

# Multiple Linear Regression in Data Mining

## **Contents**

- 2.1. A Review of Multiple Linear Regression
- 2.2. Illustration of the Regression Process
- 2.3. Subset Selection in Linear Regression

Perhaps the most popular mathematical model for making predictions is the multiple linear regression model. You have already studied multiple regression models in the “Data, Models, and Decisions” course. In this note we will build on this knowledge to examine the use of multiple linear regression models in data mining applications. Multiple linear regression is applicable to numerous data mining situations. Examples are: predicting customer activity on credit cards from demographics and historical activity patterns, predicting the time to failure of equipment based on utilization and environment conditions, predicting expenditures on vacation travel based on historical frequent flier data, predicting staffing requirements at help desks based on historical data and product and sales information, predicting sales from cross selling of products from historical information and predicting the impact of discounts on sales in retail outlets.

In this note, we review the process of multiple linear regression. In this context we emphasize (a) the need to split the data into two categories: the training data set and the validation data set to be able to validate the multiple linear regression model, and (b) the need to relax the assumption that errors follow a Normal distribution. After this review, we introduce methods for identifying subsets of the independent variables to improve predictions.

## 2.1 A Review of Multiple Linear Regression

In this section, we review briefly the multiple regression model that you encountered in the DMD course. There is a continuous random variable called the dependent variable,  $Y$ , and a number of independent variables,  $x_1, x_2, \dots, x_p$ . Our purpose is to predict the value of the dependent variable (also referred to as the response variable) using a linear function of the independent variables. The values of the independent variables (also referred to as predictor variables, regressors or covariates) are known quantities for purposes of prediction, the model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon, \quad (2.1)$$

where  $\varepsilon$ , the “noise” variable, is a Normally distributed random variable with mean equal to zero and standard deviation  $\sigma$  whose value we do not know. We also do not know the values of the coefficients  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ . We estimate all these  $(p + 2)$  unknown values from the available data.

The data consist of  $n$  rows of observations also called cases, which give us values  $y_i, x_{i1}, x_{i2}, \dots, x_{ip}; i = 1, 2, \dots, n$ . The estimates for the  $\beta$  coefficients are computed so as to minimize the sum of squares of differences between the

fitted (predicted) values at the observed values in the data. The sum of squared differences is given by

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2$$

Let us denote the values of the coefficients that minimize this expression by  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ . These are our estimates for the unknown values and are called OLS (ordinary least squares) estimates in the literature. Once we have computed the estimates  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  we can calculate an unbiased estimate  $\hat{\sigma}^2$  for  $\sigma^2$  using the formula:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2 \\ &= \frac{\text{Sum of the residuals}}{\# \text{observations} - \# \text{coefficients}}. \end{aligned}$$

We plug in the values of  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  in the linear regression model (1) to predict the value of the dependent value from known values of the independent values,  $x_1, x_2, \dots, x_p$ . The predicted value,  $\hat{Y}$ , is computed from the equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p.$$

Predictions based on this equation are the best predictions possible in the sense that they will be unbiased (equal to the true values on the average) and will have the smallest expected squared error compared to any unbiased estimates *if* we make the following assumptions:

1. **Linearity** The expected value of the dependent variable is a linear function of the independent variables, i.e.,

$$E(Y|x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p.$$

2. **Independence** The “noise” random variables  $\varepsilon_i$  are independent between all the rows. Here  $\varepsilon_i$  is the “noise” random variable in observation  $i$  for  $i = 1, \dots, n$ .
3. **Unbiasness** The noise random variable  $\varepsilon_i$  has zero mean, i.e.,  $E(\varepsilon_i) = 0$  for  $i = 1, 2, \dots, n$ .
4. **Homoskedasticity** The standard deviation of  $\varepsilon_i$  equals the same (unknown) value,  $\sigma$ , for  $i = 1, 2, \dots, n$ .

5. **Normality** The “noise” random variables,  $\varepsilon_i$ , are Normally distributed.

An important and interesting fact for our purposes is that even if we drop the assumption of normality (Assumption 5) and allow the noise variables to follow arbitrary distributions, these estimates are very good for prediction. We can show that predictions based on these estimates are the best linear predictions in that they minimize the expected squared error. In other words, amongst all linear models, as defined by equation (1) above, the model using the least squares estimates,

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p,$$

will give the smallest value of squared error on the average. We elaborate on this idea in the next section.

