Chapter 12: Strategies for Variable Selection (Class 1 of 2)

Class 19, 4/15/09 W

Slide 1 Chapter 12: Strategies for Variable Selection (Class 1 of 2)

NOTES:

HW 12 due Friday 4/24/09

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

- WIMBA sessions:
  - Tonight & every Weds 10-11 pm
  - Thursday Noon - 1 pm (log on from anywhere)
- New Homework due dates
  - 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
  - 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

Slide 2 HW 12 due Friday 4/24/09

NOTES:

HW13: 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 HW13: 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	0.66	17
3	Tubifex tubifex	Oligochaete (annelid)	0.27	0.46	29.7
4	Hyalella azteca	Crustacean	0.32	0.46	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	[illegible]	Crustacean (amphipod)	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	0.04
19	Macrophthalmus japonicus	Crustacean	2050	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:

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:

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

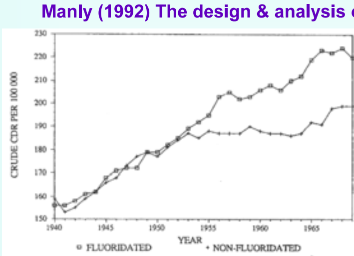


Figure 1.2. Crude cancer death rates per 100 000 of population for ten fluoridated and ten non-fluoridated cities in the United States, 1940–69. Fluoridation of cities took place over the period 1952–56.

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:

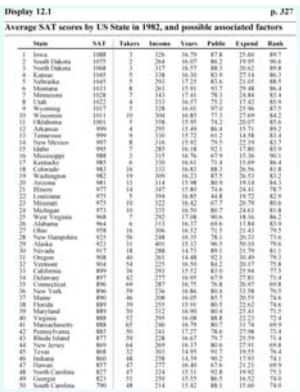
Gallagher's addenda

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:


Display 12.1 p. 327

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

	State	SAT	Takers	Income	Years	Public	Expend	Rank

Case Study 12.1
SAT Scores

Slide 14

NOTES:

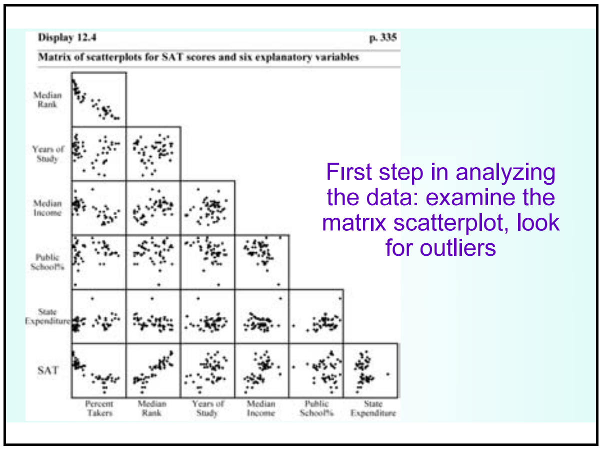

Display 12.2 p. 328

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

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)

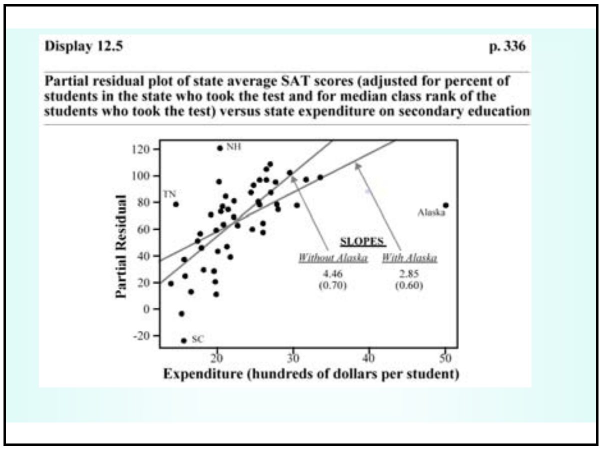
Slide 15

NOTES:



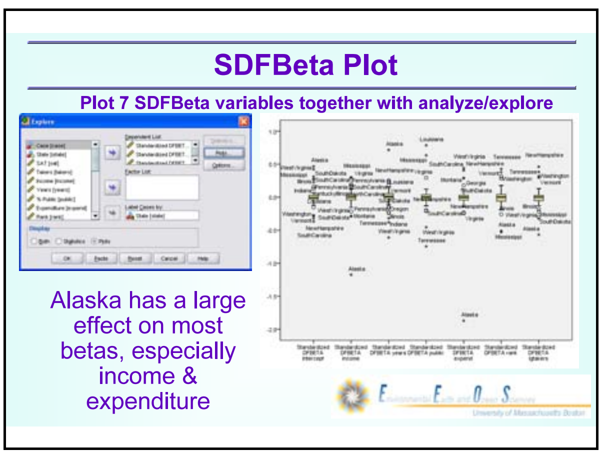
Slide 16

NOTES:



