400	1.2
17. <i>Alvinicella</i>	Polychaete (annelid)	380	3400	0.4
18. <i>Alvinicella</i>	Polychaete (annelid)	880	4700	0.64
19. <i>Alvinicella</i>	Crustacean	2050	4680	2.1
20. <i>Macoma</i>	Bivalve mollusc	5.1	4.49	20
21. <i>Macoma</i>	Bivalve mollusc	19.9	3.24	6.8
22. <i>Macoma</i>	Bivalve mollusc	380	4.3	3.4

Slide 4

NOTES:

HW13: Cammen model

Answer each question and address each issue.

- 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 HW13: Cammen model

NOTES:

Homework Presentations

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



Slide 6 Homework Presentations

NOTES:

<p>Chapter 12: Strategies for variable selection</p>	<p>Slide 7 Chapter 12: Strategies for variable selection</p> <p>NOTES:</p>
<p>Using multiple regression to test causal models</p> <p>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</p> <p>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</p>	<p>Slide 8 Using multiple regression to test causal models</p> <p>NOTES:</p>
<p>Regression errors & artifacts</p> <ul style="list-style-type: none"> ● A) Covariates are often necessary <ul style="list-style-type: none"> ▸ Fluoride & cancer (Manly 1992) ▸ Storks & babies ● B) Multicollinearity: <ul style="list-style-type: none"> ▸ 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 <ul style="list-style-type: none"> ▸ Math ability & gender ▸ The Bell Curve 	<p>Slide 9 Regression errors & artifacts</p> <p>NOTES:</p>

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

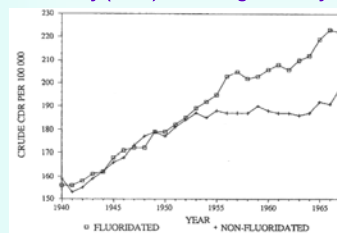


Figure 1.2. Crude cancer death rates per 100,000 of population for ten fluoridated and ten non-fluoridated cities in the United States, 1940-69. Fluoridation of cities took place over the period 1952-56.

Why ?

Yiamouyiannis & Burk 1977: 10 largest non-fl. Cities and matched fluoridated 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 modelling (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:

Display 12.1 p. 327

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

State	SAT	Take	Income	Years	Public	Expend	Rank
1. Iowa	1089	2	526	16.79	87.8	25.60	85.7
2. North Dakota	1079	3	284	16.77	89.7	24.91	86.4
3. North Dakota	1068	3	117	16.17	88.3	26.62	89.9
4. Kansas	1067	5	338	16.30	91.9	23.14	92.4
5. Nebraska	1065	5	293	17.23	83.6	23.00	88.3
6. Minnesota	1053	8	263	17.91	92.7	26.48	89.4
7. Montana	1052	4	345	16.47	79.2	27.42	93.9
8. Utah	1023	4	233	16.17	79.2	27.42	93.9
9. Wisconsin	1011	10	394	16.87	73.3	27.69	84.2
10. Wisconsin	1011	10	394	16.87	73.3	27.69	84.2
11. Oklahoma	1000	5	156	17.93	83.2	26.07	87.6
12. Arkansas	999	4	293	14.45	88.9	13.71	89.3
13. Tennessee	999	6	136	17.72	83.7	24.17	93.4
14. New Mexico	993	7	285	16.14	92.1	17.80	83.9
15. Idaho	989	3	133	16.76	87.9	27.36	92.1
16. Kentucky	981	6	130	16.41	71.4	15.49	89.4
17. Colorado	981	6	130	16.41	89.2	26.36	93.4
18. Washington	981	19	309	16.23	87.5	26.13	87.1
19. Illinois	971	11	114	17.86	86.9	14.14	86.3
20. Missouri	971	14	147	17.86	86.9	14.14	86.3
21. Louisiana	971	9	284	16.87	44.8	19.72	82.9
22. Michigan	971	10	325	16.42	67.7	26.79	90.9
23. Michigan	971	10	325	16.42	67.7	26.79	90.9
24. West Virginia	968	7	282	17.08	81.6	18.16	89.2
25. Alabama	964	6	131	16.37	69.6	13.84	93.9
26. Ohio	959	16	266	16.52	73.2	24.43	79.4
27. Alaska	923	13	489	15.12	96.5	50.00	76.6
28. Nevada	923	13	289	14.73	89.3	12.76	82.1
29. Oregon	906	40	263	14.48	92.2	30.49	79.3
30. Vermont	906	34	223	16.30	82.2	17.79	87.1
31. California	897	42	275	16.97	83.9	27.83	71.4
32. Connecticut	897	42	275	16.97	83.9	27.83	71.4
33. Maryland	896	40	266	16.05	81.7	26.11	76.1
34. New York	896	19	136	16.86	80.4	13.18	76.1
35. Maine	896	40	266	16.05	81.7	26.11	76.1
36. Florida	889	19	275	15.91	80.5	22.62	74.9
37. Massachusetts	889	19	275	15.91	80.5	22.62	74.9
38. Maryland	889	40	266	16.05	81.7	26.11	76.1
39. Delaware	888	43	246	16.79	80.7	13.14	69.9
40. Rhode Island	877	39	228	16.17	76.7	25.19	73.4
41. New Jersey	877	39	228	16.17	76.7	25.19	73.4
42. Texas	868	27	303	14.95	92.7	19.03	74.9
43. Indiana	868	48	236	14.39	17.93	23.49	67.1
44. Florida	877	47	234	15.11	92.8	19.92	79.3
45. Georgia	877	51	236	15.35	86.5	19.52	74.9
46. South Carolina	790	48	214	17.42	88.1	17.60	74.9