The Normal distribution assumption was required to derive confidence intervals for predictions. In data mining applications we have two distinct sets of data: the training data set and the validation data set that are both representative of the relationship between the dependent and independent variables. The training data is used to estimate the regression coefficients  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ . The validation data set constitutes a “hold-out” sample and is not used in computing the coefficient estimates. This enables us to estimate the error in our predictions without having to assume that the noise variables follow the Normal distribution. We use the training data to fit the model and to estimate the coefficients. These coefficient estimates are used to make predictions for each case in the validation data. The prediction for each case is then compared to value of the dependent variable that was actually observed in the validation data. The average of the square of this error enables us to compare different models and to assess the accuracy of the model in making predictions.

## 2.2 Illustration of the Regression Process

We illustrate the process of Multiple Linear Regression using an example adapted from Chatterjee, Hadi and Price from on estimating the performance of supervisors in a large financial organization.

The data shown in Table 2.1 are from a survey of clerical employees in a sample of departments in a large financial organization. The dependent variable is a performance measure of effectiveness for supervisors heading departments in the organization. Both the dependent and the independent variables are totals of ratings on different aspects of the supervisor’s job on a scale of 1 to 5 by 25 clerks reporting to the supervisor. As a result, the minimum value for

each variable is 25 and the maximum value is 125. These ratings are answers to survey questions given to a sample of 25 clerks in each of 30 departments. The purpose of the analysis was to explore the feasibility of using a questionnaire for predicting effectiveness of departments thus saving the considerable effort required to directly measure effectiveness. The variables are answers to questions on the survey and are described below.

- $Y$  Measure of effectiveness of supervisor.
- $X1$  Handles employee complaints
- $X2$  Does not allow special privileges.
- $X3$  Opportunity to learn new things.
- $X4$  Raises based on performance.
- $X5$  Too critical of poor performance.
- $X6$  Rate of advancing to better jobs.

The multiple linear regression estimates as computed by the StatCalc add-in to Excel are reported in Table 2.2. The equation to predict performance is

$$Y = 13.182 + 0.583X1 - 0.044X2 + 0.329X3 - 0.057X4 + 0.112X5 - 0.197X6.$$

In Table 2.3 we use ten more cases as the validation data. Applying the previous equation to the validation data gives the predictions and errors shown in Table 2.3. The last column entitled error is simply the difference of the predicted minus the actual rating. For example for Case 21, the error is equal to  $44.46 - 50 = -5.54$

We note that the average error in the predictions is small (-0.52) and so the predictions are unbiased. Further the errors are roughly Normal so that this model gives prediction errors that are approximately 95% of the time within  $\pm 14.34$  (two standard deviations) of the true value.

## 2.3 Subset Selection in Linear Regression

A frequent problem in data mining is that of using a regression equation to predict the value of a dependent variable when we have a number of variables available to choose as independent variables in our model. Given the high speed of modern algorithms for multiple linear regression calculations, it is tempting

Case	Y	X1	X2	X3	X4	X5	X6
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
8	71	75	50	55	70	66	41
9	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41
11	64	53	53	58	58	67	34
12	67	60	47	39	59	74	41
13	69	62	57	42	55	63	25
14	68	83	83	45	59	77	35
15	77	77	54	72	79	77	46
16	81	90	50	72	60	54	36
17	74	85	64	69	79	79	63
18	65	60	65	75	55	80	60
19	65	70	46	57	75	85	46
20	50	58	68	54	64	78	52

**Table 2.1:** Training Data (20 departments).

in such a situation to take a kitchen-sink approach: why bother to select a subset, just use all the variables in the model. There are several reasons why this could be undesirable.

- It may be expensive to collect the full complement of variables for future predictions.
- We may be able to more accurately measure fewer variables (for example in surveys).
- Parsimony is an important property of good models. We obtain more insight into the influence of regressors in models with a few parameters.

Multiple R-squared			0.656		
Residual SS			738.900		
Std. Dev.	Estimate			7.539	
	Coefficient	StdError	t-statistic	p-value	
Constant	13.182	16.746	0.787	0.445	
X1	0.583	0.232	2.513	0.026	
X2	-0.044	0.167	-0.263	0.797	
X3	0.329	0.219	1.501	0.157	
X4	-0.057	0.317	-0.180	0.860	
X5	0.112	0.196	0.570	0.578	
X6	-0.197	0.247	-0.798	0.439	

**Table 2.2:** Output of StatCalc.

- Estimates of regression coefficients are likely to be unstable due to multicollinearity in models with many variables. We get better insights into the influence of regressors from models with fewer variables as the coefficients are more stable for parsimonious models.
- It can be shown that using independent variables that are uncorrelated with the dependent variable will increase the variance of predictions.
- It can be shown that dropping independent variables that have small (non-zero) coefficients can reduce the average error of predictions.

Let us illustrate the last two points using the simple case of two independent variables. The reasoning remains valid in the general situation of more than two independent variables.

### 2.3.1 Dropping Irrelevant Variables

Suppose that the true equation for  $Y$ , the dependent variable, is:

$$Y = \beta_1 X_1 + \varepsilon \quad (2.2)$$

Case	Y	X1	X2	X3	X4	X5	X6	Prediction	Error
21	50	40	33	34	43	64	33	44.46	-5.54
22	64	61	52	62	66	80	41	63.98	-0.02
23	53	66	52	50	63	80	37	63.91	10.91
24	40	37	42	58	50	57	49	45.87	5.87
25	63	54	42	48	66	75	33	56.75	-6.25
26	66	77	66	63	88	76	72	65.22	-0.78
27	78	75	58	74	80	78	49	73.23	-4.77
28	48	57	44	45	51	83	38	58.19	10.19
29	85	85	71	71	77	74	55	76.05	-8.95
30	82	82	39	59	64	78	39	76.10	-5.90
Averages:								62.38	-0.52
Std Devs:								11.30	7.17

**Table 2.3:** Predictions on the validation data.

and suppose that we estimate  $Y$  (using an additional variable  $X_2$  that is actually irrelevant) with the equation:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \quad (2.3)$$

We use data  $y_i, x_{i1}, x_{i2}, i = 1, 2, \dots, n$ . We can show that in this situation the least squares estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  will have the following expected values and variances:

$$E(\hat{\beta}_1) = \beta_1, \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(1 - R_{12}^2) \sum_{i=1}^n x_{i1}^2}$$

$$E(\hat{\beta}_2) = 0, \quad \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{(1 - R_{12}^2) \sum_{i=1}^n x_{i2}^2},$$

where  $R_{12}$  is the correlation coefficient between  $X_1$  and  $X_2$ .

We notice that  $\hat{\beta}_1$  is an unbiased estimator of  $\beta_1$  and  $\hat{\beta}_2$  is an unbiased estimator of  $\beta_2$ , since it has an expected value of zero. If we use Model (2) we obtain that

$$E(\hat{\beta}_1) = \beta_1, \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_{i1}^2}.$$

Note that in this case the variance of  $\hat{\beta}_1$  is lower.

The variance is the expected value of the squared error for an unbiased estimator. So we are worse off using the irrelevant estimator in making predictions. Even if  $X_2$  happens to be uncorrelated with  $X_1$  so that  $R_{12}^2 = 0$  and the variance of  $\hat{\beta}_1$  is the same in both models, we can show that the variance of a prediction based on Model (3) will be worse than a prediction based on Model (2) due to the added variability introduced by estimation of  $\beta_2$ .

Although our analysis has been based on one useful independent variable and one irrelevant independent variable, the result holds true in general. **It is always better to make predictions with models that do not include irrelevant variables.**

### 2.3.2 Dropping independent variables with small coefficient values

Suppose that the situation is the reverse of what we have discussed above, namely that Model (3) is the correct equation, but we use Model (2) for our estimates and predictions ignoring variable  $X_2$  in our model. To keep our results simple let us suppose that we have scaled the values of  $X_1$ ,  $X_2$ , and  $Y$  so that their variances are equal to 1. In this case the least squares estimate  $\hat{\beta}_1$  has the following expected value and variance:

$$E(\hat{\beta}_1) = \beta_1 + R_{12}\beta_2, \quad Var(\hat{\beta}_1) = \sigma^2.$$

