In this chapter, we give a brief overview of several analytics methods that were referenced throughout the book. We do not cover all of the analytics methods discussed, but we try to cover many different methods that can be used for descriptive, predictive, and prescriptive analytics. Tutorials and resources for recommended software options can be found in the Online Companion for this book (www.dynamic-ideas.com/analytics-edge). For a more technical and comprehensive understanding of the methods presented here, we give several references in Section 21.7.

21.1 Linear Regression

Linear regression is one of the most common methods for making predictions. It is used to determine how an outcome variable, called the *dependent variable*, can best be expressed as a linear combination of a set of known input variables, called the *independent variables*. The dependent variable is typically denoted by y and the independent variables are denoted by $x_1, x_2, \ldots, x_k$, where k is the number of different independent variables. We are interested in finding the best possible coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ such that our predicted values:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

are as close as possible to the actual y values. There are many different ways to define the "best" coefficients. For basic linear regression, this is achieved by minimizing the sum of the squared differences between the actual values, y, and the predictions $\hat{y}$. These differences, $(y - \hat{y})$, are often called *error terms* or *residuals*.

As an example, we saw in Chapter 1 that Orley Ashenfelter developed the following multiple linear regression equation to predict wine quality:

> Wine Quality = -12.145 + (0.001173 × Winter Rainfall) + (0.616 × Average Growing Season Temperature) - (0.00386 × Harvest Rainfall) + (0.0238 × Age of Vintage).

Here, the dependent variable is "Wine Quality" and the four independent variables are "Winter Rainfall," "Average Growing Season Temperature," "Harvest Rainfall" and "Age of Vintage." By using a data set and minimizing the sum of the squared differences between the actual values and the predicted values, Ashenfelter discovered that the best coefficients were $\beta_0 = -12.145$, $\beta_1 = 0.001173$, $\beta_2 = 0.616$, $\beta_3 = -0.00386$ and $\beta_4 = 0.0238$.

Evaluating a Model

Once a linear regression model is constructed, it is important to evaluate the quality of the model. When a linear regression model is created with a statistical package, a lot of additional information is typically generated about the model in addition to the coefficient estimates, or the β values.

Significance

One important piece of information to consider when evaluating the model is the significance of the independent variables. In a linear regression model, a coefficient is considered *significant* if the coefficient estimate is significantly different from zero according to the data used to build the model. A coefficient of zero means that the value of the independent variable does not change the prediction for the dependent variable. If a coefficient is not significantly different than zero, then we should probably remove the variable from the model, since it is not helping predict the dependent variable.

More specifically, a *standard error* value can be computed for each coefficient. This value gives a measure of how much the "true" coefficient is likely to vary from the estimate value, given the data used to build the model. Keep in mind that the standard error only holds for the data used to build the model. If the dataset is not representative of future data, the model might not perform as well.

A *t-value* for a coefficient is also often computed, which is the coefficient estimate divided by the standard error (it will be negative if the coefficient is negative, and positive if the coefficient is positive). A larger absolute t-value means that the coefficient is more significant (we want independent variables with a small standard error relative to the coefficient estimate). So we ultimately want independent variables with large absolute t-values. A probability that a coefficient is actually zero given the data can also be computed (called a p-value), and the linear regression model output in any software package will often make it easy to tell which variables are significant. For a more mathematically rigorous treatment of these topics, see the references in Section 21.7.

Removing Independent Variables

If there are insignificant independent variables in the model, we should consider removing these variables, for several reasons. One is that having too many independent variables compared to the number of data observations can cause the model to *overfit* to the data used to build the model. This leads to a model that does really well at predicting the outcome for data it has seen before, but really poorly at predicting the outcome for new observations.

If there is significantly more data than independent variables, then overfitting is not really a concern, and leaving in independent variables that are insignificant will not hurt the predictive ability of the model. But it might still be better to remove the insignificant variables for another important reason: simplicity and interpretability of the model. A model with unnecessary independent variables requires more data, and makes the model more complex than it needs to be. Simpler models that are just as accurate should be preferred to more complex models that do not improve the accuracy of prediction. However, a modeler should be careful when removing independent variables because there are interactions going on between the independent variables that can change when a variable is removed. We discuss this more below.

The Metric R^2

Another important evaluation of a linear regression model is to compute the R-squared, or R^2, value. The R^2 of a linear regression model can be computed with the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where n is the number of data observations, y_i is the actual value of the dependent variable for observation i, $\hat{y_i}$ is the predicted value of the dependent variable for observation i, and $\bar{y}$ is the mean value of the dependent variable, for all observations. The numerator of the fraction is called the *sum of squared errors (SSE)* and the denominator of the fraction is called the *total sum of squares (SST)*.

The R^2 value can be interpreted as the proportion of the total variation of the dependent variable that is accounted for by the independent variables used in the regression equation. A different interpretation of R^2 is that it measures the value added of the model over a "baseline" model of predicting the average of the dependent variable for all observations. It is unit-less and takes values between 0 and 1. The worst case is an R^2 of 0, because it means that using the independent variables to make predictions is no better than just predicting the mean of the dependent variable for all observations. The best case is an R^2 of 1, because it means that the model perfectly predicts the variation of the dependent variable, and we do not make any errors.

Generally, a higher R^2 value implies a better regression model, but this can be very dependent on the application area. A valuable model for a hard problem (like predicting stock prices) can have a low R^2 because the problem is so challenging, and a less valuable model for an easy problem can have a high R^2 because the problem is so easy (like predicting revenue of a product using the number of items sold). Therefore, it is important to evaluate the use of a model in practice in addition to computing the R^2 value.

Refining the Model

If it has been determined that one or more insignificant independent variables should be removed from the model, it is typically easy to rebuild the model without a few independent variables. However, the independent variables should be removed one at a time. This is due to interactions between the independent variables, specifically something called *multicollinearity*. Multicollinearity means that two independent variables are highly correlated. If there is multicollinearity in the model, two independent variables are basically representing the same thing. This can make the variables seem insignificant when they actually should be significant. By removing just one of them, it can be discovered that the other variable is significant in the model.

If there are multiple insignificant independent variables in the model, a general rule for which variable to remove first is to remove the one that is the least significant. However, expert human judgment can also be a good strategy for deciding which of two insignificant variables to remove first, based on intuition about the model.

Making Predictions

An important final step to validate a linear regression model is to make predictions on new data, which is often called *out-of-sample* data or a *test set*. If a linear regression model is not good at predicting unseen data, it might be overfit to the training data.

To make predictions, the new data should be plugged into the linear regression equation. For example, recall Ashenfelter's linear regression equation to predict the quality of wine:

Wine Quality = -12.145 + (0.001173 × Winter Rainfall) + (0.616 × Average Growing Season Temperature) - (0.00386 × Harvest Rainfall) + (0.0238 × Age of Vintage).

Suppose we have two new data points about wine prices, which Ashenfelter did not use to build his model, described in Table 21.1.

Using the linear regression equation, we can predict that for Test Point 1,

Wine Quality = -12.145 + (0.001173 × 717) + (0.616 × 16.17) - (0.00386 × 122) + (0.0238 × 4) = -1.719,

and for Test Point 2,

Table 21.1: Out-of-sample data for Ashenfelter's wine equation.

Variable	Test Point 1	Test Point 2
Winter Rainfall	717	578
Average Growing Season Temperature	16.17	16.00
Harvest Rainfall	122	74
Age of Vintage	4	3

Wine Quality = -12.145 + (0.001173 × 578) + (0.616 × 16.00) - (0.00386 × 74) + (0.0238 × 3) = -1.825.

Note that we do not need the values of the dependent variables to make predictions. This is important, because often it is necessary to make future predictions before the results are realized (for example, think about predicting the winner of the Academy Awards the week before the results are revealed). But suppose that we know that the actual Wine Quality of the first data point is -1.540, and the actual Wine Quality of the second data point is -1.996. Then we can compute our accuracy on this out-of-sample data by computing the test set R^2.

First, we can compute the sum of squared errors:

$$SSE = (-1.719 - (-1.540))^2 + (-1.825 - (-1.996))^2 = 0.061.$$

Then we can compute the total sum of squares:

$$SST = (-1.426 - (-1.540))^2 + (-1.426 - (-1.996))^2 = 0.338,$$

where -1.426 is the average wine quality of the data used to build the model. Note that we use the mean of Wine Quality in the *training set* to calculate SST. This is because we are trying to compare ourselves against how well we could do by just guessing the mean. In reality, we often have to make out-of-sample predictions before we know the outcome, and so we can not use the test set mean. To make this calculation match what will occur in practice, we should use the training set mean. (Note that this convention makes it so that the R^2 of a model on the test set can sometimes be negative!)

So the R^2 on our test set is

$$R^2 = 1 - \frac{0.061}{0.338} = 0.82$$

This is promising, but we would want to get a larger test set to be more confident of this result. Computing how good a model is at making out-of-sample predictions is crucial for evaluating the predictive power of the model.

21.2 Logistic Regression

Logistic regression is a predictive method that handles cases where the dependent variable, y, only has two possible outcomes, called *classes*. Examples of dependent variables that could be used with logistic regression are predicting whether a new business will succeed or fail, whether or not a loan will be approved, and whether a team will win or lose a game. These are all called *classification* problems, since the goal is to figure out which class each observation belongs to.

Similar to linear regression, logistic regression uses a set of independent variables to make predictions, but instead of predicting a continuous value for the dependent variable, it predicts the probability of membership in each of the possible outcomes, or classes.

The two possible outcomes are often denoted as 1 and 0. For example, if our possible outcomes were “success” or “failure,” then we could denote success by 1 and failure by 0. If our possible outcomes were “yes” or “no,” then we could denote yes by 1 and no by 0. We will refer to the two possible classes as 1 and 0 for the rest of this section.

Logistic regression consists of two steps. The first step is to compute the probabilities of the two different outcomes. Thus, for each observation i, we are computing $P(y_i = 1)$, or the probability the observation belongs to class 1, and $P(y_i = 0)$, or the probability the observation belongs to class 0. Since we only have two possible outcomes, we know that $P(y_i = 0) = 1 - P(y_i = 1)$ according to the basic properties of probability (the sum of the probabilities for all possible outcomes must equal 1). So logistic regression with two classes only needs to predict $P(y_i = 1)$ for each observation i, since we can easily compute $P(y_i = 0)$.

In the second step of logistic regression, we use a *cutoff*, or *threshold*, value to classify each observation into one of the classes. A common cutoff value is 0.5, meaning that if $P(y_i = 1) \geq 0.5$ we will classify observation i into class 1 and if $P(y = 1) < 0.5$ we will classify observation i into class 0. Simply, we will classify each observation into the class with the highest probability. However, other cutoff values can be chosen, and in some cases are more appropriate. For example, if the probability of belonging to class 1 is a low probability event, a lower cutoff value might be sufficient. We discuss this more later in this section.

The Logistic Regression Model

The probabilities in logistic regression are computed using an equation that is similar to the equation used for linear regression. In linear regression, our equation to predict y was $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$. Since in logistic regression we want our prediction to be a probability, we use a nonlinear function that will only produce values between 0 and 1:

$$P(y = 1) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k)}}$$

This is called the *logistic response function* and it predicts the probability of an observation belonging to class 1. For more technical details about the logistic response function, see the references in Section 21.7.

To make this equation easier to understand, we define what is known as the *odds* of belonging to class 1:

$$Odds = \frac{P(y = 1)}{P(y = 0)} = \frac{P(y = 1)}{1 - P(y = 1)}.$$

This metric is very popular in gambling games and many other areas. Instead of talking about the probability of an outcome, we can refer to the odds of an outcome. When the probability of class 1 is equal to the probability of