Chapter 9 Variable Selection and Model Building

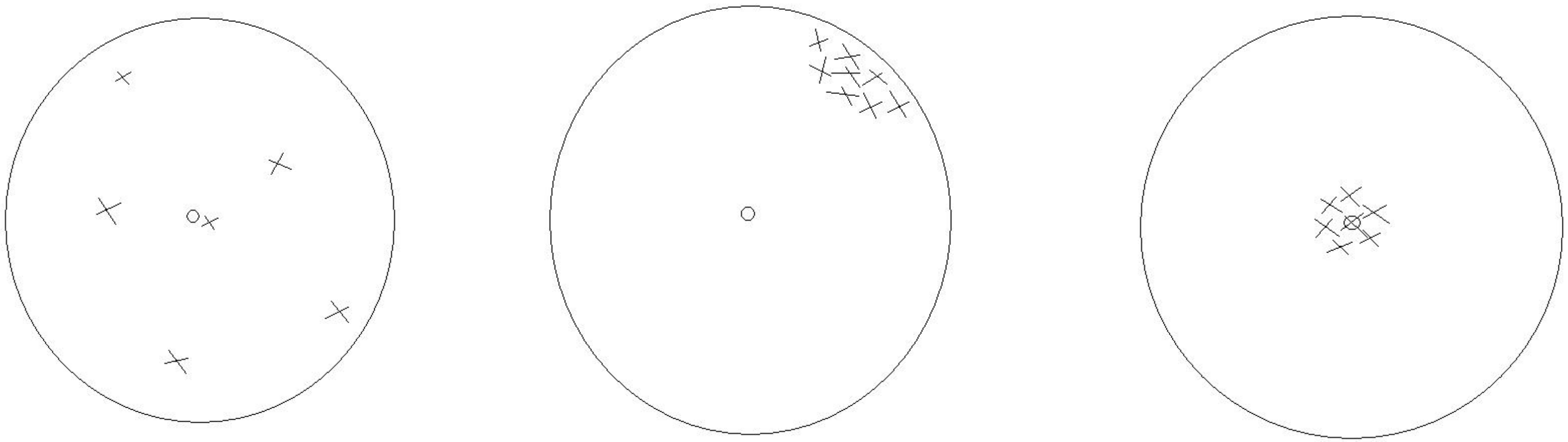
Topics:

- Understand the bias-variance tradeoff in model selection
- Become familiar with model selection criteria
- Understand when/how to use selection algorithms such as stepwise and best subsets
- Understand how to validate a model and measure prediction error

Problems: have a set of predictor variables, how do you select a subset of these that is in some way "best" for predicting the response?

- Subset size, how many explanatory variables should be used to construct the regression model
- Given the subset size, which variables should we choose?

Figure 1: Unbiased, precise and accurate archers

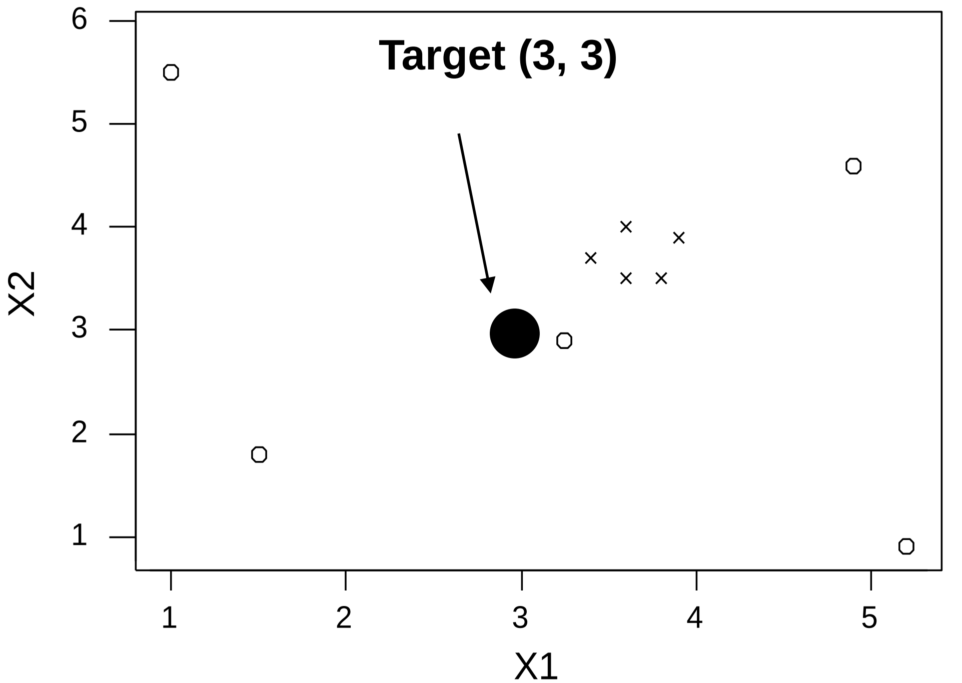


Archer A: unbiased **Archer B: precise** **Archer C: accurate**

Bias-Variance Tradeoff

Row	Gun 1 X1	Gun 1 X2	Gun 2 X1	Gun 2 X2
1	1.00	5.5	3.4	3.7
2	1.50	1.8	3.6	4.0
3	3.25	2.9	3.6	3.5
4	4.90	4.6	3.9	3.9
5	5.20	0.9	3.8	3.5

Gun 1 = circles
Gun 2 = crosses



Which gun is more accurate?

Which is more precise?

Bias-Variance Tradeoff

1. Accuracy corresponds to bias
2. Precision corresponds to variance

On average, Gun 1 hits the target (small or zero bias)
Gun 2 is always close to its average (small variance)

The best gun will have both high accuracy and precision.

Now back to statistics

Instead of choosing a gun, we're choosing an estimator—a statistic or a regression model for prediction

Our targets are the population values:

- **E{Y}-------------estimator based on what model?**
- **E{b_1}------------- what model, what estimator?**
- **E{s^2}, etc. ------what model, what estimator?**

Let's agree that we want our estimator of any parameter, on average, to be close to the true value.

Criterion: Mean Squared Error

Estimator is $\hat{Y}_i$, target is E{$Y_i|X_i$}—true mean at X_i, μ_i.

Error is: $\hat{Y}_i - \mu_i$

Mean (Expected) squared error of $\hat{Y}_i$ ---MSE{$\hat{Y}_i$} is:

$$E\{\hat{Y}_i - \mu_i\}^2$$

Famous, hot, big-time result:

$$E\{\hat{Y}_i - \mu_i\}^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\}$$

MSE = squared bias plus variance

= "accuracy" plus "precision"

Criterion: Mean Square Error

Justification of result: just add and subtract $E\{\hat{Y}_i\}$:

$$(\hat{Y}_i - \mu_i)^2 = [(E\{\hat{Y}_i\} - \mu_i) + (\hat{Y}_i - E\{\hat{Y}_i\})]^2$$

and then square the term and take expectation:

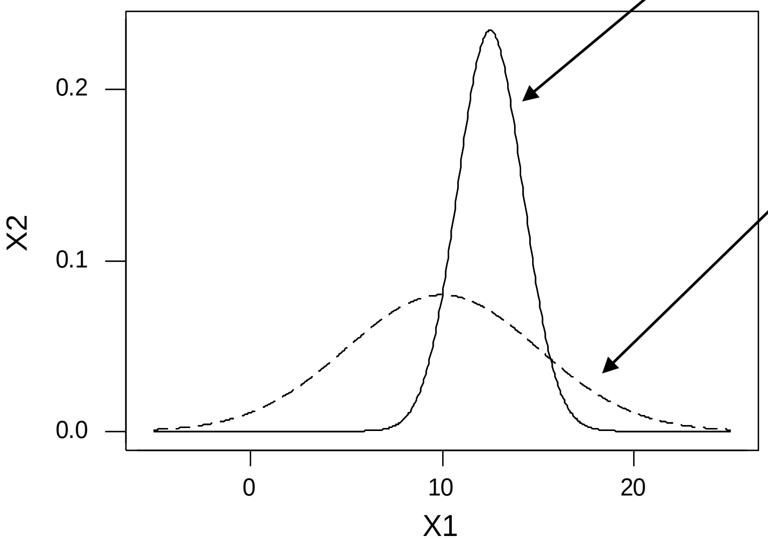
$$E\{\hat{Y}_i - \mu_i\}^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\}$$

So what's this got to do with regression?

Goal: Predict response at X_h
(We secretly know $E\{Y|\,X_h\}= 10$)

Gun 1: $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$ ---N(10, 25)

Gun 2: $\hat{Y} = b'_0 + b'_1X_1 + b'_2X_2$ -----N(12.5, 3)



Which model (gun) is better?

In terms of squared bias:

In terms of variance:

In terms of MSE:

Why Eliminate Unimportant Predictors?

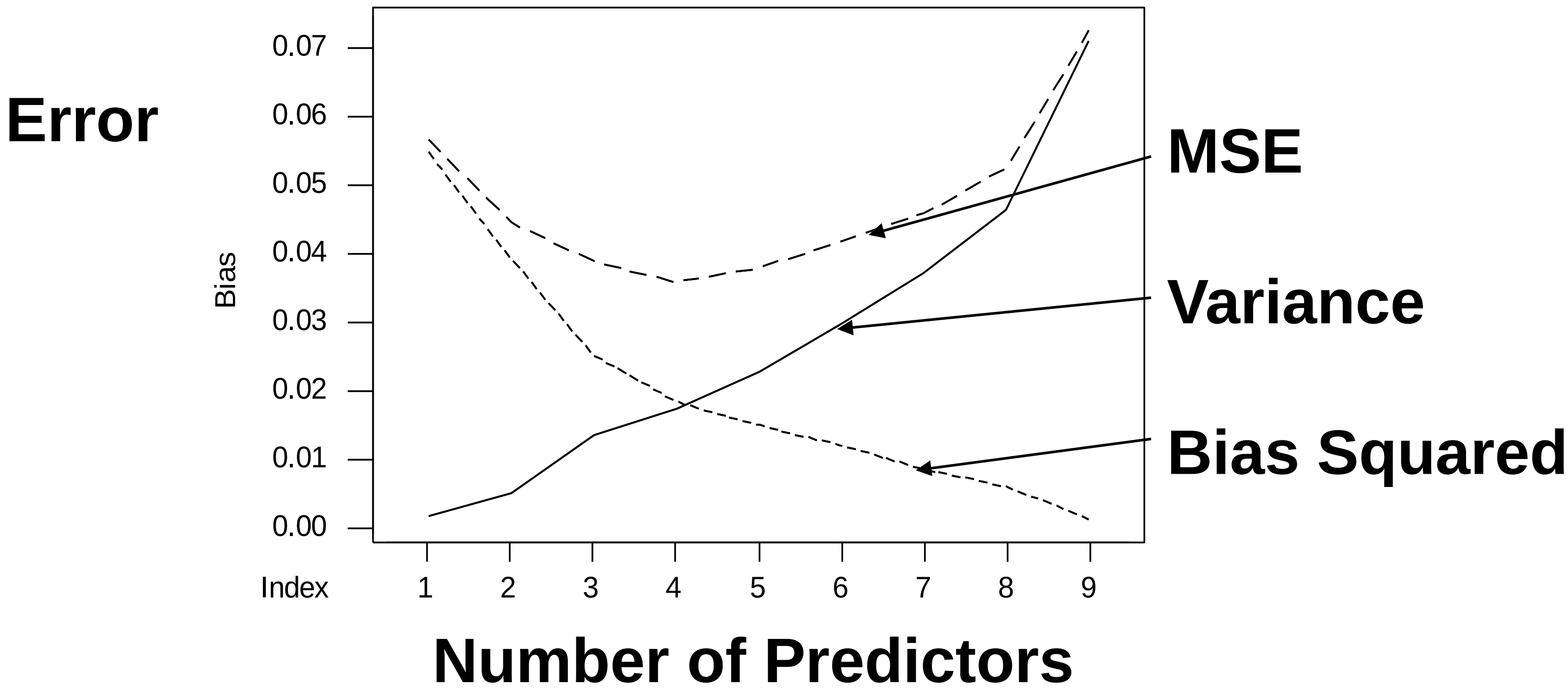
Often smaller models will have smaller MSE!

Depends on:

1. Size of true coefficients, b_i
2. Degree of multicollinearity

   So selecting a best model balances the increase in squared bias of smaller models against the increase in variance for larger models

Picture: Best Model has 4 Predictors



Notations:

- $P-1$: total possible number of predictor variables
- $p-1$: number of predictor variables selected in a regression model, p is the number of parameters in the model.
- $p-1 \leq P-1, n > p$
- For any set of $p-1$ predictors, 2^{p-1} alternative models can be constructed, including the one with no X variables.

Criteria for Model Selection

1. R_p^2 or SSE_p Criterion

- R_p^2 is the coefficient of Multiple Determination for model with $p-1$ predictors
- $R_p^2 = 1 - SSE_p/SSTO$
- Plot R_p^2 v.s $p-1$, R_p^2 will increase as $p-1$ increases.
- The R_p^2 plot will tend to level off at some point. Take the model to be the one where there is no more "meaningful" increase in R_p^2.
- A drawback to R^2 is that the addition of any variable to the model (significant or not) will increase R^2.

2. $R^2_{a,p}$ or MSE_p Criterion

$$R^2_{a,p} = 1 - \frac{SSE_p/(n-p)}{SSTO/(n-1)}$$

$$= 1 - \frac{MSE_p}{SSTO/(n-1)}$$

- $R^2_{a,p}$ increases if and only if MSE_p decreases. This is the same as using MSE.
- Select the subset with the largest $R^2_{a,p}$

3. Mallow's C_p criterion

- Mallow's criterion tries to find the model that minimizes

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} E[(\hat{y}_i - E(y_i))^2]$$

- Mallows found an estimate for this criterion called C_p with

$$C_p = \frac{SSE_p}{MSE_{(Full)}} - (n - 2p).$$

The full model is good at prediction, but if there is multicollinearity, our interpretations of the parameter estimates may not makes sense. A subset model is good if there is not substantial bias in the predicted values (relative to the full model). The C_p criterion looks at the ratio of error SS for the model with p variables to the MSE of the full model, then adds a penalty for the number of variables. SSE_p is based on a specific choice of $p-1$ predictors; while $MSE_{(Full)}$ is based on the full set of variables.

- Adequately fitted model have $C_p \approx p$. Models with lack of fit have $C_p > p$. In considering possible models we would generally consider any subset with $C_p \leq p$.

- Select as the "best" subset, the one with the smallest C_p value.

4. $PRESS_p$ Criterion

- $PRESS_p$ (Prediction Sums of Squares) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses y_i.
- The error sum of squares, $SSE = \sum(y_i - \hat{y}_i)^2$ is also such a measure.
- The $PRESS$ measure differs from SSE in that each fitted value $\hat{y}_i$ for the $PRESS$ criterion is obtained by deleting the ith case from the data set, estimating the regression function for the subset model from the remaining $n-1$ cases, and then using the regression function to obtain the predicted value $\hat{y}_{i(i)}$ for the $i$the case.

$$PRESS_p = \sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2$$

- Models with a small PRESS statistic are considered good candidates.

5. AIC_p and SBC_p

These criteria are motivated from information theory (AIC) and from Bayesian statistics (SBC). They are Criterions based on log(likelihood) plus a penalty for more complexity. We want to choose models that minimize AIC and SBC.

$$AIC_p = n\ln SSE_p - n\ln n + 2p$$

$$SBC_p = n\ln SSE_p - n\ln n + [\ln(n)]p$$

Comments

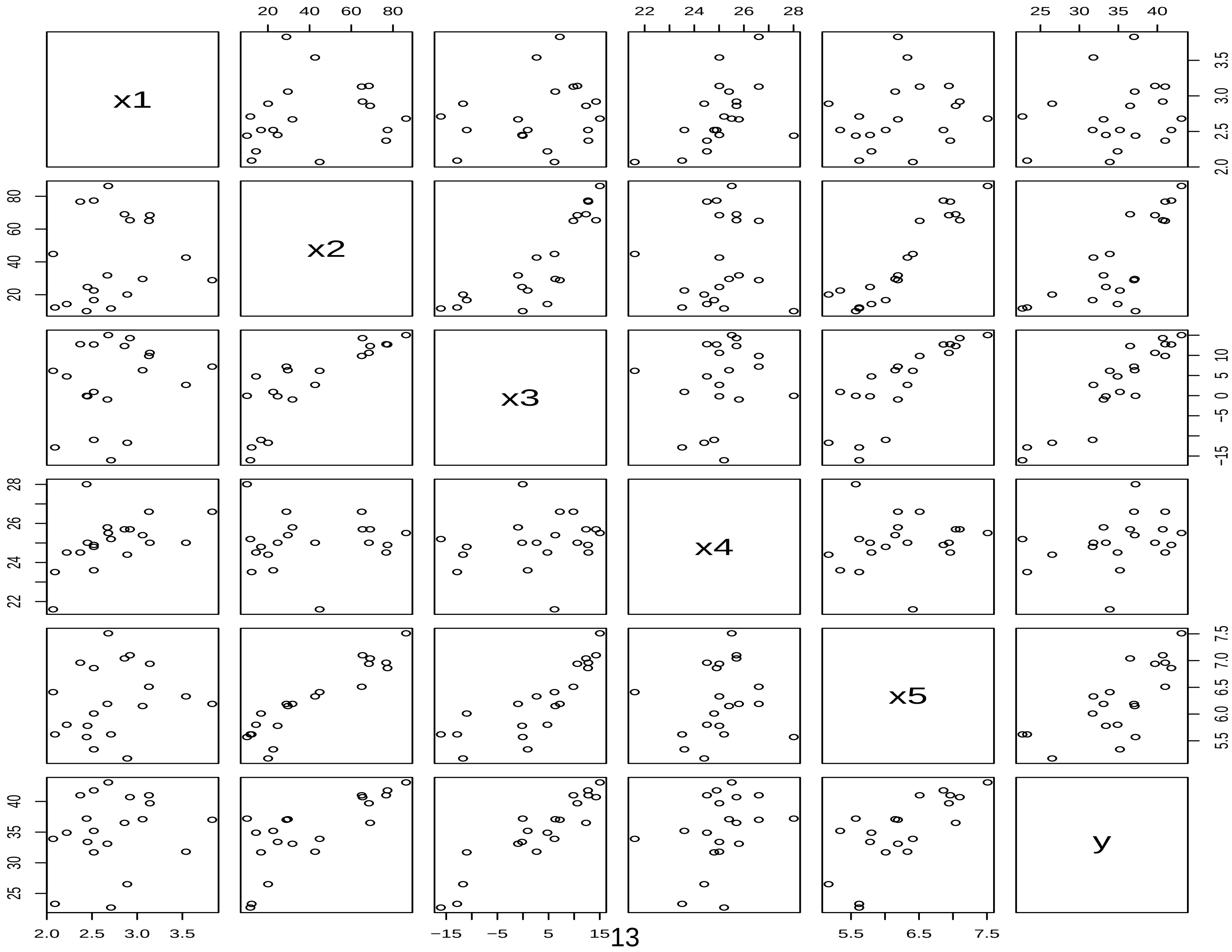
- The different criteria will not always give the identical answer.
- The all subsets method is good for identifying a collection of possible models. One should not necessarily use the model that is declared “best” by any method.
- There might be several subsets that provide a good fit. The final selection of a model should involve residual analysis and knowledge of the subject matter.

Example:

Coleman report data (Christensen)

- y: the mean verbal test score for sixth graders
- x_1: staff salaries per pupil
- x_2: percentage of sixth graders whose fathers have white collar job
- x_3: a composite measure of socioeconomic status
- x_4: the mean of verbal test scores given to the teachers
- x_5: the mean educational level of the sixth grader's mothers (one unit equals two school years)

Figure 2: Scatterplot of Coleman report data



Correlation between y and the predictor variables.

	x_1	x_2	x_3	x_4	x_5
Correlation with y	0.192	0.753	0.927	0.334	0.733

- Of the five variables, x_3 has the highest correlation. It explains more of the y variable than any other single variable.
- x_2 and x_5 also have reasonably high correlations with y.
- Low correlations exist between y and both x_1 and x_4

Table 1: Selection by different criteria

Vars	R^2	AdjR^2	C_p	$\sqrt{MSE}$	x_1	x_2	x_3	x_4	x_5
1	86.0	85.2	5.0	2.2392			×		
1	56.8	54.4	48.6	3.9299		×			
1	53.7	51.2	53.1	4.0654					×
2	88.7	87.4	2.8	2.0641			×	×	
2	86.2	84.5	6.7	2.2866			×		×
2	86.0	84.4	6.9	2.2993		×	×		
3	90.1	88.2	2.8	1.9974	×		×	×	
3	88.9	86.8	4.6	2.1137			×	×	×
3	88.7	86.6	4.8	2.1272		×	×	×	
4	90.2	87.6	4.7	2.0514	×		×	×	×
4	90.1	87.5	4.8	2.0603	×	×	×	×	
4	89.2	86.3	6.1	2.1499		×	×	×	×
5	90.6	87.3	6.0	2.0743	×	×	×	×	×

Example: Hospital was interested in understanding factors that affect survival time following a liver operation. $n = 108$, 54 were held out for validation studies (to be discussed later).

- Y: log of survival time
- X_1: blood clotting score
- X_2: prognostic index
- X_3: enzyme function test score
- X_4: liver function text score

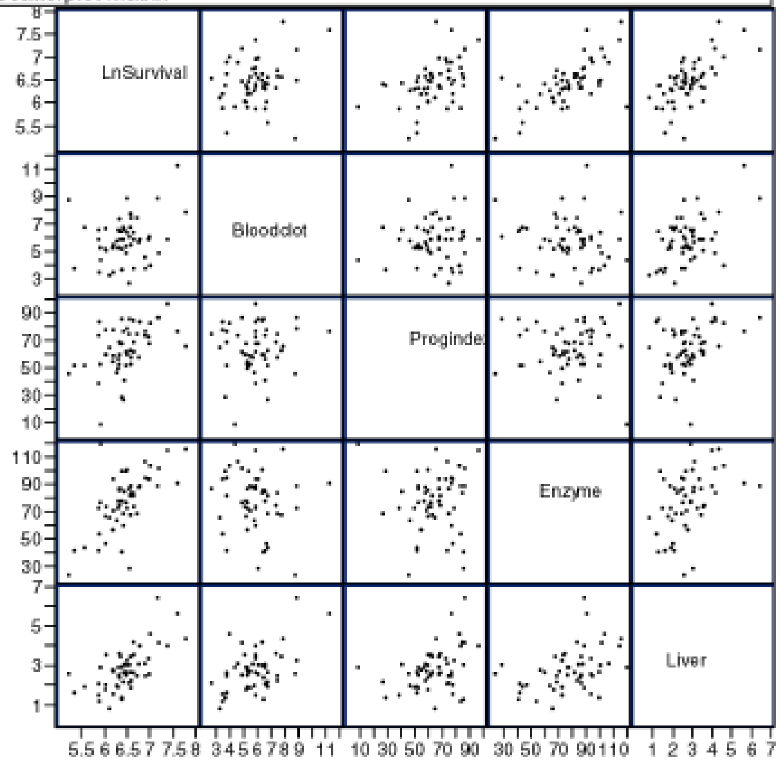
Surgical Unit Example with 4 Predictors

Multivariate

Correlations

	LnSurvival	Bloodclot	Progindex	Enzyme	Liver
LnSurvival	1.0000	0.2462	0.4699	0.6539	0.6493
Bloodclot	0.2462	1.0000	0.0901	-0.1496	0.5024
Progindex	0.4699	0.0901	1.0000	-0.0236	0.3690
Enzyme	0.6539	-0.1496	-0.0236	1.0000	0.4164
Liver	0.6493	0.5024	0.3690	0.4164	1.0000

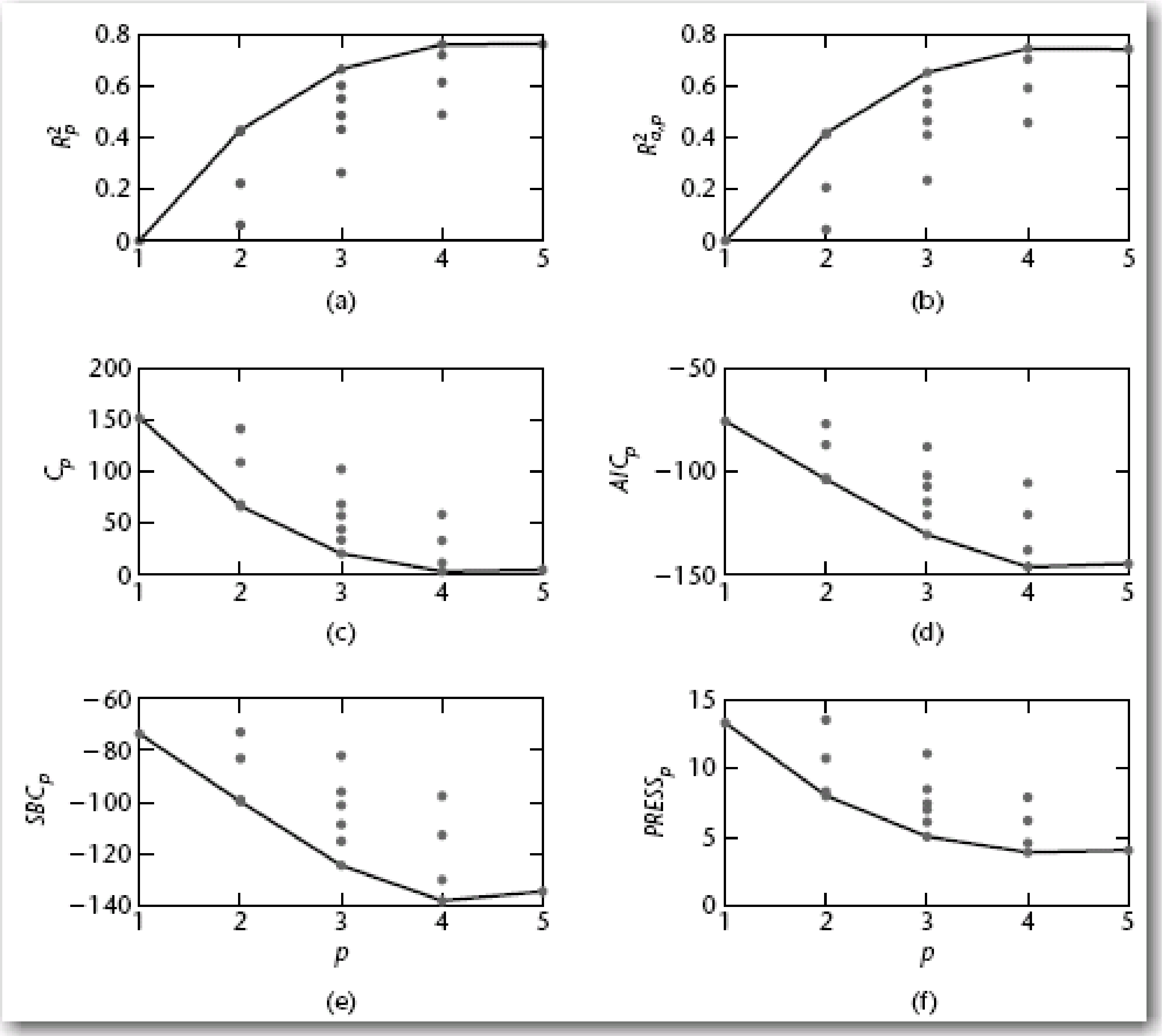
Scatterplot Matrix



Surgical Unit Example with 4 Predictors

X Variables in Model	(1) p	(2) SSE_p	(3) R_p^2	(4) $R_{a,p}^2$	(5) C_p	(6) AIC_p	(7) SBC_p	(8) $PRESS_p$
None	1	12.808	0.000	0.000	151.498	−75.703	−73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	−77.079	−73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	−87.178	−83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	−103.827	−99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	−103.262	−99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	−88.162	−82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	−114.658	−108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	−102.067	−96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	−130.483	−124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	−107.324	−101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	−121.113	−115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	0.743	3.391	−146.161	−138.205	3.914
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	−105.748	−97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	−120.844	−112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	−138.023	−130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	−144.590	−134.645	4.069

Surgical Unit Example with 4 Predictors



Comments:

- As the number of predictors increases, the number of possible models blows up! We need clever computer algorithms to find the really good models.

Two approaches:

- If $p - 1$ is less than 30, use best subsets procedures: These algorithms can use clever search paths to find all of the top models without having to evaluate all $2^{(p-1)}$ possible models.

- If $p - 1$ is greater than 30, use stepwise procedures: These are “greedy” algorithms that first find the best single term model. Given that term, add the next best term, and so on.

Stepwise Regression analysis

- A computationally available method for subset selection
- Evaluate the variables one at a time and look at a sequence of models
- Backwards elimination (start with full model)
- Forward elimination (start with intercept model)
- Stepwise methods (variables can be both added and deleted)

Backwards elimination

- Begins with the full model and sequentially eliminates from the model the least important variable. Importance of the variable is judged by the size t or F statistic.

- $$F_i^* = \frac{MSR(x_i|x_1, \cdots, x_{p-1} \quad \text{except} \quad x_i)}{MSE(x_1, x_2, \cdots, x_{p-1})}, \quad \text{for} \quad i = 1, 2, \cdots, p-1.$$

  Find the smallest F_i^*, If the smallest $F_i^* < F - out$ (predetermined value), remove x_i.

- After the variable with the smallest F statistic is dropped, the model is refitted and the F statistic is recalculated. Again, the variable with the smallest F statistic is dropped

- Process ends when all of F statistics are greater than some predetermined level (predetermined value can change depending on the step).

Table 2: Backwards elimination of y on 5 predictors with $n = 20$, coleman data, predetermined value is 2

Step		const	x_1	x_2	x_3	x_4	x_5	R^2	$\sqrt{MSE}$
1	$\hat{\beta}$	19.95	-1.8	0.044	0.556	1.11	-1.8	90.63	2.07
	t_{obs}		-1.45	0.82	5.98	2.56	-0.89		
2	$\hat{\beta}$	15.47	-1.7		0.582	1.03	-0.5	90.18	2.05
	t_{obs}		-1.41		6.75	2.46	-0.41		
3	$\hat{\beta}$	12.12	-1.7		0.553	1.04		90.07	2.00
	t_{obs}		-1.47		11.27	2.56			
4	$\hat{\beta}$	14.58			0.542	0.75		88.73	2.06
	t_{obs}				10.82	2.05			

Forward selection

- Begins with an initial model (could be intercept only) and adds variables to the model one at a time. Importance of the variable is judged by the size t or F statistic.

- $$F_k^* = \frac{MSR(x_k)}{MSE(x_k)}$$
  enter the variable with the largest F_k^* provided this $F_k^* > F-IN$ (predetermined value) or the corresponding P-value is less than a predetermined α

- One variable in the regression equation, say x_h. Compute all two variable regression equation between y and x_h and x_k for $k \neq h$, calculate
  $$F_k^* = \frac{MSR(x_k|x_h)}{MSE(x_k, x_h)},$$
  enter the variable with the largest F_k^* value provided this $F_k^* > F - IN$

- Procedure ends when none of the F statistic is greater than a predetermined level.

Table 3: Forward selection of y on 5 predictors with $n = 20$, coleman data, predetermined value is 2

Step		const	x_1	x_2	x_3	x_4	x_5	R^2	$\sqrt{MSE}$
1	$\hat{\beta}$	33.32			0.560			85.96	2.24
	t_{obs}				10.50				
4	$\hat{\beta}$	14.58			0.542	0.75		88.73	2.06
	t_{obs}				10.82	2.05			

Stepwise methods

- Alternate between forward selection and backwards elimination
- Arrive at model by dropping a variable, check to see if any variable can be added to the model
- Arrive at a model by adding a variable, check to see if any variable can be dropped
- The value of the F statistic required for dropping a variable is allowed to be different from the value required for adding a variable
- Usually start with an initial model that contains only an intercept
- Stepwise methods gives the same result as forward selection if starting from an initial model; gives the same result as backward elimination if starting from a full model for coleman data

Stepwise methods:

- Step 1: No variable in the regression equation, compute all one variable regression equation between y and $p-1$ predictors and calculate

$$F_k^* = \frac{MSR(x_k)}{MSE(x_k)}$$

enter the variable with the largest F_k^* provided this $F_k^* > F-IN$ (predetermined value) or the corresponding P-value is less than a predetermined α

- Step 2: 1 variable in the regression equation, say x_{k1}. Compute all two variable regression equation between y and x_{k1} and x_k for $k \neq k_1$, calculate

$$F_k^* = \frac{MSR(x_k|x_{k1})}{MSE(x_k, x_{k1})},$$

enter the variable with the largest F_k^* value provided this $F_k^* > F-IN$ (predetermined value) or the corresponding P-value is less than a predetermined α

- Step 3, two variables in regression equation, say x_{k1} and x_{k2}. Determine if any of the variables previously entered should be removed from the regression equation due to the addition of the latest variable.

  —Calculate

$$F_{k1}^{*} = \frac{MSR(x_{k1}|x_{k2})}{MSE(x_{k1}, x_{k2})}$$

  —If the F_{k1}^{*} falls below a predetermined value called F-out or the corresponding P-value is greater than a predetermined α, then x_{k1} is removed from the model

- Suppose there are $r-1$ variables in the regression equation, compute

$$F_k^* = \frac{MSR(x_k|x_{k1}, x_{k2}, \cdots, x_{k,r-1})}{MSE(x_k, x_{k1}, \cdots, x_{k,r-1})}$$

enter the variable with the largest F_k^* value provided $F_k^* > F-in$

—Suppose x_{kr} is added at the above step, compute

$$F_{ki}^* = \frac{MSR(x_{ki}|x_{k1}, \cdots, x_{kr} \quad \text{except} \quad x_{ki})}{MSE(x_{k1}, x_{k2}, \cdots, x_{kr})},$$

$$\text{for} \quad i = 1, 2, \cdots, r-1,$$

find the smallest F_{ki}^*, If the smallest $F_{ki}^* < F-out$, then remove x_{ki} from the equation.

- Go to next step to try to enter another variable, keep gong until no new variable can be entered.

Model selection and case deletion

- Outliers tend to be cases with large residuals

  —-eliminating the largest residuals obviously makes the SSE and MSE smaller

- Variable selection methods tend to identify as good reduced models those with small MSEs

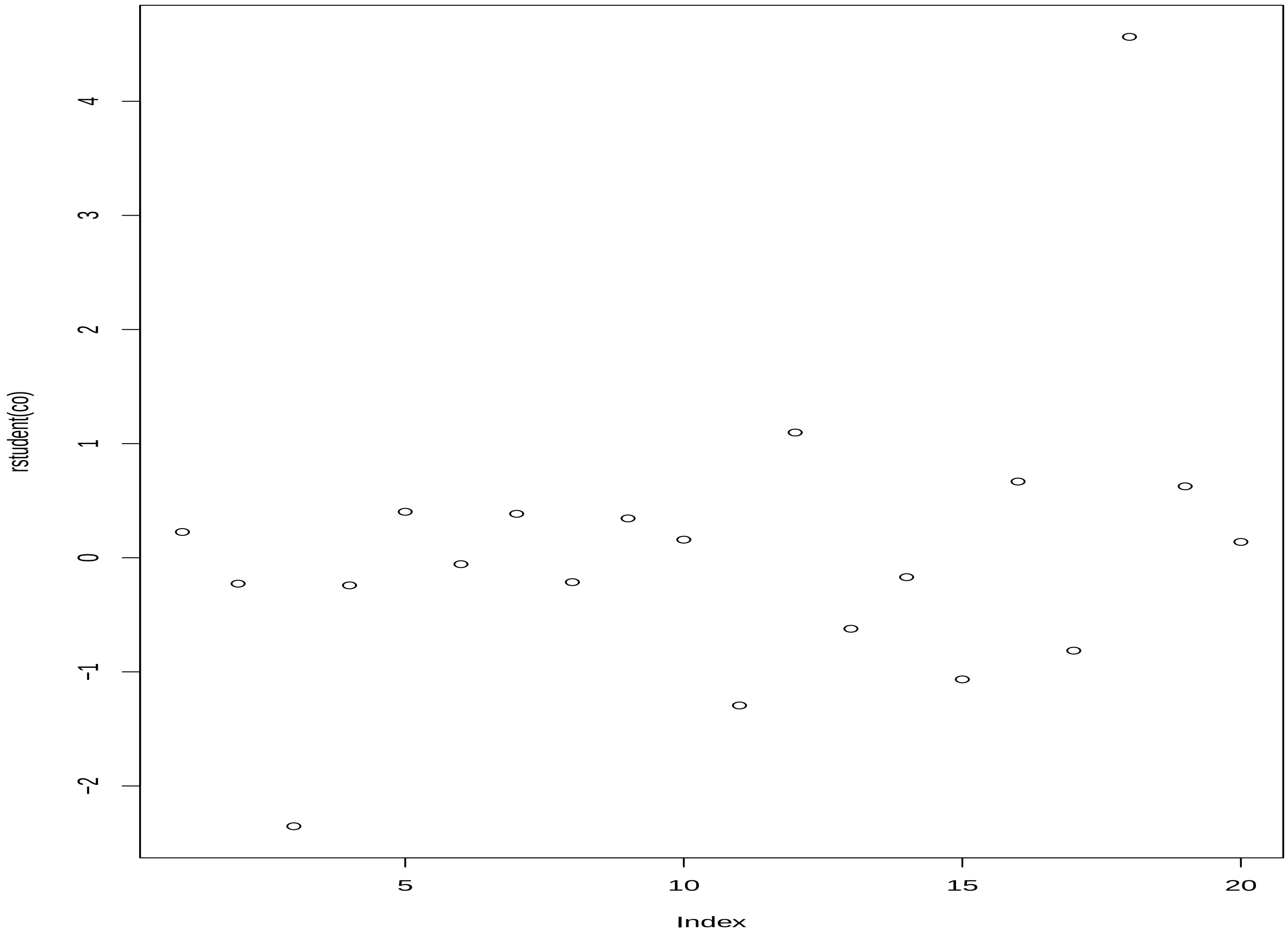
  —-Delete outliers if they are from recording errors (such as obvious typos), experimental accident (drop the tube) etc,.

  —-Usually after deleting outliers, new data will produce new outliers

Example: Coleman data.

```
> outlierTest(co)
   rstudent unadjusted p-value Bonferonni p
18 4.564631         0.00053079     0.010616
```

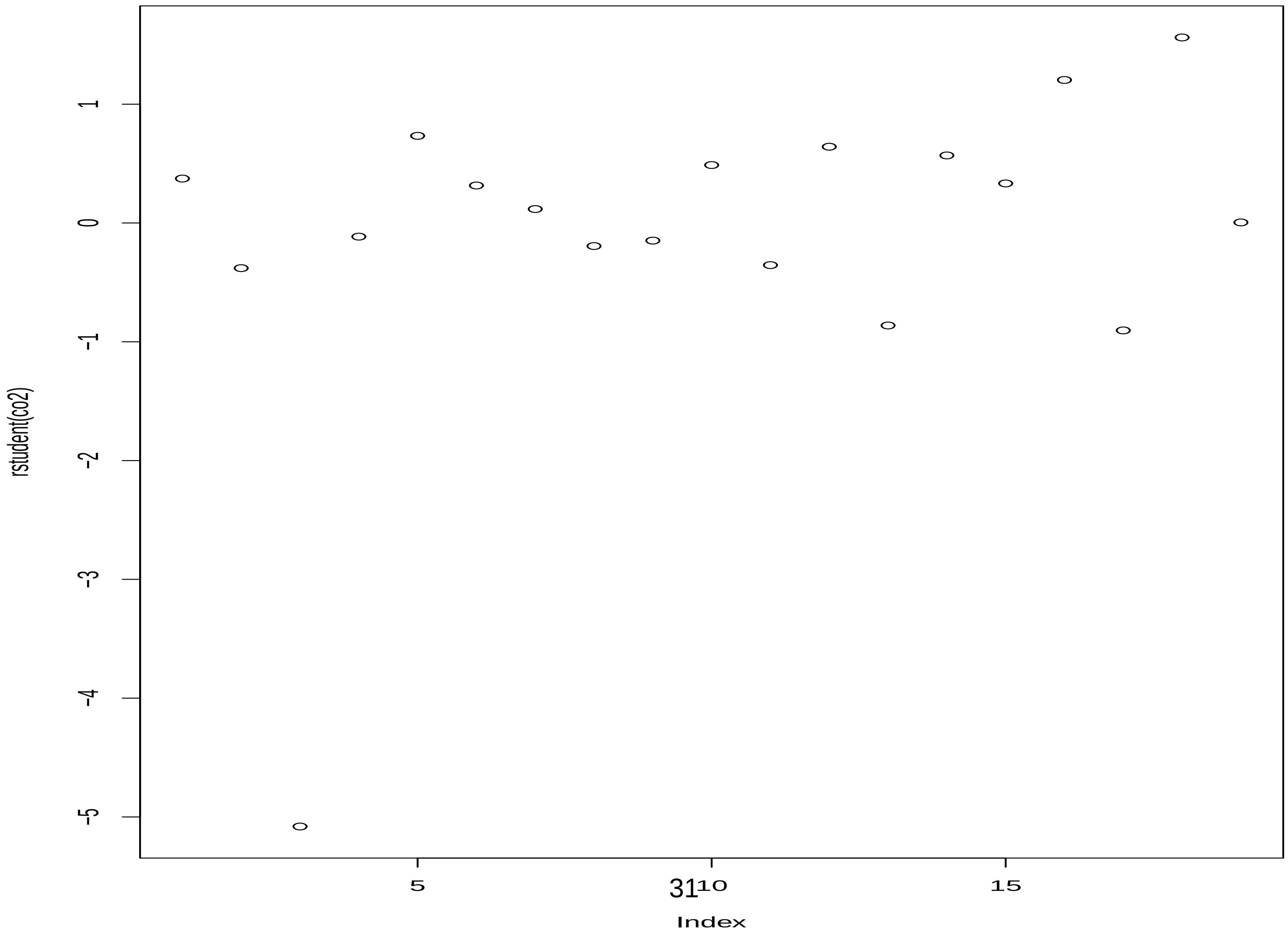
Figure 3: rstudent of Coleman report data, case #18 was identified as an outlier



After case #18 been deleted, case #3 becomes a new outlier

```
> outlierTest(co2)
  rstudent unadjusted p-value Bonferonni p
3 -5.08053          0.00027041    0.0051379
```

Figure 4: Plot of rstudent of Coleman report data after case #18 been deleted, case # 3 was identified as an outlier



Both variable selection and case deletion

- Cause the resulting model to appear better than it probably should
- Tend to give MSEs that are unrealistically small
- Prediction intervals are unrealistically narrow and test statistics are unrealistically large
- Test performed after variable selection or outlier deletion should be viewed as the greatest reasonable evidence against the null hypothesis, with the understanding that more appropriate tests would probably display a lower level of significance.

Example: Coleman data, case 18 deleted

- Case 18 was identified as an influential point
- After case 18 deleted, the full model is the best model as measured by either the C_p statistic or the adjusted R^2 value.
- This is a far cry from the full data analysis in which the models with x_3, x_4 and with x_1, x_3, x_4 had the smallest C_p statistics. After deleting case 18, models x_3, x_4 and x_1, x_3, x_4 are only the seventh and fifth best models.

Table 4: Best subset regression

Vars	R^2	AdjR^2	C_p	$\sqrt{MSE}$	x_1	x_2	x_3	x_4	x_5
1	86.0	85.2	5.0	2.2392			$\times$		
1	56.8	54.4	48.6	3.9299		$\times$			
1	53.7	51.2	53.1	4.0654					$\times$
2	88.7	87.4	2.8	2.0641			$\times$	$\times$	
2	86.2	84.5	6.7	2.2866			$\times$		$\times$
2	86.0	84.4	6.9	2.2993		$\times$	$\times$		
3	90.1	88.2	2.8	1.9974	$\times$		$\times$	$\times$	
3	88.9	86.8	4.6	2.1137			$\times$	$\times$	$\times$
3	88.7	86.6	4.8	2.1272		$\times$	$\times$	$\times$	
4	90.2	87.6	4.7	2.0514	$\times$		$\times$	$\times$	$\times$
4	90.1	87.5	4.8	2.0603	$\times$	$\times$	$\times$	$\times$	
4	89.2	86.3	6.1	2.1499		$\times$	$\times$	$\times$	$\times$
5	90.6	87.3	6.0	2.0743	$\times$	$\times$	$\times$	$\times$	$\times$

Table 5: Best subset regression: Case 18 deleted

Vars	R^2	Adjusted R^2	C_p	$\sqrt{MSE}$	x_1	x_2	x_3	x_4	x_5
1	89.6	89.0	21.9	1.9653			$\times$		
1	56	53.4	140.8	4.0397		$\times$			
1	53.4	50.6	150.2	4.1596					$\times$
2	92.3	91.3	14.3	1.7414			$\times$	$\times$	
2	91.2	90.1	18.2	1.8635			$\times$		$\times$
2	89.8	88.6	23.0	2.0020		$\times$	$\times$		
3	93.7	92.4	11.4	1.6293			$\times$	$\times$	$\times$
3	93.5	92.2	12.1	1.6573	$\times$		$\times$	$\times$	
3	92.3	90.8	16.1	1.7942		$\times$	$\times$	$\times$	
4	95.2	93.8	8.1	1.4766		$\times$	$\times$	$\times$	$\times$
4	94.7	93.2	9.8	1.5464	$\times$		$\times$	$\times$	$\times$
4	93.5	91.6	14.1	1.7143	$\times$	$\times$	$\times$	$\times$	
5	96.3	94.9	6.0	1.3343	$\times$	$\times$	$\times$	$\times$	$\times$

Model Selection Techniques Only Narrow the Field

Final choice of a model based on:

- p-values, residual plots, other diagnostics
- Parsimony (Occam's Razor): Simple models work best
- The sniff (giggle) test: does the model agree with expectations or theory? Do the signs make sense? Can you explain the results?
- Model validation studies

Model Validation

- The real test of a model or theory: How well does the model predict future observations?

- Problem with your model: the residuals are closer to the observations than they should be! So MSE is too small!!!!
—-Why? Because picked the model that best predicts your data set. Your measure of predictive ability is biased.

- Optimism Principle: A model chosen by some selection process provides a more optimistic explanation of data used in its derivation than it does of other data that will arise in a similar fashion.

Getting an unbiased view

- Way 1: Collect n^* new observations and compute the mean squared prediction error:

$$\text{MSPR} = \frac{\sum_{i=1}^{n^*}(y_i - \hat{y}_i)^2}{n^*}$$

—y_i is the response variable in the ith validation case

—$\hat{y}_i$ is the predicted value for the ith validation case based on the model building data set

—n^* is the number of cases in the validation data set.

- Way 2: Cross-validation

—Keep n^* cases out of the data set (at random!).

—Base regression on the $n - n^*$ cases in the training set.

—Computer the MSPR for the n^* cases in the validation set (or test set).

—Usually $n^* \approx n/2$.

- Way 3: K-fold cross-validation (sample size n is small)

— Break data into K roughly equal parts.

—Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K-1$ subsamples are used as training data.

—The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data.

—The K results from the folds can then be averaged to produce a single estimation.

—When $K = n$, the K-fold cross-validation estimate is identical to leave one out cross-validation.

Example: pages 373, 374

In the surgical unit example (utilize all 8 predictors), three models were favored by the various model-selection criteria.

Model 1: favored by SBC_p and $PRESS_p$ criteria:

$$y_i' = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_8 x_{i8} + \epsilon_i$$

Model 2: favored by C_p criterion:

$$y_i' = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_5 x_{i5} + \beta_8 x_{i8} + \epsilon_i$$

Model 3: favored by $R^2_{a,p}$ and AIC_p criteria.

$$y_i' = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_8 x_{i8} + \epsilon_i$$

Table 6: Some results for Models 1-3 based on model-building and validation data set–Surgical Unit Example

Statistic	Model 1 (training)	Model 1 (validation)	Model 2 (training)	Model 2 (validation)	Model 3 (training)	Model 3 (validation)
SSE_p	2.1788	3.7951	2.0820	3.7288	2.0052	3.6822
$PRESS_p$	2.7378	4.5219	2.7827	4.6536	2.7723	4.8981
MSE_p	0.0445	0.0775	0.0434	0.0777	0.0427	0.0783
$MSPR$	0.0773		0.0764		0.0794	

- $PRESS_p$ value is always larger than SSE_p because the regression fit for the ith case when this case is deleted in fitting can never be as good as that when the ith case is included.
- A $PRESS_p$ value reasonably close to SSE_p supports the validity of the fitted regression model and of MSE_p as an indicator of the predictive capability of this model.
- All three of the candidate models have $PRESS_p$ values that are reasonably close to SSE_p.

Table 7: Some results for Models 1-3 based on model-building and validation data set–Surgical Unit Example

Statistic	Model 1 (training)	Model 1 (validation)	Model 2 (training)	Model 2 (validation)	Model 3 (training)	Model 3 (validation)
SSE_p	2.1788	3.7951	2.0820	3.7288	2.0052	3.6822
$PRESS_p$	2.7378	4.5219	2.7827	4.6536	2.7723	4.8981
MSE_p	0.0445	0.0775	0.0434	0.0777	0.0427	0.0783
$MSPR$	0.0773		0.0764		0.0794	

- MSPR for the 54 cases in the validation data set for each of the three models are 0.0773, 0.0764, and 0.0794.
- The mean squared prediction error generally will be larger than MSE_p based on the training data set because entirely new data are involved in the validation data set.
- The fact that $MSPR$ does not differ too greatly from MSE_p implies that the error mean square MSE_p based on the training data set is a reasonably valid indicator of the predictive ability of the fitted regression model.
- The closeness of the three MSPR values suggest that the three candidate models perform comparably in terms of predictive accuracy.

Select a Model

- A review of Table 9.4 in the textbook shows that most of the estimated coefficients agree quite closely, however, for Model 3
  —$b_5 = -0.0035$ (the coefficient of age) for the training data
  —$b_5 = 0.0025$ for the validation data.

- This is certainly a cause for concern, and it raises doubts about the validity of Model 3. Model 3 was eliminated from further consideration.

- The final selection was based on the principle of parsimony. While Model 1 and 2 performed comparably in the validation study. Model 1 achieves this level of performance with one fewer parameter. For this reason, Model 1 was ultimately chosen by the investigator as the final model.