Regression Analysis
Model Selection

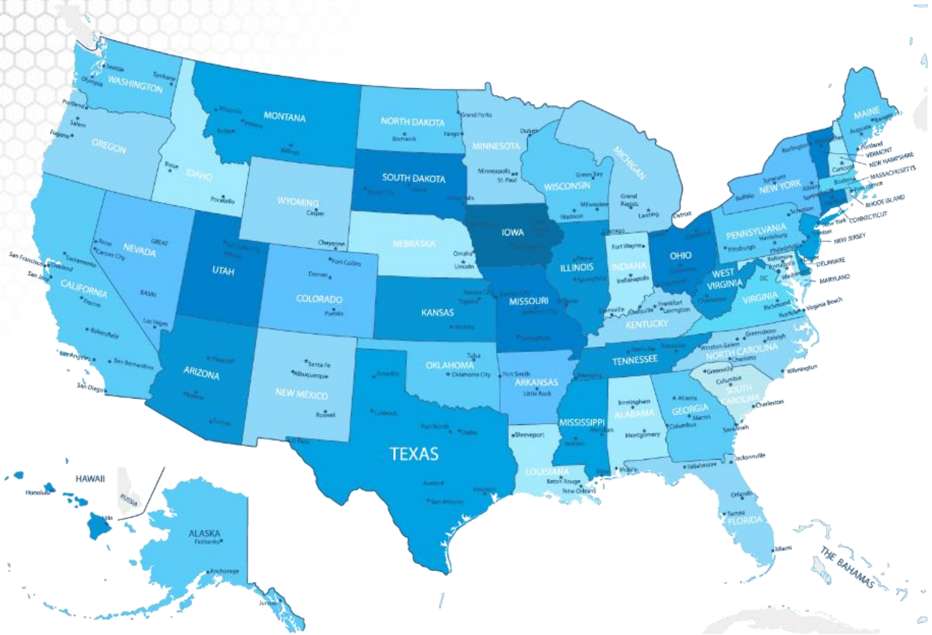
Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Model Search: Data Examples

Georgia Tech

About This Lesson

Ranking States by SAT Performance



SAT Mean Score by State – Year 1982
790 (South Carolina) – 1088 (Iowa)

- ***Which variables are associated with state average SAT scores?***
- ***After accounting for selection biases, how do the states rank?***
- ***Which states perform best for the amount of money they spend?***

Compare All Models

```
library(leaps)
out = leaps(datasat[,-c(1,2)], sat, method = "Cp")
cbind(as.matrix(out$which),out$Cp)
  1 2 3 4 5 6
1 0 0 0 0 0 1   34.026834
1 1 0 0 0 0 0   47.639512
1 0 1 0 0 0 0 187.387572
1 0 0 1 0 0 0 269.647903
1 0 0 0 1 0 0 306.188562
1 0 0 0 0 1 0 307.076043
⋮
6 1 1 1 1 1 1   7.000000

best.model = which(out$Cp==min(out$Cp))
cbind(as.matrix(out$which), out$Cp)[best.model,]
       1        2        3        4        5        6
0.000000 0.000000 1.000000 1.000000 1.000000 1.000000 3.581157
```

The output includes all 64 combinations of predictors with specification of which predictors are in the model and the Cp score value for each model.

The best model with respect to Mallow's Cp criterion:
years, public, expend, rank (last four predictors in the input dataset)

Does not allow for specification of controlling variables!!!

Stepwise Regression

 # Forward Stepwise Regression

step(lm(sat~log(takers)+rank), scope=list(lower=sat~log(takers)+rank,
upper=sat~log(takers)+rank+expend+years+income+public), direction="forward")

Start: AIC=346.7
sat ~ log(takers) + rank

	Df	Sum of Sq	RSS	AIC
+ expend	1	13149.5	32380	331.66
+ years	1	9827.2	35703	336.55
<none>			45530	346.70
+ income	1	1305.3	44224	347.25
+ public	1	15.9	45514	348.69

Step: AIC=331.66
sat ~ log(takers) + rank + expend

	Df	Sum of Sq	RSS	AIC
+ years	1	5743.5	26637	323.90
<none>			32380	331.66
+ public	1	421.0	31959	333.01
+ income	1	317.3	32063	333.17