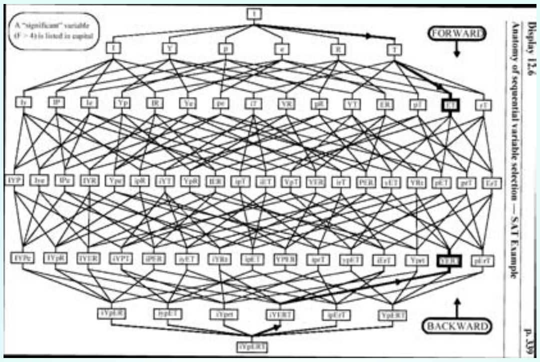
Slide 17

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 $A_i \cap A_j = \emptyset$ 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 $1 \le j \le n$.

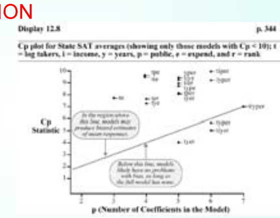
Slide 22 Bayes' Theorem

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:

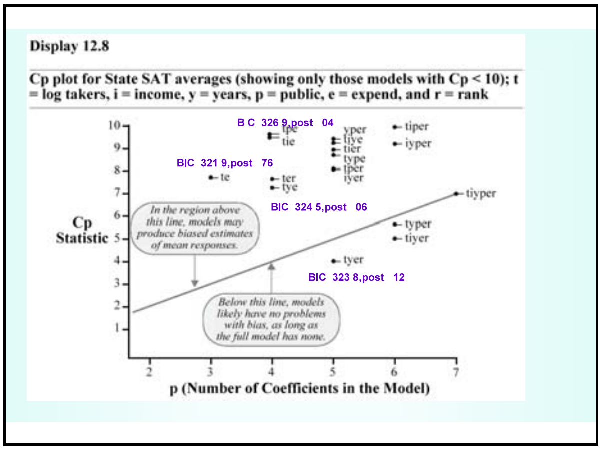
SPSS regression syntax

/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 .
```

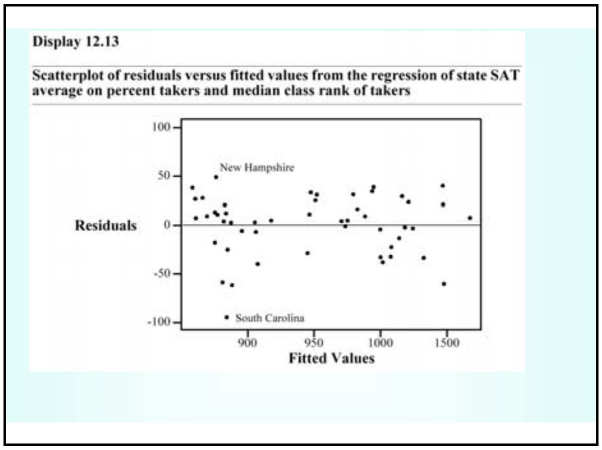
Slide 24 SPSS regression syntax

NOTES:



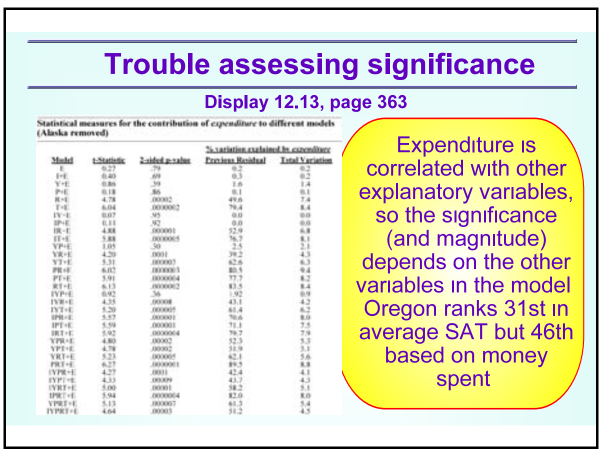
Slide 25

NOTES:



Slide 26

NOTES:



Slide 27 Trouble assessing significance

NOTES:

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

Slide 28 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-sided p-value	% variation explained by expenditure: Previous Residual	Total Variation
E	0.27	.79	0.2	0.2
I+E	0.40	.69	0.3	0.2
Y+E	0.86	.39	1.6	1.4
P+E	0.18	.86	0.1	0.1
R+E	4.78	.00002	49.6	7.4
T+E	6.04	.0000002	79.4	8.4
IY+E	0.07	.95	0.0	0.0
IP+E	0.11	.92	0.0	0.0
IR+E	4.88	.000001	52.9	6.8
IT+E	5.88	.0000005	76.7	8.1
YP+E	1.05	.30	2.5	2.1
YR+E	4.20	.0001	39.2	4.3
YT+E	5.31	.000003	62.6	6.3
PR+E	6.02	.0000003	80.5	9.4
PT+E	5.91	.0000004	77.7	8.2
RT+E	6.13	.0000002	83.5	8.4
IYP+E	0.92	.36	1.92	0.9
IYR+E	4.35	.00008	43.1	4.2
IYT+E	5.20	.000005	61.4	6.2
IPR+E	5.57	.000001	70.6	8.0
IPT+E	5.59	.000001	71.1	7.5
IRT+E	5.92	.0000004	79.7	7.9
YPR+E	4.80	.00002	52.3	5.3
YPT+E	4.78	.00002	51.9	5.1
YRT+E	5.23	.000005	62.1	5.6
PRT+E	6.27	.0000001	89.5	8.8
IYPR+E	4.27	.0001	42.4	4.1
IYPT+E	4.33	.00009	43.7	4.3
IYRT+E	5.00	.00001	58.2	5.1
IPRT+E	5.94	.0000004	82.0	8.0
YPRT+E	5.13	.000007	61.3	5.4
IYPRT+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

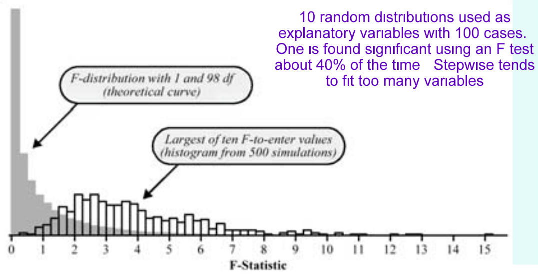
Slide 30 Covariates: overfitting & multicollinearity

NOTES:

Slide 31

Display 12.7

Simulated distribution of the largest of ten F-statistics



NOTES:

Slide 32 Gallagher's overfitting.sps

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.
```

NOTES:

Slide 33 Results of Stepwise Selection

Results of Stepwise Selection

31 Random predictor variables

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	Upper Bound
1	(Constant)	.278	.094		2.949	.004	.091	.465
	V7	.301	.093	.311	3.237	.002	.116	.485
2	(Constant)	.270	.093		2.904	.005	.085	.454
	V7	.335	.093	.346	3.615	.000	.151	.519
	V31	-.182	.086	-.203	-2.118	.037	-.353	-.012
3	(Constant)	.296	.092		3.211	.002	.113	.479
	V7	.383	.094	.396	4.075	.000	.197	.570
	V31	-.209	.086	-.232	-2.438	.017	-.378	-.039
	V13	.186	.090	.200	2.074	.041	.008	.365
4	(Constant)	.310	.091		3.415	.001	.130	.490
	V7	.411	.093	.425	4.402	.000	.226	.597
	V31	-.254	.087	-.283	-2.926	.004	-.426	-.082
	V13	.237	.091	.254	2.590	.011	.055	.418
	V3	.182	.087	.203	2.101	.038	.010	.354
5	(Constant)	.313	.089		3.508	.001	.136	.491
	V7	.420	.092	.434	4.563	.000	.237	.602
	V31	-.262	.085	-.292	-3.065	.003	-.431	-.092
	V13	.226	.090	.243	2.511	.014	.047	.405
	V3	.200	.086	.224	2.340	.021	.030	.370
	V2	.188	.091	.187	2.069	.041	.008	.368

a. Dependent Variable: V1

Backward
(added V23, V19)

25	(Constant)	.295	.088		3.354	.001	.120	.469
	V2	.183	.090	.183	2.046	.044	.005	.362
	V3	.174	.085	.194	2.056	.043	.006	.342
	V7	.421	.090	.435	4.675	.000	.242	.600
	V13	.210	.089	.225	2.364	.020	.033	.386
	V19	-.166	.089	-.167	-1.870	.065	-.342	.010
	V23	-.145	.083	-.156	-1.746	.084	-.309	.020
	V31	-.242	.084	-.270	-2.880	.005	-.410	-.075

a. Dependent Variable: V1

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)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ [d]
Failure (survival) time	number of failures [e]

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

NOTES: