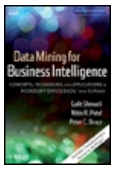

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Second Edition

by Galit Shmueli, Nitin R. Patel and Peter C. Bruce

Chapter 6: Multiple Linear Regression

In this chapter we introduce linear regression models for the purpose of prediction. We discuss the differences between fitting and using regression models for the purpose of inference (as in classical statistics) and for prediction. A predictive goal calls for evaluating model performance on a validation set and for using predictive metrics. We then raise the challenges of using many predictors and describe variable selection algorithms that are often implemented in linear regression procedures.

6.1 Introduction

The most popular model for making predictions is the *multiple linear regression model* encountered in most introductory statistics classes and textbooks. This model is used to fit a linear relationship between a quantitative *dependent variable Y* (also called the *outcome* or *response variable*) and a set of *predictors* $X_1, X_2, \ldots, X_p$ (also referred to as *independent variables, input variables, regressors*, or *covariates*). The assumption is that in the population of interest, the following relationship holds:

$$(6.1)\quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

where $\beta_0, \ldots, \beta_p$ are *coefficients* and ε is the *noise* or *unexplained* part. The data, which are a sample from this population, are then used to estimate the coefficients and the variability of the noise.

The two popular objectives behind fitting a model that relates a quantitative outcome with predictors are for understanding the relationship between these factors and for predicting the outcomes of new cases. The classical statistical approach has focused on the first objective: fitting the best model to the data in an attempt to learn about the underlying relationship in the population. In data mining, however, the focus is typically on the second goal: predicting new observations. Important differences between the approaches stem from the fact that in the classical statistical world we are interested in drawing conclusions from a limited supply of data and in learning how reliable those conclusions might be. In data mining, by contrast, data are typically plentiful, so the performance and reliability of our model can easily be established by applying it to fresh data.

Multiple linear regression is applicable to numerous data mining situations. Examples are predicting customer activity on credit cards from their 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 flyer 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. Although a linear regression model is used for both goals, the modeling step and performance assessment differ depending on the goal. Therefore, the choice of model is closely tied to whether the goal is explanatory or predictive.

6.2 Explanatory versus Predictive Modeling

Both explanatory and predictive modeling involve using a dataset to fit a model (i.e., to estimate coefficients), checking model validity, assessing its performance, and comparing to other models. However, there are several major differences between the two:

1. A good explanatory model is one that fits the data closely, whereas a good predictive model is one that predicts new cases accurately.
2. In explanatory models (classical statistical world, scarce data) the entire dataset is used for estimating the best-fit model, to maximize the amount of information that we have about the hypothesized relationship in the population. When the goal is to predict outcomes of new cases (data mining, plentiful data), the data are typically split into a training set and a validation set. The training set is used to estimate the model, and the validation, or *holdout*, set is used to assess this model's performance on new, unobserved data.
3. Performance measures for explanatory models measure how close the data fit the model (how well the model approximates the data), whereas in predictive models performance is measured by predictive accuracy (how well the model predicts new cases).

For these reasons it is extremely important to know the goal of the analysis before beginning the modeling process. A good predictive model can have a looser fit to the data on which it is based, and a good explanatory model can have low prediction accuracy. In the remainder of this chapter we focus on predictive models because these are more popular in

data mining and because most textbooks focus on explanatory modeling.

6.3 Estimating the Regression Equation and Prediction

The coefficients β_0 β_p and the standard deviation of the noise (σ) determine the relationship in the population of interest. Since we only have a sample from that population, these coefficients are unknown. We therefore estimate them from the data using a method called *ordinary least squares* (OLS). This method finds values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ that minimize the sum of squared deviations between the actual values (Y) and their predicted values based on that model ($\hat{Y}$).

To predict the value of the dependent value from known values of the predictors, x_1, x_2, …, x_p, we use sample estimates for β_0, …, β_p in the linear regression model (6.1) since β_0, …, β_p cannot be observed directly unless we have available the entire population of interest. The predicted value, $\hat{Y}$, is computed from the equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$. Predictions based on this equation are the best predictions possible in the sense that they will be unbiased (equal to the true values on average) and will have the smallest average squared error compared to any unbiased estimates *if* we make the following assumptions:

1. The noise e (or equivalently, the dependent variable) follows a normal distribution.
2. The linear relationship is correct.
3. The cases are independent of each other.
4. The variability in Y values for a given set of predictors is the same regardless of the values of the predictors *(homoskedasticity)*.

An important and interesting fact for the predictive goal is that *even if we drop the first assumption and allow the noise to follow an arbitrary distribution, these estimates are very good for prediction*, in the sense that among all linear models, as defined by equation (6.1), the model using the least-squares estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$, will have the smallest average squared errors. An assumption of a normal distribution is required in the classical implementation of multiple linear regression to derive confidence intervals for predictions. In this classical world, data are scarce and the same data are used to fit the regression model and to assess its reliability (with confidence limits). In data mining applications we have two distinct sets of data: The training dataset and the validation dataset are both representative of the relationship between the dependent and independent variables. The training data is used to fit the model and estimate the regression coefficients β_0, β_1, …, β_p. The validation dataset constitutes a holdout sample and is not used in computing the coefficient estimates. The estimates are then used to make predictions for each case in the validation data. This enables us to estimate the error in our predictions by using the validation set without having to assume that the noise follows a normal distribution. The prediction for each case is then compared to the 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 prediction accuracy of the model.

Example: Predicting the Price of Used Toyota Corolla Automobiles

A large Toyota car dealership offers purchasers of new Toyota cars the option to buy their used car as part of a trade-in. In particular, a new promotion promises to pay high prices for used Toyota Corolla cars for purchasers of a new car. The dealer then sells the used cars for a small profit. To ensure a reasonable profit, the dealer needs to be able to predict the price that the dealership will get for the used cars. For that reason, data were collected on all previous sales of used Toyota Corollas at the dealership. The data include the sales price and other information on the car, such as its age, mileage, fuel type, and engine size. A description of each of these variables is given in Table 6.1. A sample of this dataset is shown in Table 6.2.

Table 6.1: VARIABLES IN THE TOYOTA COROLLA EXAMPLE

Variable	Description
Price	Offer price in euros
Age	Age in months as of August 2004

Kilometers	Accumulated kilometers on odometer
Fuel Type	Fuel type (*Petrol, Diesel, CMC*)
HP	Horsepower
Metallic	Metallic color? (Yes = 1, No = 0)
Automatic	Automatic (Yes = 1, No = 0)
CC	Cylinder volume in cubic centimeters
Doors	Number of doors
QuartTax	Quarterly road tax in euros
Weight	Weight in kilograms

Table 6.2: PRICES AND ATTRIBUTES FOR A SAMPLE OF 30 USED TOYOTA COROLLA CARS

Price	Age	Kilometers	Fuel Type	HP	Metallic	Automatic	CC	Doors	Quart Tax	Weigh
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185
19950	22	43610	Petrol	192	0	0	1800	3	100	1185
19600	25	32189	Petrol	192	0	0	1800	3	100	1185
21500	31	23000	Petrol	192	1	0	1800	3	100	1185
22500	32	34131	Petrol	192	1	0	1800	3	100	1185
22000	28	18739	Petrol	192	0	0	1800	3	100	1185
22750	30	34000	Petrol	192	1	0	1800	3	100	1185
17950	24	21716	Petrol	110	1	0	1600	3	85	1105
16750	24	25563	Petrol	110	0	0	1600	3	19	1065
16950	30	64359	Petrol	110	1	0	1600	3	85	1105
15950	30	67660	Petrol	110	1	0	1600	3	85	1105
16950	29	43905	Petrol	110	0	1	1600	3	100	1170
15950	28	56349	Petrol	110	1	0	1600	3	85	1120
16950	28	32220	Petrol	110	1	0	1600	3	85	1120
16250	29	25813	Petrol	110	1	0	1600	3	85	1120
15950	25	28450	Petrol	110	1	0	1600	3	85	1120
17495	27	34545	Petrol	110	1	0	1600	3	85	1120
15750	29	41415	Petrol	110	1	0	1600	3	85	1120
11950	39	98823	CNG	110	1	0	1600	5	197	1119

The total number of records in the dataset is 1000 cars (we use the first 1000 cars from the dataset ToyotoCorolla.xls). After partitioning the data into training and validation sets (at a 60%: 40% ratio), we fit a multiple linear regression model between price (the dependent variable) and the other variables (as predictors) using the training set only. Figure 6.1 shows the estimated coefficients (as computed by XLMiner). Notice that the Fuel Type predictor has three categories (Petrol,

Diesel, and CNG), and we therefore have two dummy variables in the model [e.g., Petrol (0/1) and Diesel (0/1); the third, CNG (0/1), is redundant given the information on the first two dummies]. These coefficients are then used to predict prices of used Toyota Corolla cars based on their age, mileage, and so on. Figure 6.2 shows a sample of 20 of the predicted prices for cars in the validation set, using the estimated model. It gives the predictions and their errors (relative to the actual prices) for these 20 cars. On the right we get overall measures of predictive accuracy. Note that the average error is $111. A boxplot of the residuals (Figure 6.3) shows that 50% of the errors are approximately between ±$850. This might be small relative to the car price but should be taken into account when considering the profit. Such measures are used to assess the predictive performance of a model and to compare models. We discuss such measures in the next section. This example also illustrates the point about the relaxation of the normality assumption. A histogram or probability plot of prices shows a right-skewed distribution. In a classical modeling case where the goal is to obtain a good fit to the data, the dependent variable would be transformed (e.g., by taking a natural log) to achieve a more "normal" variable. Although the fit of such a model to the training data is expected to be better, it will not necessarily yield a significant predictive improvement. In this example the average error in a model of log(price) is —$160, compared to $111 in the original model for price.

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-2327.281494	1622.562866	0.15210986	81481950000
Age	-134.137619	4.77474403	0	5888770000
Mileage	-0.0199055	0.00236949	0	172544200
Fuel_Type_Diesel	129.2410126	536.7660523	0.80982733	2427870
Fuel_Type_Petrol	2670.873291	520.0211792	0.0000004	670008.4375
Horse_Power	33.95512009	5.37533283	0	339071900
Metalic_Color	-38.04909897	120.321022	0.75196105	716922.5
Automatic	224.9384003	269.0696716	0.40356547	10970180
CC	0.0209207	0.0959821	0.8275463	1553226
Doors	-3.00326943	61.79518509	0.96125734	17263280
Quarterly_Tax	22.90351105	2.48583364	0	221851400
Weight	12.9385519	1.51249933	0	136067800

Residual df	588
Multiple R-squared	0.861344575
Std. Dev. estimate	1363.600464
Residual SS	1093331000

FIGURE 6.1: ESTIMATED COEFFICIENTS FOR REGRESSION MODEL OF PRICE VS. CAR ATTRIBUTES

Predicted Value	Actual Value	Residual
16199	13750	-2449
16686	13950	-2736
16266	16900	634
16236	18600	2364
20534	20950	416
20520	19600	-920
19860	21500	1640
19504	22500	2996
20385	22000	1615
16993	16950	-43
16106	16950	844
16099	16250	151
15789	15750	-39
15590	15950	360
15660	14950	-710
15668	14750	-918
15300	16750	1450
17919	19000	1081
17242	17950	708
19148	21950	2802

(*a*)

Total sum of squared errors	RMS Error	Average Error
795600925.2	1410.319933	110.9145714

(*b*)

FIGURE 6.2: *(A)* PREDICTED PRICES (AND ERRORS) FOR 20 CARS IN VALIDATION SET AND *(B)* SUMMARY PREDICTIVE MEASURES FOR ENTIRE VALIDATION SET

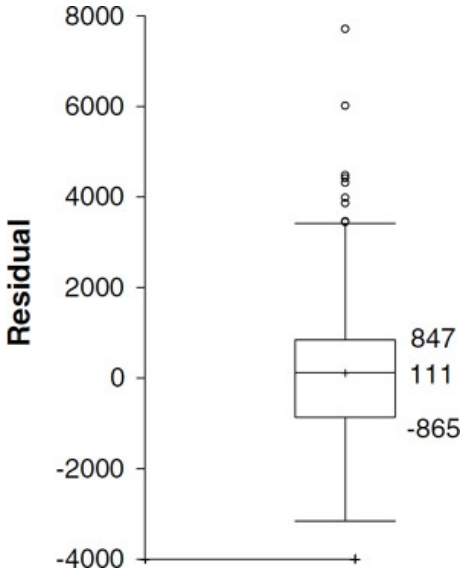

FIGURE 6.3: BOXPLOT OF MODEL RESIDUALS (BASED ON VALIDATION SET)

6.4 Variable Selection in Linear Regression

Reducing the Number of Predictors

A frequent problem in data mining is that of using a regression equation to predict the value of a dependent variable when we have many variables available to choose as predictors in our model. Given the high speed of modern algorithms for multiple linear regression calculations, it is tempting 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 or not feasible to collect a full complement of predictors for future predictions.
- We may be able to measure fewer predictors more accurately (e.g., in surveys).
- The more predictors there are, the higher the chance of missing values in the data. If we delete or impute cases with missing values, multiple predictors will lead to a higher rate of case deletion or imputation.
- *Parsimony* is an important property of good models. We obtain more insight into the influence of predictors in models with few parameters.
- Estimates of regression coefficients are likely to be unstable, due to *mul-ticollinearity* in models with many variables. (Multicollinearity is the presence of two or more predictors sharing the same linear relationship with the outcome variable.) Regression coefficients are more stable for parsimonious models. One very rough rule of thumb is to have a number of cases n larger than $5(p + 2)$, where p is the number of predictors.
- It can be shown that using predictors that are uncorrelated with the dependent variable increases the variance of predictions.
- It can be shown that dropping predictors that are actually correlated with the dependent variable can increase the average error (bias) of predictions.

The last two points mean that there is a trade-off between too few and too many predictors. In general, accepting some bias can reduce the variance in predictions. This *bias-variance trade-off* is particularly important for large numbers of predictors because 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 and also exhibit at least moderate correlation with other variables. Dropping such variables will improve the predictions, as it will reduce the prediction variance. This type of bias-variance trade-off is a basic aspect of most data mining procedures for prediction and classification. In light of this, methods for reducing the number of predictors p to a smaller set are often used.

How to Reduce the Number of Predictors

The first step in trying to reduce the number of predictors should always be to use domain knowledge. It is important to

understand what the various predictors are measuring and why they are relevant for predicting the response. With this knowledge the set of predictors should be reduced to a sensible set that reflects the problem at hand. Some practical reasons for predictor elimination are expense of collecting this information in the future, inaccuracy, high correlation with another predictor, many missing values, or simply irrelevance. Also helpful in examining potential predictors are summary statistics and graphs, such as frequency and correlation tables, predictor-specific summary statistics and plots, and missing value counts.

The next step makes use of computational power and statistical significance. In general, there are two types of methods for reducing the number of predictors in a model. The first is an exhaustive search for the "best" subset of predictors by fitting regression models with all the possible combinations of predictors. The second is to search through a partial set of models. We describe these two approaches next.

Exhaustive Search The idea here is to evaluate all 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. Criteria for evaluating and comparing models are based on the fit to the training data. One popular criterion is the *adjusted* R^2, which is defined as

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}(1-R^2),$$

where R^2 is the proportion of explained variability in the model (in a model with a single predictor, this is the squared correlation). Like R^2, higher values of adjusted R^2 indicate better fit. Unlike R^2, which does not account for the number of predictors used, adjusted R^2 uses a penalty on the number of predictors. This avoids the artificial increase in R^2 that can result from simply increasing the number of predictors but not the amount of information. It can be shown that using R^2_{adj} to choose a subset is equivalent to picking the subset that minimizes $\hat{\sigma}^2$.

Another criterion that is often used for subset selection is known as *Mallow's* C_p. This criterion assumes that the full model (with all predictors) is unbiased, although it may have predictors that if dropped would reduce prediction variability. With this assumption we can show that ifa subset model is unbiased, the average C_p value equals the number of parameters p + 1(= number of predictors +1), the size of the subset. So a reasonable approach to identifying subset models with small bias is to examine those with values of C_p that are near p + 1. C_p is also an estimate of the error[1] for predictions at the x values observed in the training set. Thus, good models are those that have values of C_p near p + 1 and that have small p (i.e., are of small size). C_p is computed from the formula

$$C_p = \frac{\text{SSR}}{\hat{\sigma}^2_{\text{full}}} + 2(p+1) - n$$

where σ^2 is the estimated value of σ^2 in the full model that includes all predictors, and SSR is the sum of squares of Regression given in the ANOVA table. It is important to remember that the usefulness of this approach depends heavily on the reliability of the estimate of σ^2 for the full model. This requires that the training set contain a large number of observations relative to the number of pre-dictors. Finally, a useful point to note is that for a fixed size of subset, R^2, R^2_{adj}, and C_p all select the same subset. In fact, there is no difference between them in the order of merit they ascribe to subsets of a fixed size.

Figure 6.4 gives the results of applying an exhaustive search on the Toyota Corolla price data (with the 11 predictors). It reports the best model with a single predictor, 2 predictors, and so on. It can be seen that the R^2_{adj} increases until 6 predictors are used (number of coefficients = 7) and then stabilizes. The C_p indicates that a model with 9-11 predictors is good. The dominant predictor in all models is the age of the car, with horsepower and mileage playing important roles as well.

#Coeffs	RSS	Cp	R-Sq	Adj. R-Sq	Model (Constant present in all models)											
					1	2	3	4	5	6	7	8	9	10	11	12
2	1996467712	477.71	0.75	0.75	Constant	Age										
3	1672546432	305.51	0.79	0.79	Constant	Age	HP									
4	1438242432	181.50	0.82	0.82	Constant	Age	HP	Weight								
5	1258062976	86.59	0.84	0.84	Constant	Age	Mileage	HP	Weight							
6	1181816320	47.59	0.85	0.85	Constant	Age	Mileage	Petrol	QuartTax	Weight						
7	1095153024	2.98	0.86	0.86	Constant	Age	Mileage	Petrol	HP	QuartTax	Weight					
8	1093753344	4.23	0.86	0.86	Constant	Age	Mileage	Petrol	HP	Automatic	QuartTax	Weight				
9	1093557120	6.12	0.86	0.86	Constant	Age	Mileage	Petrol	HP	Metallic	Automatic	QuartTax	Weight			
10	1093422592	8.05	0.86	0.86	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	QuartTax	Weight		
11	1093335424	10.00	0.86	0.86	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	QuartTax	Weight	
12	1093331072	12.00	0.86	0.86	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	Doorsuart	Tax	Weight

FIGURE 6.4: EXHAUSTIVE SEARCH RESULT FOR REDUCING PREDICTORS IN TOYOTA COROLLA EXAMPLE

Popular Subset Selection Algorithms The second method of finding the best subset of predictors relies on a partial, iterative search through the space of all possible regression models. The end product is one best subset of predictors (although there do exist variations of these methods that identify several close-to-best choices for different sizes of predictor subsets). This approach is computationally cheaper, but it has the potential of missing "good" combinations of predictors. None of the methods guarantee that they yield the best subset for any criterion, such as adjusted R^2. They are reasonable methods for situations with large numbers of predictors, but for moderate numbers of predictors the exhaustive search is preferable.

Three popular iterative search algorithms are forward selection, backward elimination, and stepwise regression. In *forward selection* we start with no predictors and then add predictors one by one. Each predictor added is the one (among all predictors) that has the largest contribution to R^2 on top of the predictors that are already in it. The algorithm stops when the contribution of additional predictors is not statistically significant. The main disadvantage of this method is that the algorithm will miss pairs or groups of predictors that perform very well together but perform poorly as single predictors. This is similar to interviewing job candidates for a team project one by one, thereby missing groups of candidates who perform superiorly together, but poorly on their own.

In *backward elimination* we start with all predictors and then at each step eliminate the least useful predictor (according to statistical significance). The algorithm stops when all the remaining predictors have significant contributions. The weakness of this algorithm is that computing the initial model with all predictors can be time consuming and unstable. *Stepwise regression* is like forward selection except that at each step we consider dropping predictors that are not statistically significant, as in backward elimination.

Note In XLMiner, unlike other popular software packages (SAS, Minitab, etc.), these three algorithms yield a table similar to the one that the exhaustive search yields rather than a single model. This allows the user to decide on the subset size after reviewing all possible sizes based on criteria such as R^2_{adj} and C_p.

For the Toyota Corolla price example, forward selection yields exactly the same results as those found in an exhaustive search: For each number of predictors the same subset is chosen (it therefore gives a table identical to the one in Figure 6.4). Notice that this will not always be the case. In comparison, backward elimination starts with the full model and then drops predictors one by one in this order: Doors, CC, Diesel, Metallic, Automatic, QuartTax, Petrol, Weight, and Age (see Figure 6.5). The R^2_{adj} and C_p measures indicate exactly the same subsets as those suggested by the exhaustive search. In other words, it correctly identifies Doors, CC, Diesel, Metallic, and Automatic as the least useful predictors. Backward elimination would yield a different model than that of the exhaustive search only if we decided to use fewer than six predictors. For instance, if we were limited to two predictors, backward elimination would choose Age and Weight, whereas an exhaustive search shows that the best pair of predictors is actually Age and HP.

#Coeffs	RSS	Cp	R-Sq	Adj. R-Sq	Probability	Model (Constant present in all models)										
						1	2	3	4	5	6	7	8	9	10	11
2	1996467712	477.712341	0.74681	0.7463861	0	Constant	Age									
3	1780184064	363.393707	0.77424	0.7734821	0	Constant	Age	Weight								
4	1482806272	205.462128	0.81195	0.8110051	0	Constant	Age	Petrol	Weight							
5	1310214400	114.64119	0.83384	0.8327225	0	Constant	Age	Petrol	QuartTax	Weight						
6	1181816320	47.5879288	0.85012	0.8488613	8E-08	Constant	Age	Mileage	Petrol	QuartTax	Weight					
7	1095153024	2.97988558	0.86111	0.8597082	0.962122	Constant	Age	Mileage	Petrol	HP	QuartTax	Weight				
8	1093753344	4.22712946	0.86129	0.8596509	0.993999	Constant	Age	Mileage	Petrol	HP	Automatic	QuartTax	Weight			
9	1093557120	6.12159872	0.86132	0.8594386	0.989111	Constant	Age	Mileage	Petrol	HP	Metallic	Automatic	QuartTax	Weight		
10	1093422592	8.0492487	0.86133	0.8592177	0.975677	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	QuartTax	Weight	
11	1093335424	10.0023689	0.86134	0.8589899	0.961197	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	uartTax	Weight
12	1093331072	12.0000286	0.86134	0.8587507	1	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	Doors	QuartTax

FIGURE 6.5: BACKWARD ELIMINATION RESULT FOR REDUCING PREDICTORS IN TOYOTA COROLLA

EXAMPLE

The results for stepwise regression can be seen in Figure 6.6. It chooses the same subsets as forward selection for subset sizes of 1-7 predictors. However, for 8-10 predictors, it chooses a different subset than that chosen using the other methods: It decides to drop Doors, Quart Tax, and Weight. This means that it fails to detect the best subsets for 8-10 predictors. R^2_{adj} is largest at 6 predictors (the same 6 as were selected by the other models), but C_p indicates that the full model with 11 predictors is the best fit.

#Coeffs	RSS	Cp	R-Sq	Adj. R-Sq	Probability	Model (Constant present in all models)										
						1	2	3	4	5	6	7	8	9	10	11
2	1996467712	477.712341	0.74681	0.7463861	0	Constant	Age									
3	1672546432	305.505524	0.78789	0.7871783	0	Constant	Age	HP								
4	1438242560	181.495499	0.8176	0.816685	0	Constant	Age	HP	Weight							
5	1258062976	86.5938416	0.84045	0.8393808	0	Constant	Age	Mileage	HP	Weight						
6	1188944640	51.4215813	0.84922	0.8479497	2E-08	Constant	Age	Mileage	HP	QuartTax	Weight					
7	1095153024	2.97988558	0.86111	0.8597082	0.962122	Constant	Age	Mileage	Petrol	HP	QuartTax	Weight				
8	1093753344	4.22712946	0.86129	0.8596509	0.993999	Constant	Age	Mileage	Petrol	HP	Automatic	QuartTax	Weight			
9	1468513408	207.775345	0.81376	0.8112433	0	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC		
10	1451250000	200.491074	0.81595	0.8131461	0	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	Doors	
11	1229398784	83.1780624	0.84409	0.8414415	0	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	Doors	QuartTax
12	1093331072	12.0000286	0.86134	0.8587507	1	Constant	Age	Mileage	Diesel	Petrol	HP	Metallic	Automatic	CC	Doors	QuartTax

FIGURE 6.6: STEPWISE SELECTION RESULT FOR REDUCING PREDICTORS IN TOYOTA COROLLA EXAMPLE

This example shows that the search algorithms yield fairly good solutions, but we need to carefully determine the number of predictors to retain. It also shows the merits of running a few searches and using the combined results to determine the subset to choose. There is a popular (but false) notion that stepwise regression is superior to backward elimination and forward selection because of its ability to add and to drop predictors. This example shows clearly that it is not always so.

Finally, additional ways to reduce the dimension of the data are by using principal components (Chapter 4) and regression trees (Chapter 9).

[1]In particular, it is the sum of the MSE (mean squared error) standardized by dividing by σ^2.

Problems

6.1 **Predicting Boston Housing Prices**. The file BostonHousing.xls contains information collected by the U.S. Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The goal is to predict the median house price in new tracts based on information such as crime rate, pollution, and number of rooms. The dataset contains 14 predictors, and the response is the median house price (MEDV). Table 6.3 describes each of the predictors and the response.

a. Why should the data be partitioned into training and validation sets? For what will the training set be used? For what will the validation set be used?

   Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM.

b. Write the equation for predicting the median house price from the predictors in the model.

c. What median house price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error?

d. Reduce the number of predictors:
   i. Which predictors are likely to be measuring the same thing among the 14 predictors? Discuss the relationships among INDUS, NOX, and TAX.

   ii. Compute the correlation table for the 13 numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multicollinearity. Choose which ones to remove based on this table.

   iii. Use an exhaustive search to reduce the remaining predictors as follows: First, choose the top three models. Then run each of these models separately on the training set, and compare their predictive accuracy for the validation set. Compare RMSE and average error, as well as lift charts. Finally, describe the best model.

Table 6.3: DESCRIPTION OF VARIABLES FOR BOSTON HOUSING EXAMPLE

CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25, 000 ft^2
INDUS	Proportion of nonretail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10, 000
PTRATIO	Pupil/teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% Lower status of the population
MEDV	Median value of owner-occupied homes in $1000s

6.2 **Predicting Software Reselling Profits**. Tayko Software is a software catalog firm that sells games and educational software. It started out as a software manufacturer and then added third-party titles to its offerings. It recently revised its collection of items in a new catalog, which it mailed out to its customers. This mailing yielded 1000 purchases. Based on these data, Tayko wants to devise a model for predicting the spending amount that a purchasing customer will yield. The file Tayko.xls contains information on 1000 purchases. Table 6.4 describes the variables to be used in the problem (the Excel file contains additional variables).

a. Explore the spending amount by creating a pivot table for the categorical variables and computing the average and standard deviation of spending in each category.

b. Explore the relationship between spending and each of the two continuous predictors by creating two scatterplots (SPENDING vs. FREQ, and SPENDING vs. LAST-UPDATE). Does there seem to be a linear relationship?

c. To fit a predictive model for SPENDING:
   i. Partition the 1000 records into training and validation sets.
   ii. Run a multiple linear regression model for SPENDING versus all six predictors. Give the estimated predictive equation.
   iii. Based on this model, what type of purchaser is most likely to spend a large amount of money?
   iv. If we used backward elimination to reduce the number of predictors, which predictor would be dropped first from the model?
   v. Show how the prediction and the prediction error are computed for the first purchase in the validation set.
   vi. Evaluate the predictive accuracy of the model by examining its performance on the validation set.
   vii. Create a histogram of the model residuals. Do they appear to follow a normal distribution? How does this affect the predictive performance of the model?

Table 6.4: DESCRIPTION OF VARIABLES FOR TAYKO SOFTWARE EXAMPLE

FREQ	Number of transactions in the preceding year
LAST-UPDATE	Number of days since last update to customer record
WEB	Whether customer purchased by Web order at least once
GENDER	Male or female
ADDRESS-RES	Whether it is a residential address
ADDRESS-US	Whether it is a U.S. address
SPENDING (response)	Amount spent by customer in test mailing (in dollars)

6.3 **Predicting Airfares on New Routes**. Several new airports have opened in major cities, opening the market for new routes (a route refers to a pair of airports), and Southwest has not announced whether it will cover routes to/from these cities. In order to price flights on these routes, a major airline collected information on 638 air routes in the United States. Some factors are known about these new routes: the distance traveled, demographics of the city where the new airport is located, and whether this city is a vacation destination. Other factors are yet unknown (e.g., the number of passengers who will travel this route). A major unknown factor is whether Southwest or another discount airline will travel on these new routes. Southwest's strategy (point-to-point routes covering only major cities, use of secondary airports, standardized fleet, low fares) has been very different from the model followed by the older and bigger airlines (hub-and-spoke model extending to even smaller cities, presence in primary airports, variety in fleet, pursuit of high-end business travelers). The presence of discount airlines is therefore believed to reduce the fares greatly.

The file Airfares.xls contains real data that were collected for the third quarter of 1996. They consist of the predictors and response listed in Table 6.5. Note that some cities are served by more than one airport, and in those cases the airports are distinguished by their three-letter code.

a. Explore the numerical predictors and response (FARE) by creating a correlation table and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE?

b. Explore the categorical predictors (excluding the first four) by computing the percentage of flights in each category. Create a pivot table with the average fare in each category. Which categorical predictor seems best for predicting FARE?

c. Find a model for predicting the average fare on a new route:
   i. Convert categorical variables (e.g., SW) into dummy variables. Then, partition the data into training and validation sets. The model will be fit to the training data and evaluated on the validation set.
   ii. Use stepwise regression to reduce the number of predictors. You can ignore the first four predictors (S_CODE, S_CITY, E_CODE, E_CITY). Report the estimated model selected.
   iii. Repeat (ii) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (ii) in terms of the predictors that are in the model.
   iv. Compare the predictive accuracy of both models (ii) and (iii) using measures such as RMSE and average error and lift charts.
   v. Using model (iii), predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28, 760, E_INCOME = $27, 664, S_POP = 4, 557, 004, E_POP = 3, 195, 503, SLOT = Free, GATE = Free, PAX = 12782, DISTANCE = 1976 miles.
   vi. Predict the reduction in average fare on the route if in (b) Southwest decides to cover this route [using model (iii)].
   vii. In reality, which of the factors will not be available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)? Which ones can be estimated? How?
   viii. Select a model that includes only factors that are available before flights begin to operate on the new route. Use an exhaustive search to find such a model.
   ix. Use the model in (viii) to predict the average fare on a route with characteristics COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28, 760, E_INCOME = $27, 664, S_POP = 4, 557, 004, E_POP = 3, 195, 503, SLOT = Free, GATE = Free, PAX = 12782, DISTANCE = 1976 miles.
   x. Compare the predictive accuracy of this model with model (iii). Is this model good enough, or is it worthwhile reevaluating the model once flights begin on the new route?

Table 6.5: DESCRIPTION OF VARIABLES FOR AIRFARE EXAMPLE

S_CODE	Starting airport's code
S_CITY	Starting city
E_CODE	Ending airport's code
E_CITY	Ending city

COUPON	Average number of coupons (a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route
NEW	Number of new carriers entering that route between Q3-96 and Q2-97
VACATION	Whether (Yes) or not (No) a vacation route
SW	Whether (Yes) or not (No) Southwest Airlines serves that route
HI	Herfindahl index: measure of market concentration
S_INCOME	Starting city's average personal income
E_NCOME	Ending city's average personal income
S_POP	Starting city's population
E_POP	Ending city's population
SLOT	Whether or not either endpoint airport is slot controlled (this is a measure of airport congestion)
GATE	Whether or not either endpoint airport has gate constraints (this is another measure of airport congestion)
DISTANCE	Distance between two endpoint airports in miles
PAX	Number of passengers on that route during period of data collection
FARE	Average fare on that route

d. In competitive industries, a new entrant with a novel business plan can have a disruptive effect on existing firms. If a new entrant's business model is sustainable, other players are forced to respond by changing their business practices. If the goal of the analysis was to evaluate the effect of Southwest Airlines' presence on the airline industry rather than predicting fares on new routes, how would the analysis be different? Describe technical and conceptual aspects.

6.4 **Predicting Prices of Used Cars**. The file ToyotaCorolla.xls contains data on used cars (Toyota Corolla) on sale during late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications. (The example in Section 4.2.1 is a subset of this dataset.)

Data Preprocessing. Create dummy variables for the categorical predictors (Fuel Type and Metallic). Split the data into training (50%), validation (30%), and test (20%) datasets.

Run a multiple linear regression using the *Prediction* menu in XLMiner with the output variable Price and input variables Age_08_04, KM, Fuel_Type, HP, Automatic, Doors, Quarterly_Tax, Mfg_Guarantee, Guarantee_Period, Airco, Automatic_Airco, CD_Player, Powered_Windows, Sport_Model, and Tow_Bar.

a. What appear to be the three or four most important car specifications for predicting the car's price?

b. Using metrics you consider useful, assess the performance of the model in predicting prices.