Regression Analysis
Model Selection

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

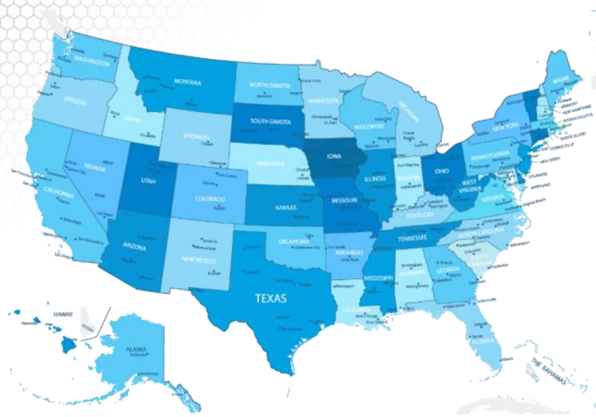
Data Examples

Georgia Tech

About This Lesson

Georgia Tech

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?***

Response & Predicting Variables

The response variable is:

Y = State average SAT score (verbal and quantitative combined)

The predicting variables are:

takers	% of eligible students (high school seniors) in state who took the exam
rank	Median percentile ranking of test takers in their secondary school classes
income	Median income of families of test takers (in $00's)
years	Average years test takers had in social/natural sciences and humanities
public	% of test takers who attended public schools
expend	State expenditure on secondary schools (in $00's/student)

Regression Analysis

regression.line = lm(sat ~ log(takers) + rank + income + years + public + expend)
summary(regression.line)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	407.53990	282.76325	1.441	0.15675	
log(takers)	-38.43758	15.95214	-2.410	0.02032	*
rank	4.11427	2.50166	1.645	0.10734	
income	-0.03588	0.13011	-0.276	0.78407	
years	17.21811	6.32007	2.724	0.00928	**
public	-0.11301	0.56239	-0.201	0.84168	
expend	2.56691	0.80641	3.183	0.00271	**

Test for Statistical Significance
p-values

$\hat{\beta}_{takers}$ ≈ 0.02
$\hat{\beta}_{rank}$ > 0.1
$\hat{\beta}_{income}$ > 0.1
$\hat{\beta}_{years}$ < 0.01
$\hat{\beta}_{public}$ > 0.1
$\hat{\beta}_{expend}$ < 0.01

Shall we discard the predicting variables with regression coefficients that are not statistically significant?

➔ NO. Perform variable selection.

Inference on Subset of Coefficients

regression.red = lm(sat ~ log(takers) + rank)
anova(regression.red, regression.line)

Model 1: sat ~ log(takers) + rank
Model 2: sat ~ log(takers) + rank + income + years + public + expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	45530				
2	43	26585	4	18945	7.6604	9.42e-05 ***

Testing for a subset of regression coefficients:

H_0: Reduced Model (takers and rank only)

vs.

H_A: Full Model

Partial F Test: F-value = 7.6604, P-value ≈ 0

Inference on Subset of Coefficients

regression.red = lm(sat ~ log(takers) + rank)
anova(regression.red, regression.line)

Model 1: sat ~ log(takers) + rank
Model 2: sat ~ log(takers) + rank + income + years + public + expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	45530				
2	43	26585	4	18945	7.6604	9.42e-05 ***

- **Controlling and explanatory variables**: log(takers) and rank need to be in the model.
- **Partial F test for explanatory variables**: at least one predicting variable has explanatory power. Which ones?

➔ Perform variable selection!!!

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.

Bankruptcy Data

Data Sample:

- 25 telecommunication firms that declared bankruptcy 2000–2002
- 25 telecommunication firms that did not declare bankruptcy, "matched" according to the asset size of the bankrupt firms

Replicate Experimental Data Setting:

→ Matching firms to be comparable with respect to meaningful factors

→ Allowing for causal inference

Response & Predicting Variables

The response variable is:

Y = Whether the firm declared bankruptcy

The predicting variables are:

WC.TA Working capital as a percentage of total assets (in %)

RE.TA Retained earnings as a percentage of total assets (in %)

EBIT.TA Earnings before interest and taxes as a percentage of total assets (in %)

S.TA Sales as a percentage of total assets (in %)

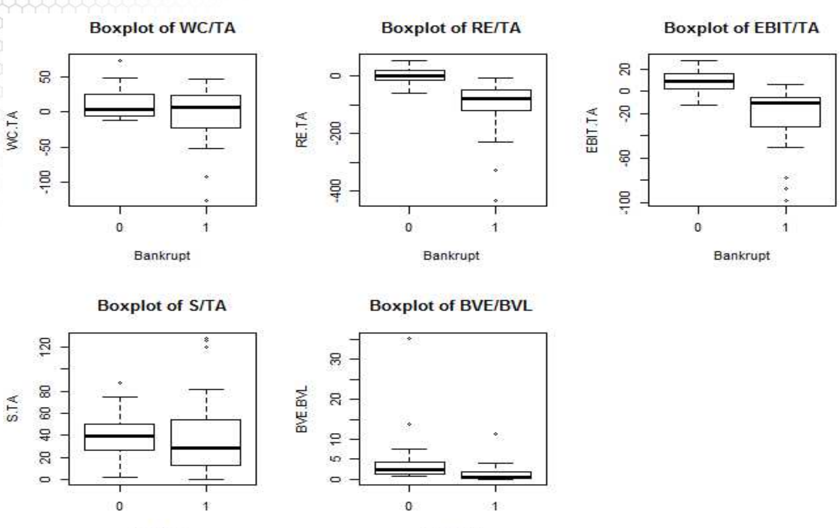
BE.T Book value of equity divided by book value of total liabilities

Exploratory Data Analysis

Read the data from the file## Exploratory analysis

```
bankruptcy = read.table("bankruptcy.dat", sep="\t", header=T, row.names=NULL)
attach(bankruptcy)

## Exploratory analysis
par(mfrow=c(2,3))

boxplot(split(WC.TA,Bankrupt), style.bxp="old", xlab="Bankrupt", ylab="WC.TA",
main="Boxplot of WC/TA")

boxplot(split(RE.TA,Bankrupt), style.bxp="old", xlab="Bankrupt", ylab="RE.TA",
main="Boxplot of RE/TA")

boxplot(split(EBIT.TA,Bankrupt), style.bxp="old", xlab="Bankrupt", ylab="EBIT.TA",
main="Boxplot of EBIT/TA")

boxplot(split(S.TA,Bankrupt), style.bxp="old", xlab="Bankrupt", ylab="S.TA",
main="Boxplot of S/TA")

boxplot(split(BVE.BVL,Bankrupt), style.bxp="old", xlab="Bankrupt", ylab="BVE.BVL",
main="Boxplot of BVE/BVL")
```

Exploratory Data Analysis



Regression Analysis

```
bank1 = glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA +
BVE.BVL, family=binomial)
summary(bank1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.42646	6.35770	1.168	0.243
WC.TA	-0.15587	0.12208	-1.277	0.202
RE.TA	-0.07605	0.06311	-1.205	0.228
EBIT.TA	-0.49111	0.32260	-1.522	0.128
S.TA	-0.08040	0.09216	-0.872	0.383
BVE.BVL	-2.07764	1.47488	-1.409	0.159

```
gstat = bank1$null.deviance - deviance(bank1)
cbind(gstat, 1 - pchisq(gstat, length(coef(bank1))-1))
```

```
     gstat
[1,] 57.46799 4.049594e-11
```

Test for Statistical Significance
All p-values > 0.1
None of the coefficients are statistically significant.

Test for Overall Regression
p-value ≈ 0
The overall regression has predictive power.

Summary