Case Study 12.1
SAT Scores



Slide 14

NOTES:

Display 12.2 p. 328

State SAT scores after adjustment (in points above or below average)

SAT'S ADJUSTED FOR % TAKING EXAM AND THEIR MEDIAN CLASS RANK

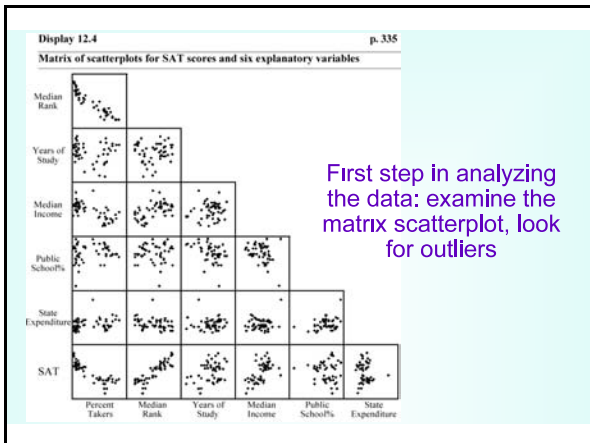
SAT'S ADJUSTED FOR % TAKING EXAM, RANK, AND EXPENDITURE

Rank	State	Adjusted SAT score	Rank	State	Adjusted SAT score
1	New Hampshire	45	1	New Hampshire	45
2	Massachusetts	44	2	Massachusetts	44
3	Connecticut	43	3	Connecticut	43
4	California	42	4	California	42
5	Washington	41	5	Washington	41
6	Minnesota	40	6	Minnesota	40
7	Illinois	39	7	Illinois	39
8	New York	38	8	New York	38
9	Florida	37	9	Florida	37
10	North Carolina	36	10	North Carolina	36
11	Ohio	35	11	Ohio	35
12	Michigan	34	12	Michigan	34
13	Wisconsin	33	13	Wisconsin	33
14	South Dakota	32	14	South Dakota	32
15	Idaho	31	15	Idaho	31
16	Montana	30	16	Montana	30
17	Utah	29	17	Utah	29
18	Alaska	28	18	Alaska	28
19	Arizona	27	19	Arizona	27
20	Nebraska	26	20	Nebraska	26
21	West Virginia	25	21	West Virginia	25
22	Delaware	24	22	Delaware	24
23	Alabama	23	23	Alabama	23
24	Georgia	22	24	Georgia	22
25	South Carolina	21	25	South Carolina	21
26	Mississippi	20	26	Mississippi	20
27	Arkansas	19	27	Arkansas	19
28	Louisiana	18	28	Louisiana	18
29	Texas	17	29	Texas	17
30	Kentucky	16	30	Kentucky	16
