

Chapter 12 of *The Sleuth* is about strategies for variable selection. This chapter deals with a fundamental statistical difficulty, sometimes known as the *bias-variance* tradeoff. A model with many explanatory variables will have low bias, but high variance. There may exist estimates within the model that are, in some sense, close to the “true” model, but the procedure of estimating these modeling parameters will be highly variable from data set to data set. The estimated model from any given data set may be inaccurate. On the other hand, a model with too few explanatory variables will have low variance but can be quite biased if the best parameter estimates result in a model far from the truth. In a regression setting, the desire to reduce both bias and variance results in a tension between wishing to add more variables to reduce bias and to remove more variables from the model to reduce variability.

The main idea to keep in mind is that in most situations, there are a large number of models that are nearly equal in their ability to explain observed data. The goal of identifying the single best model that is very close to the “true” is generally unachievable. Different methods of variable selection will often wind up with different sets of included variables. The goal needs to be to find a good model recognizing that other models are also good. Great care needs to be taken in interpretation, especially in attributing causality or statistical significance.

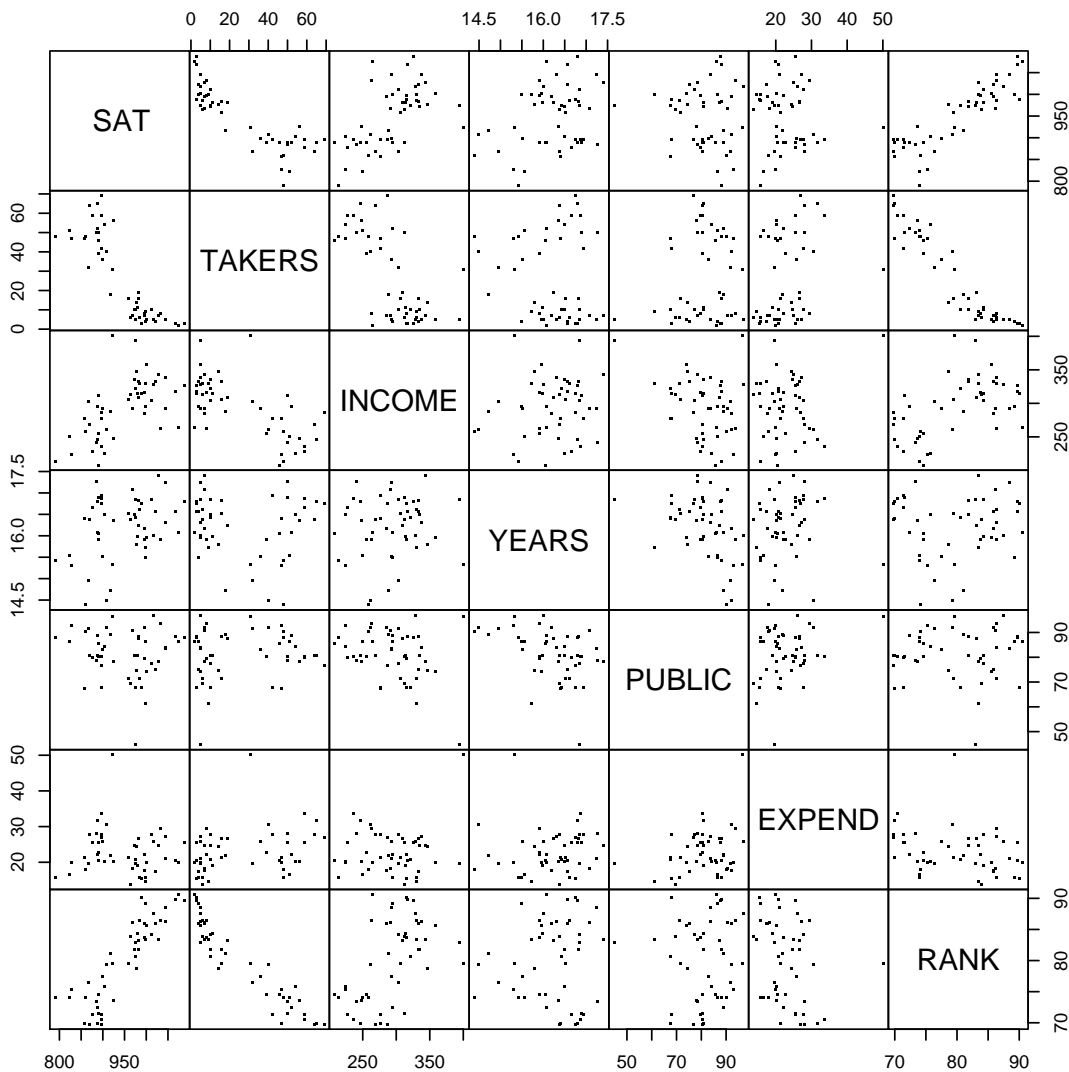
Chapter 12 describes several methods of variable selection. The remainder of this document shows how to implement some of these ideas in R. We will follow the SAT example.

```
> case1201 = read.table("sleuth/case1201.csv", header = T, sep = ",")
> attach(case1201)
```

Pairwise Scatterplots

In R, the function `pairs` will draw pairwise scatterplots for each pair of columns in a matrix. You can affect the output by changing several input variables. Here is an example with the SAT data resulting in a plot similar to that in Display 12.4. Notice that we do not want to include the first column of the data set which is the state name.

```
> pairs(case1201[, -1], gap = 0, pch = ".")
```



It is most informative to look at the row of plots for which SAT is the response. We see here that the percent takers seems to have a nonlinear relationship with SAT. There are potential outliers among public school percentage and state expenditure per student. *The Sleuth* argues for excluding Alaska from the analysis because its unusually high state expenditure per student is quite influential. Louisiana which has the low percentage of students in public schools is not as influential. Here is R code that creates a new data set without Alaska and with the percentage of students taking the exam log transformed.

```
> keep <- STATE != "Alaska"
> x <- data.frame(SAT = SAT[keep], ltakers = log(TAKERS[keep]),
+   income = INCOME[keep], years = YEARS[keep], public = PUBLIC[keep],
+   expend = EXPEND[keep], rank = RANK[keep])
> detach(case1201)
> attach(x)
```

Sequential Variable Selection Techniques

There are three common sequential variable selection methods. In *forward selection*, you begin with a small model and keep adding new variables, one at a time, picking the best variable by some criterion, until you cannot get a significantly better model. In *backward elimination*, you begin with a large model and continue removing existing variables, one at a time,

removing the least significant variable at each step, until all variables are significant. In *stepwise* regression, you alternate between forward and backward steps until you do not change.

Model Selection Methods and Criteria In *The Sleuth*, they give an example where the criterion for adding or removing a variable is of the square of the *t* statistic is at least 4. R has different option built in. There are several methods in common use to objectively distinguish between models with different sets of explanatory variables. No single method has been shown to be universally best. Most methods take the form of a measure of goodness-of-fit plus a penalty for each parameter used. The book describes the Cp statistic, Akaike's Information Criterion (AIC), and the Bayes Information Criterion (BIC). R uses AIC by default, but it is easy to make it use BIC instead.

Forward Selection

```
> step(lm(SAT ~ 1), SAT ~ ltakers + income + years + public + expend +
+ rank, direction = "forward")
```

Start: AIC= 419.42
SAT ~ 1

	Df	Sum of Sq	RSS	AIC
+ ltakers	1	199007	46369	340
+ rank	1	190297	55079	348
+ income	1	102026	143350	395
+ years	1	26338	219038	416
<none>			245376	419
+ public	1	1232	244144	421
+ expend	1	386	244991	421

Step: AIC= 339.78
SAT ~ ltakers

	Df	Sum of Sq	RSS	AIC
+ expend	1	20523	25846	313
+ years	1	6364	40006	335
<none>			46369	340
+ rank	1	871	45498	341
+ income	1	785	45584	341
+ public	1	449	45920	341

Step: AIC= 313.14
SAT ~ ltakers + expend

	Df	Sum of Sq	RSS	AIC
+ years	1	1248.2	24597.6	312.7
+ rank	1	1053.6	24792.2	313.1
<none>			25845.8	313.1
+ income	1	53.3	25792.5	315.0
+ public	1	1.3	25844.5	315.1

Step: AIC= 312.71
SAT ~ ltakers + expend + years

	Df	Sum of Sq	RSS	AIC
+ rank	1	2675.5	21922.1	309.1
<none>			24597.6	312.7

```
+ public 1      287.8 24309.8  314.1
+ income 1       19.2 24578.4  314.7
```

```
Step: AIC= 309.07
SAT ~ ltakers + expend + years + rank
```

	Df	Sum of Sq	RSS	AIC
<none>			21922.1	309.1
+ income	1	505.4	21416.7	309.9
+ public	1	185.0	21737.1	310.7

```
Call:
lm(formula = SAT ~ ltakers + expend + years + rank)
```

```
Coefficients:
(Intercept)      ltakers      expend      years      rank
  399.115      -38.100       3.996     13.147     4.400
```

Backward Elimination

```
> step(lm(SAT ~ ltakers + income + years + public + expend + rank),
+      direction = "backward")
```

```
Start: AIC= 311.88
SAT ~ ltakers + income + years + public + expend + rank
```

	Df	Sum of Sq	RSS	AIC
- public	1	20	21417	310
- income	1	340	21737	311
<none>			21397	312
- ltakers	1	2150	23547	315
- years	1	2532	23928	315
- rank	1	2679	24076	316
- expend	1	10964	32361	330

```
Step: AIC= 309.93
SAT ~ ltakers + income + years + expend + rank
```

	Df	Sum of Sq	RSS	AIC
- income	1	505	21922	309
<none>			21417	310
- ltakers	1	2552	23968	313
- years	1	3011	24428	314
- rank	1	3162	24578	315
- expend	1	12465	33882	330

```
Step: AIC= 309.07
SAT ~ ltakers + years + expend + rank
```

	Df	Sum of Sq	RSS	AIC
<none>			21922	309
- rank	1	2676	24598	313
- years	1	2870	24792	313
- ltakers	1	5094	27016	317
- expend	1	13620	35542	331

Call:
`lm(formula = SAT ~ ltakers + years + expend + rank)`

Coefficients:

(Intercept)	ltakers	years	expend	rank
399.115	-38.100	13.147	3.996	4.400

Stepwise Regression

`> step(lm(SAT ~ ltakers + income + years + public + expend + rank),
+ direction = "both")`

Start: AIC= 311.88
SAT ~ ltakers + income + years + public + expend + rank

	Df	Sum of Sq	RSS	AIC
- public	1	20	21417	310
- income	1	340	21737	311
<none>			21397	312
- ltakers	1	2150	23547	315
- years	1	2532	23928	315
- rank	1	2679	24076	316
- expend	1	10964	32361	330

Step: AIC= 309.93
SAT ~ ltakers + income + years + expend + rank

	Df	Sum of Sq	RSS	AIC
- income	1	505	21922	309
<none>			21417	310
+ public	1	20	21397	312
- ltakers	1	2552	23968	313
- years	1	3011	24428	314
- rank	1	3162	24578	315
- expend	1	12465	33882	330

Step: AIC= 309.07
SAT ~ ltakers + years + expend + rank

	Df	Sum of Sq	RSS	AIC
<none>			21922	309
+ income	1	505	21417	310
+ public	1	185	21737	311
- rank	1	2676	24598	313
- years	1	2870	24792	313
- ltakers	1	5094	27016	317
- expend	1	13620	35542	331

Call:
`lm(formula = SAT ~ ltakers + years + expend + rank)`

Coefficients:

(Intercept)	ltakers	years	expend	rank
399.115	-38.100	13.147	3.996	4.400