Step: AIC=323.9
sat ~ log(takers) + rank + expend + years

	Df	Sum of Sq	RSS	AIC
<none>			26637	323.90
+ income	1	26.6165	26610	325.85
+ public	1	4.5743	26632	325.89

Call:
lm(formula = sat ~ log(takers) + rank + expend + years)

Coefficients:

(Intercept)	log(takers)	rank	expend	years
388.425	-38.015	4.004	2.423	17.857

Selected model: *expend* and *years*, with confounding variables log(*takers*) and *rank*

Stepwise Regression (cont’d)

 # Backward Stepwise Regression

full = lm(sat ~ log(takers) + rank + expend + years + income + public)
minimum = lm(sat ~ log(takers) + rank)
step(full, scope=list(lower=minimum, upper=full), direction="backward")

Start: AIC=327.8
sat ~ log(takers) + rank + expend + years + income + public

	Df	Sum of Sq	RSS	AIC
- public	1	25.0	26610	325.85
- income	1	47.0	26632	325.89
<none>			26585	327.80
- years	1	4588.8	31174	333.77
- expend	1	6264.4	32850	336.38

Step: AIC=325.85
sat ~ log(takers) + rank + expend + years + income

	Df	Sum of Sq	RSS	AIC
- income	1	26.6	26637	323.90
<none>			26610	325.85
- years	1	5452.8	32063	333.17
- expend	1	7430.3	34040	336.16

Step: AIC=323.9
sat ~ log(takers) + rank + expend + years

	Df	Sum of Sq	RSS	AIC
<none>			26637	323.90
- years	1	5743.5	32380	331.66
- expend	1	9065.8	35703	336.55

Call:
lm(formula = sat ~ log(takers) + rank + expend + years)

Coefficients:

(Intercept)	log(takers)	rank	expend	years
388.425	-38.015	4.004	2.423	17.857

Georgia Tech

Stepwise Regression (cont'd)

 # Backward Stepwise Regression

full = lm(sat ~ log(takers) + rank + expend + years + income + public)
minimum = lm(sat ~ log(takers) + rank)
step(full, scope=list(lower=minimum, upper=full), direction="backward")

Start: AIC=327.8
sat ~ log(takers) + rank + expend + years + income + public

	Df	Sum of Sq	RSS	AIC
- public	1	25.0	26610	325.85
- income	1	47.0	26632	325.89
<none>			26585	327.80
- years	1	4588.8	31174	333.77
- expend	1	6264.4	32850	336.38

Step: AIC=325.85
sat ~ log(takers) + rank + expend + years + income

	Df	Sum of Sq	RSS	AIC
- income	1	26.6	26637	323.90
<none>			26610	325.85
- years	1	5452.8	32063	333.17
- expend	1	7430.3	34040	336.16

- Selected model includes
  - *expend* and *years*
  - confounding variables log(*takers*) and *rank*
- The same model was selected using forward regression
  - Generally, for a large number of predictors, the two methods will select different models

Step: AIC=323.9
sat ~ log(takers) + rank + expend + years

	Df	Sum of Sq	RSS	AIC
<none>			26637	323.90
- years	1	5743.5	32380	331.66
- expend	1	9065.8	35703	336.55

Call:
lm(formula = sat ~ log(takers) + rank + expend + years)

Coefficients:

(Intercept)	log(takers)	rank	expend	years
388.425	-38.015	4.004	2.423	17.857

Predicting Bankruptcy

- Effective bankruptcy prediction is useful for investors and analysts, allowing for accurate evaluation of a firm’s prospects.

- Roughly 40 years ago, Ed Altman showed that publicly available financial indicators can be used to distinguish between firms that are about to go bankrupt and those that are not.

Which financial indicators are associated with bankruptcy for telecommunications firms?

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University and was inspired by the honors thesis of Jeffrey Lui.