31	Indiana	15	31	Indiana	15
32	Ohio	14	32	Ohio	14
33	Illinois	13	33	Illinois	13
34	Michigan	12	34	Michigan	12
35	Wisconsin	11	35	Wisconsin	11
36	Minnesota	10	36	Minnesota	10
37	California	9	37	California	9
38	Washington	8	38	Washington	8
39	Oregon	7	39	Oregon	7
40	Arizona	6	40	Arizona	6
41	Colorado	5	41	Colorado	5
42	Idaho	4	42	Idaho	4
43	Montana	3	43	Montana	3
44	Utah	2	44	Utah	2
45	Alaska	1	45	Alaska	1
46	South Carolina	0	46	South Carolina	0
47	Georgia	-1	47	Georgia	-1
48	Florida	-2	48	Florida	-2
49	Alabama	-3	49	Alabama	-3
50	Mississippi	-4	50	Mississippi	-4
51	Arkansas	-5	51	Arkansas	-5
52	Louisiana	-6	52	Louisiana	-6
53	Texas	-7	53	Texas	-7
54	Kentucky	-8	54	Kentucky	-8
55	Indiana	-9	55	Indiana	-9
56	Ohio	-10	56	Ohio	-10
57	Illinois	-11	57	Illinois	-11
58	Michigan	-12	58	Michigan	-12
59	Wisconsin	-13	59	Wisconsin	-13
60	Minnesota	-14	60	Minnesota	-14
61	California	-15	61	California	-15
62	Washington	-16	62	Washington	-16
63	Oregon	-17	63	Oregon	-17
64	Arizona	-18	64	Arizona	-18
65	Colorado	-19	65	Colorado	-19
66	Idaho	-20	66	Idaho	-20
67	Montana	-21	67	Montana	-21
68	Utah	-22	68	Utah	-22
69	Alaska	-23	69	Alaska	-23
70	South Carolina	-24	70	South Carolina	-24



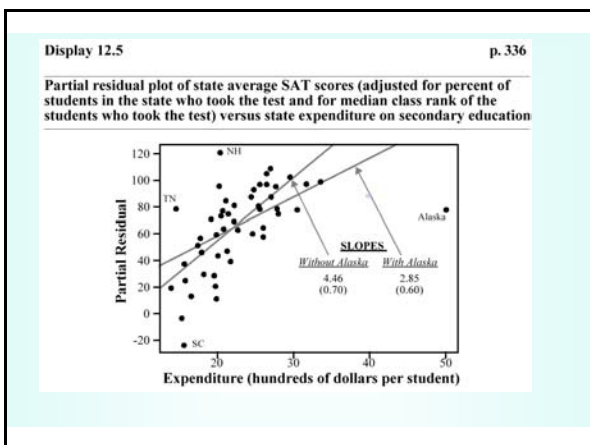
Slide 15

NOTES:



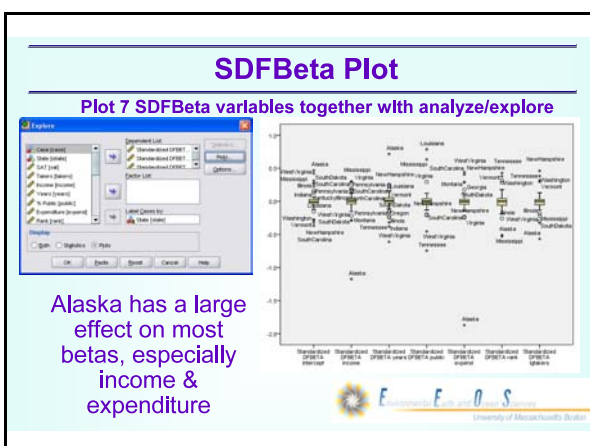
Slide 16

NOTES:



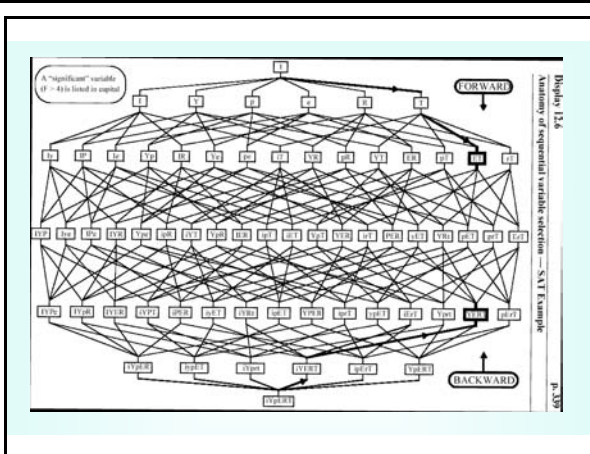
Slide 17

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 C_p
 - $C_p = p + (n-p)(\sigma^2 - \sigma^2_{full}) / \sigma^2_{full}$
- Akaike information content
- Schwarz Bayesian Information Content (BIC)
 - $\ln(\log(\sigma^2) + p \log(n))$.
 - 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 ^a									
Change Statistics									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	df1	df2	Sig. F Change	Selection Criteria
1	.388 ^a	.148	.086	1.32344	.148	1	61	.192	Akaike Information Criterion 28.216 Schwarz Bayesian Criterion 111.619
2	.728 ^b	.532	.492	1.00379	.384	14	34	.000	Akaike Information Criterion 7.077 Schwarz Bayesian Criterion 69.935
3	.894 ^c	.758	.737	.87174	.266	44	34	.000	Akaike Information Criterion -26.298 Schwarz Bayesian Criterion -12.396

a. Predictors: (Constant), In (Precipitation), In (Area), In (Runoff), In (Discharge), In (Deposition), In (NO₃ precipitation)

b. Predictors: (Constant), In (Precipitation), In (Area), In (Runoff), In (Discharge), In (Deposition), In (NO₃ precipitation), In (Density)

c. Dependent Variable: In (NO₃)

Slide 21 SPSS selection criteria

NOTES:

Bayes' Theorem

Larsen & Marx 2nd Edition (2001)

Bayes' Theorem (Theorem 26.2 p. 65)

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

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

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

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

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

for any $i = 1, 2, \dots, n$.

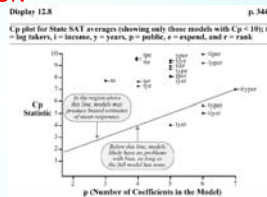
Slide 22 Bayes' Theorem

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 lgtakern income years public expend rank

/PARTIALPLOT ALL

/SCATTERPLOT=(*ZRESID ,*ZPRED)

/RESIDUALS ID(state)

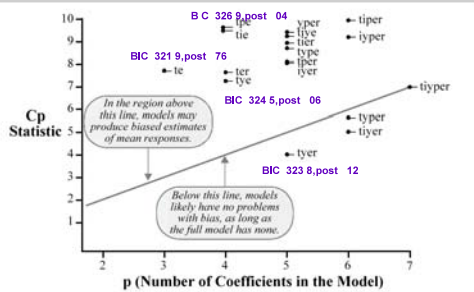
/SAVE PRED COOK MCIN ICIN RESID .

Slide 24 SPSS regression syntax

NOTES:

Display 12.8

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

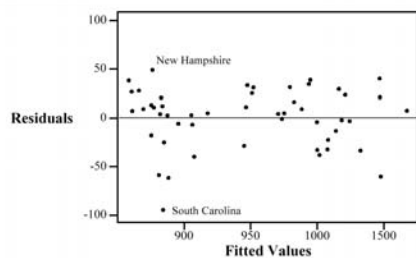


Slide 25

NOTES:

Display 12.13

Scatterplot of residuals versus fitted values from the regression of state SAT average on percent takers and median class rank of takers



Slide 26

NOTES:

Trouble assessing significance

Display 12.13, page 363

Statistical measures for the contribution of expenditure to different models (Alaska removed)

Model	t-Statistic	2-sided p-value	% variation explained by expenditure	Proportion Residual	Total Variation
0	0.27	.79	0.2	0.2	0.2
t-i	0.40	.69	0.3	0.3	0.3
t-y	0.86	.39	1.6	1.4	1.4
p-i	0.18	.86	0.1	0.1	0.1
t-i-e	4.78	.00002	49.6	7.4	7.4
t-i	6.04	.000002	76.4	8.4	8.4
t-y-i	0.07	.95	0.0	0.0	0.0
p-i-e	0.11	.92	0.0	0.0	0.0
t-i-e	4.88	.000001	52.9	6.8	6.8
t-i-e	3.88	.000005	36.7	5.1	5.1
t-y-p	1.05	.30	2.3	2.1	2.1
t-y-i-e	4.20	.0001	39.2	4.3	4.3
t-y-i	3.11	.00003	42.6	6.3	6.3
p-i-e	6.07	.000001	80.5	9.4	9.4
t-i-e	5.91	.000004	77.7	8.2	8.2
t-i-e	6.13	.000002	83.5	8.4	8.4
t-y-p-e	0.92	.36	0.2	0.0	0.0
t-y-i-e	4.32	.00008	43.1	4.2	4.2
t-y-i-e	3.20	.00005	40.4	6.2	6.2
p-i-e	3.27	.00001	36.6	5.0	5.0
t-y-i-e	3.20	.00001	36.6	7.3	7.3
t-i-e	3.92	.000004	39.7	7.6	7.6
t-y-p-e	4.80	.00002	52.3	5.3	5.3
t-y-i-e	4.78	.00002	51.9	5.3	5.3
t-y-i-e	3.23	.00001	42.1	5.6	5.6
t-y-i-e	4.27	.0001	42.4	4.1	4.1
t-y-i-e	4.23	.00009	41.7	4.3	4.3
t-y-i-e	3.00	.00001	38.2	5.1	5.1
p-i-e	3.94	.000004	42.0	8.0	8.0
t-y-p-i-e	5.13	.000005	61.3	5.4	5.4
t-y-p-i-e	4.64	.00003	51.2	4.3	4.3

Expenditure is correlated with other explanatory variables, so the significance (and magnitude) depends on the other variables in the model. Oregon ranks 31st in average SAT but 46th based on money spent.

Slide 27 Trouble assessing significance

NOTES:

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

Slide 28 Trouble assessing significance

NOTES:

Trouble assessing significance

Display 12.13, page 363

Statistical measures for the contribution of expenditure to different models
(Alaska removed)

Model	t-Statistic	2-tailed p-value	% variation explained by expenditure	Final Variation
E	0.21	.84	0.2	0.2
E+E	0.40	.69	0.3	0.2
V+E	0.80	.43	1.6	1.4
P+E	0.18	.86	0.1	0.1
R+E	4.78	.00002	49.6	7.4
T+E	0.04	.000002	78.4	8.4
IV+E	0.07	.92	0.0	0.0
DP+E	0.11	.92	0.0	0.0
DE+E	4.88	.00001	52.9	6.8
II+E	3.88	.00005	76.7	8.1
VP+E	1.05	.30	2.5	2.1
VR+E	4.20	.0001	78.2	4.3
VP+E	5.11	.00003	82.6	6.3
PR+E	6.07	.000001	83.5	9.4
PT+E	5.91	.000004	77.7	8.2
RT+E	6.17	.000002	83.5	8.4
IVP+E	0.02	.88	1.02	0.0
IVR+E	4.33	.00008	43.1	4.2
IVT+E	5.20	.00005	46.4	6.2
IPR+E	5.57	.00001	76.6	8.0
IPV+E	5.50	.00001	71.1	7.1
IRT+E	5.92	.000004	78.7	7.8
IVPR+E	4.80	.00002	52.3	5.2
VPR+E	4.78	.00002	51.9	5.1
VRP+E	5.23	.00005	42.1	4.6
PRP+E	6.27	.000001	89.5	8.8
IVPR+E	4.27	.0001	42.4	4.1
IVPT+E	4.31	.00009	41.7	4.1
IVRT+E	5.00	.00001	38.2	3.1
IPR+E	5.94	.000004	82.0	8.0
VPR+E	5.15	.00003	61.5	5.4
IVPR+E	4.64	.00003	51.2	4.5

Expenditure is correlated with other explanatory variables, so the significance (and magnitude) depends on the other variables in the model. Oregon ranks 31st in average SAT but 46th based on money spent.

Slide 29 Overfitting: why stepwise procedures should not be used to estimate p values.