Notice that  $\hat{\beta}_1$  is a biased estimator of  $\beta_1$  with bias equal to  $R_{12}\beta_2$  and its Mean Square Error is given by:

$$\begin{aligned} MSE(\hat{\beta}_1) &= E[(\hat{\beta}_1 - \beta_1)^2] \\ &= E\{[\hat{\beta}_1 - E(\hat{\beta}_1) + E(\hat{\beta}_1) - \beta_1]^2\} \\ &= [Bias(\hat{\beta}_1)]^2 + Var(\hat{\beta}_1) \\ &= (R_{12}\beta_2)^2 + \sigma^2. \end{aligned}$$

If we use Model (3) the least squares estimates have the following expected values and variances:

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1, & Var(\hat{\beta}_1) &= \frac{\sigma^2}{(1 - R_{12}^2)}, \\ E(\hat{\beta}_2) &= \beta_2, & Var(\hat{\beta}_2) &= \frac{\sigma^2}{(1 - R_{12}^2)}. \end{aligned}$$

Now let us compare the Mean Square Errors for predicting  $Y$  at  $X_1 = u_1, X_2 = u_2$ .

For Model (2), the Mean Square Error is:

$$\begin{aligned}
 MSE2(\hat{Y}) &= E[(\hat{Y} - Y)^2] \\
 &= E[(u_1\hat{\beta}_1 - u_1\beta_1 - \varepsilon)^2] \\
 &= u_1^2 MSE2(\hat{\beta}_1) + \sigma^2 \\
 &= u_1^2(R_{12}\beta_2)^2 + u_1^2\sigma^2 + \sigma^2
 \end{aligned}$$

For Model (3), the Mean Square Error is:

$$\begin{aligned}
 MSE3(\hat{Y}) &= E[(\hat{Y} - Y)^2] \\
 &= E[(u_1\hat{\beta}_1 + u_2\hat{\beta}_2 - u_1\beta_1 - u_2\beta_2 - \varepsilon)^2] \\
 &= Var(u_1\hat{\beta}_1 + u_2\hat{\beta}_2) + \sigma^2, \quad \text{because now } \hat{Y} \text{ is unbiased} \\
 &= u_1^2 Var(\hat{\beta}_1) + u_2^2 Var(\hat{\beta}_2) + 2u_1u_2 Covar(\hat{\beta}_1, \hat{\beta}_2) \\
 &= \frac{(u_1^2 + u_2^2 - 2u_1u_2R_{12})}{(1 - R_{12}^2)}\sigma^2 + \sigma^2.
 \end{aligned}$$

Model (3) can lead to lower mean squared error for many combinations of values for  $u_1, u_2, R_{12}$ , and  $(\beta_2/\sigma)^2$ . For example, if  $u_1 = 1, u_2 = 0$ , then  $MSE2(\hat{Y}) < MSE3(\hat{Y})$ , when

$$(R_{12}\beta_2)^2 + \sigma^2 < \frac{\sigma^2}{(1 - R_{12}^2)},$$

i.e., when

$$\frac{|\beta_2|}{\sigma} < \frac{1}{\sqrt{1 - R_{12}^2}}.$$

If  $\frac{|\beta_2|}{\sigma} < 1$ , this will be true for all values of  $R_{12}^2$ ; if, however, say  $R_{12}^2 > .9$ , then this will be true for  $|\beta|/\sigma < 2$ .

In general, accepting some bias can reduce MSE. This Bias-Variance trade-off generalizes to models with several independent variables and is particularly important for large values of the number  $p$  of independent variables, since in that case it is very likely that there are variables in the model that have small coefficients relative to the standard deviation of the noise term and also exhibit at least moderate correlation with other variables. Dropping such variables will improve the predictions as it will reduce the MSE.

This type of Bias-Variance trade-off is a basic aspect of most data mining procedures for prediction and classification.

### 2.3.3 Algorithms for Subset Selection

Selecting subsets to improve MSE is a difficult computational problem for large number  $p$  of independent variables. The most common procedure for  $p$  greater than about 20 is to use heuristics to select “good” subsets rather than to look for the best subset for a given criterion. The heuristics most often used and available in statistics software are step-wise procedures. There are three common procedures: forward selection, backward elimination and step-wise regression.

#### Forward Selection

Here we keep adding variables one at a time to construct what we hope is a reasonably good subset. The steps are as follows:

1. Start with constant term only in subset  $S$ .
2. Compute the reduction in the sum of squares of the residuals (SSR) obtained by including each variable that is not presently in  $S$ . We denote by  $SSR(S)$  the sum of square residuals given that the model consists of the set  $S$  of variables. Let  $\hat{\sigma}^2(S)$  be an unbiased estimate for  $\sigma$  for the model consisting of the set  $S$  of variables. For the variable, say,  $i$ , that gives the largest reduction in SSR compute

$$F_i = \text{Max}_{i \notin S} \frac{SSR(S) - SSR(S \cup \{i\})}{\hat{\sigma}^2(S \cup \{i\})}$$

If  $F_i > F_{in}$ , where  $F_{in}$  is a threshold (typically between 2 and 4) add  $i$  to  $S$

3. Repeat 2 until no variables can be added.

#### Backward Elimination

1. Start with all variables in  $S$ .
2. Compute the increase in the sum of squares of the residuals (SSR) obtained by excluding each variable that is presently in  $S$ . For the variable, say,  $i$ , that gives the smallest increase in SSR compute

$$F_i = \text{Min}_{i \in S} \frac{SSR(S - \{i\}) - SSR(S)}{\hat{\sigma}^2(S)}$$

If  $F_i < F_{out}$ , where  $F_{out}$  is a threshold (typically between 2 and 4) then drop  $i$  from  $S$ .

3. Repeat 2 until no variable can be dropped.

Backward Elimination has the advantage that all variables are included in  $S$  at some stage. This addresses a problem of forward selection that will never select a variable that is better than a previously selected variable that is strongly correlated with it. The disadvantage is that the full model with all variables is required at the start and this can be time-consuming and numerically unstable.

### Step-wise Regression

This procedure is like Forward Selection except that at each step we consider dropping variables as in Backward Elimination.

Convergence is guaranteed if the thresholds  $F_{out}$  and  $F_{in}$  satisfy:  $F_{out} < F_{in}$ . It is possible, however, for a variable to enter  $S$  and then leave  $S$  at a subsequent step and even rejoin  $S$  at a yet later step.

As stated above these methods pick one best subset. There are straightforward variations of the methods that do identify several close to best choices for different sizes of independent variable subsets.

None of the above methods guarantees that they yield the best subset for any criterion such as adjusted  $R^2$ . (Defined later in this note.) They are reasonable methods for situations with large numbers of independent variables but for moderate numbers of independent variables the method discussed next is preferable.

### All Subsets Regression

The idea here is to evaluate all subsets. Efficient implementations use branch and bound algorithms of the type you have seen in DMD for integer programming to avoid explicitly enumerating all subsets. (In fact the subset selection problem can be set up as a quadratic integer program.) We compute a criterion such as  $R_{adj}^2$ , the adjusted  $R^2$  for all subsets to choose the best one. (This is only feasible if  $p$  is less than about 20).

#### 2.3.4 Identifying subsets of variables to improve predictions

The All Subsets Regression (as well as modifications of the heuristic algorithms) will produce a number of subsets. Since the number of subsets for even moderate values of  $p$  is very large, we need some way to examine the most promising subsets and to select from them. An intuitive metric to compare subsets is  $R^2$ . However since  $R^2 = 1 - \frac{SSR}{SST}$  where  $SST$ , the Total Sum of Squares, is the Sum of Squared Residuals for the model with just the constant term, if we use it as a criterion we will always pick the full model with all  $p$  variables. One approach is therefore to select the subset with the largest  $R^2$  for each possible

size  $k$ ,  $k = 2, \dots, p + 1$ . The size is the number of coefficients in the model and is therefore one more than the number of variables in the subset to account for the constant term. We then examine the increase in  $R^2$  as a function of  $k$  amongst these subsets and choose a subset such that subsets that are larger in size give only insignificant increases in  $R^2$ .

Another, more automatic, approach is to choose the subset that maximizes,  $R_{adj}^2$ , a modification of  $R^2$  that makes an adjustment to account for size. The formula for  $R_{adj}^2$  is

$$R_{adj}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2).$$

It can be shown that using  $R_{adj}^2$  to choose a subset is equivalent to picking the subset that minimizes  $\hat{\sigma}^2$ .

Table 2.4 gives the results of the subset selection procedures applied to the training data in the Example on supervisor data in Section 2.2.

Notice that the step-wise method fails to find the best subset for sizes of 4, 5, and 6 variables. The Forward and Backward methods do find the best subsets of all sizes and so give identical results as the All subsets algorithm. The best subset of size 3 consisting of {X1, X3} maximizes  $R_{adj}^2$  for all the algorithms. This suggests that we may be better off in terms of MSE of predictions if we use this subset rather than the full model of size 7 with all six variables in the model. Using this model on the validation data gives a slightly higher standard deviation of error (7.3) than the full model (7.1) but this may be a small price to pay if the cost of the survey can be reduced substantially by having 2 questions instead of 6. This example also underscores the fact that we are basing our analysis on small (tiny by data mining standards!) training and validation data sets. Small data sets make our estimates of  $R^2$  unreliable.

A criterion that is often used for subset selection is known as Mallows's  $C_p$ . This criterion assumes that the full model is unbiased although it may have variables that, if dropped, would improve the  $MSE$ . With this assumption we can show that if a subset model is unbiased  $E(C_p)$  equals  $k$ , the size of the subset. Thus a reasonable approach to identifying subset models with small bias is to examine those with values of  $C_p$  that are near  $k$ .  $C_p$  is also an estimate of the sum of MSE (standardized by dividing by  $\sigma^2$ ) for predictions (the fitted values) at the x-values observed in the training set. Thus good models are those that have values of  $C_p$  near  $k$  and that have small  $k$  (i.e. are of small size).  $C_p$  is computed from the formula:

$$C_p = \frac{SSR}{\hat{\sigma}_{Full}^2} + 2k - n,$$

<b>SST=</b> 2149.000		<i>Fin=</i> 3.840									
		<i>Fout=</i> 2.710									
<i>Forward, backward, and all subsets selections</i>											
					<b>Models</b>						
<b>Size</b>	<b>SSR</b>	<b>RSq</b>	<b>RSq (adj)</b>	<b>Cp</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
2	874.467	0.593	0.570	-0.615	Constant X1						
3	786.601	0.634	0.591	-0.161	Constant X1 X3						
4	759.413	0.647	0.580	1.361	Constant X1 X3 X6						
5	743.617	0.654	0.562	3.083	Constant X1 X3 X5 X6						
6	740.746	0.655	0.532	5.032	Constant X1 X2 X3 X5 X6						
7	738.900	0.656	0.497	7.000	Constant X1 X2 X3 X4 X5 X6						
<i>Stepwise Selection</i>											
					<b>Models</b>						
<b>Size</b>	<b>SSR</b>	<b>RSq</b>	<b>RSq (adj)</b>	<b>Cp</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
2	874.467	0.593	0.570	-0.615	Constant X1						
3	786.601	0.634	0.591	-0.161	Constant X1 X3						
4	783.970	0.635	0.567	1.793	Constant X1 X2 X3						
5	781.089	0.637	0.540	3.742	Constant X1 X2 X3 X4						
6	775.094	0.639	0.511	5.637	Constant X1 X2 X3 X4 X5						
7	738.900	0.656	0.497	7.000	Constant X1 X2 X3 X4 X5 X6						

**Table 2.4:** Subset Selection for the example in Section 2.2

where  $\hat{\sigma}_{Full}^2$  is the estimated value of  $\sigma^2$  in the full model that includes all the variables. It is important to remember that the usefulness of this approach depends heavily on the reliability of the estimate of  $\sigma^2$  for the full model. This requires that the training set contains a large number of observations relative to the number of variables. We note that for our example only the subsets of size 6 and 7 seem to be unbiased as for the other models  $C_p$  differs substantially from  $k$ . This is a consequence of having too few observations to estimate  $\sigma^2$  accurately in the full model.