Georgia Tech

Compare All Models

```
library(bestglm)
input.Xy <- as.data.frame(cbind(WC.TA, RE.TA, EBIT.TA, S.TA,
BVE.BVL,Bankrupt))
bestBIC <- bestglm(input.Xy, IC="BIC", family=binomial)

bank2 = glm(Bankrupt~RE.TA+EBIT.TA+BVE.BVL,
family=binomial, epsilon=1e-14, maxit=500, x=T)
summary(bank2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.29478	1.12323	-0.262	0.7930
RE.TA	-0.05627	0.02745	-2.050	0.0404 *
EBIT.TA	-0.16763	0.09270	-1.808	0.0706 .
BVE.BVL	-0.62975	0.39435	-1.597	0.1103

The best model selected with respect to BIC: *RE.TA, EBIT.TA, BVE.BVL*

- *RE.TA* is now statistically significant at $\alpha = 0.05$
- Not all coefficients are statistically significant

- RE.TA is associated with a decrease in the odds of going bankrupt in the next year by 5.5% holding all else fixed
- EBIT.TA is associated with a decrease in the odds of going bankrupt by 15%

Georgia Tech

Compare All Models (cont'd)

 # Testing for subset of regression coefficients

```
gstat = deviance(bank2) - deviance(bank1)
cbind(gstat, 1-pchisq(gstat,length(coef(bank1))-length(coef(bank2))))
        gstat
[1,] 4.040336 0.1326332
```

The null (reduced model) is not rejected

Remove Outlier

bankrupt2 = bankruptcy[-1,]
attach(bankrupt2)
bank3 = glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA + BVE.BVL, family=binomial,
maxit=500, data=bankrupt2)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(bank3)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	265.467	576281.709	0	1
WC.TA	-4.297	12439.717	0	1
RE.TA	-1.516	5131.146	0	1
EBIT.TA	-17.043	35543.170	0	1
S.TA	-2.859	7408.747	0	1
BVE.BVL	-77.540	184903.001	0	1

The model fits perfectly. This is complete separation, and the solution is to simplify the model if that is possible.

Compare All Models: Without Outlier

input.Xy <- as.data.frame(cbind(WC.TA, RE.TA, EBIT.TA, S.TA, BVE.BVL,Bankrupt))
bestBIC <- bestglm(input.Xy, IC="BIC", family=binomial)

bank4 = glm(Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family=binomial, maxit=500)
summary(bank4)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.09166	1.47135	-0.062	0.9503
RE.TA	-0.08229	0.04230	-1.945	0.0517 .
EBIT.TA	-0.26783	0.15854	-1.689	0.0912 .
BVE.BVL	-1.21810	0.76536	-1.592	0.1115

exp(coef(bank2)[-1])

RE.TA	EBIT.TA	BVE.BVL
0.9452862	0.8456655	0.5327273

exp(coef(bank4)[-1])

RE.TA	EBIT.TA	BVE.BVL
0.9210091	0.7650371	0.2957930

The best model selected with respect to BIC:
WC.TA, *RE.TA, EBIT.TA, BVE.BVL*

Stepwise Regression: Without Outlier

bank3.select=step(bank3, direction="backward")
summary(bank3.select)

Start: AIC=12
Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA + BVE.BVL

	Df	Deviance	AIC
- S.TA	1	0.0000	10.000
<none>		0.0000	12.000
- WC.TA	1	9.3839	19.384
- RE.TA	1	10.7362	20.736
- EBIT.TA	1	14.7992	24.799
- BVE.BVL	1	19.0267	29.027

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	255.413	728539.823	0	1
WC.TA	-9.542	23920.936	0	1
RE.TA	-5.152	15669.825	0	1
EBIT.TA	-28.983	90578.211	0	1
BVE.BVL	-103.614	225264.760	0	1

Step: AIC=10
Bankrupt ~ WC.TA + RE.TA + EBIT.TA + BVE.BVL

	Df	Deviance	AIC
<none>		0.0000	10.000
- WC.TA	1	9.3841	17.384
- RE.TA	1	12.8531	20.853
- EBIT.TA	1	14.8672	22.867
- BVE.BVL	1	19.1321	27.132

Stepwise regression selects the same four predictors as the best subset selection approach using BIC.

Georgia Tech

Summary