NOTES:

Covariates: overfitting & multicollinearity

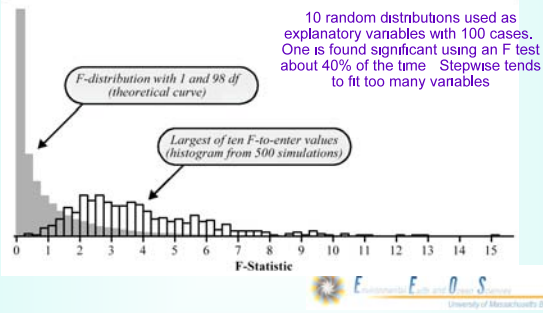
- Overfitting.sps
 - ▶ 32 random variables, 100 cases
 - ▶ Stepwise, forward & backward regression will usually always find a significant regression
 - ▶ One solution: use 40 times as many cases as covariates (>1200 for a 32-variable model!)
- More guns, less crime
 - ▶ Ayres, I and JJ Donohue (2003) Shooting down the more guns, less crime hypothesis. Stanford Law Review.
 - ▶ Including many covariates, many correlated with the key explanatory variable (gun control laws) produces an artifact, showing an effect when none existed
- Peterson's voucher studies
 - ▶ Kreuger critique: Including pre-test scores as a covariate produces an effect when none existed

Slide 30 Covariates: overfitting & multicollinearity

NOTES:

Display 12.7

Simulated distribution of the largest of ten F-statistics



Slide 31

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 #1 = 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 32 Gallagher's overfitting.sps

NOTES:

Results of Stepwise Selection

31 Random predictor variables

Model	Unstandardized Coefficients		Standardized Coefficients		t		Sig.		95% Confidence Interval for B	
	B	Std. Error	Beta		t				Lower Bound	Upper Bound
1	(Constant)	275			2.849	.004	.001	.485		
	V1	.301	.003	.111	3.237	.002	.001	.495		
2	(Constant)	275			2.854	.004	.001	.484		
	V1	.328	.003	.146	3.815	.000	.001	.519		
	V2	.154	.003	.053	2.116	.037	.001	.214		
3	(Constant)	266			3.211	.002	.001	.479		
	V1	.303	.004	.106	4.875	.000	.001	.570		
	V2	.189	.006	.122	2.838	.007	.006	.339		
	V13	.189	.006	.122	2.838	.007	.006	.339		
4	(Constant)	275			3.475	.001	.001	.492		
	V1	.411	.003	.425	4.402	.000	.001	.587		
	V13	.204	.007	.133	2.926	.004	.005	.419		
	V19	.227	.001	.204	2.980	.011	.005	.419		
	V2	.189	.007	.051	2.111	.038	.001	.254		
5	(Constant)	273			3.559	.001	.001	.491		
	V1	.420	.002	.424	4.963	.000	.001	.602		
	V13	.262	.005	.192	5.005	.000	.001	.602		
	V19	.226	.002	.243	2.911	.014	.007	.405		
	V2	.200	.006	.224	2.940	.007	.006	.379		
	V23	.189	.001	.187	2.959	.001	.001	.368		

a. Dependent Variable: Y1

Backward
(added V23, V19) →

25	(Constant)	266			3.204	.001	.001	.489		
	V1	.182	.006	.163	2.886	.004	.005	.362		
	V2	.174	.005	.184	2.959	.003	.005	.362		
	V1	.421	.000	.425	4.875	.000	.001	.587		
	V13	.200	.009	.225	2.264	.020	.020	.388		
	V19	.188	.009	.187	1.870	.065	.242	.010		
	V23	.146	.003	.146	1.746	.084	.369	.030		
	V21	.261	.004	.278	2.896	.005	.005	.318		

a. Dependent Variable: Y1

Slide 33 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 34 Harrell (2002, p. 56-57) on stepwise

NOTES:

Overfitting: too many covariates

Harrell (2001, p. 60)

"When a model is fitted that is too complex, that is it has too many free parameters to estimate for the amount of information in the data, the worth of the model (e.g., R^2) will be exaggerated and future observed values will not agree with predicted values. In this situation **overfitting** is said to be present, and some of the findings of the analysis come from fitting noise or finding spurious associations between X and Y"

Slide 35 Overfitting: too many covariates

NOTES:

Number of cases needed for regression (1 of 2)

Harrell (2001, p. 61)

- Number of predictors should be less than $m/10$ or $m/20$ where m is the limiting sample size shown below
- Candidate variables must include all variables screened for association with response, including nonlinear terms and interactions

TABLE 4.1: Limiting Sample Sizes for Various Response Variables

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ ^c
Ordinal (k categories)	$n - \frac{1}{n^2} \sum_{i=1}^k n_i^2$ ^d
Failure (survival) time	number of failures ^e

Slide 36 Number of cases needed for regression (1 of 2)

NOTES: