Questions for Case 3.3

1. What were the challenges the Town of Cary was facing?
2. What was the proposed solution?
3. What were the results?
4. What other problems and data analytics solutions do you foresee for towns like Cary?

Source: "Municipality Puts Wireless Water Meter-Reading Data To Work (SAS® Analytics)—The Town of Cary, North Carolina Uses SAS Analytics to Analyze Data from Wireless Water Meters, Assess Demand, Detect Problems and Engage Customers." Copyright © 2016 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

3.6 REGRESSION MODELING FOR INFERENTIAL STATISTICS

Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics. Historically speaking, the roots of regression date back to the 1920s and 1930s, to the earlier work on inherited characteristics of sweet peas by Sir Francis Galton and subsequently by Karl Pearson. Since then, regression has become the statistical technique for characterization of relationships between explanatory (input) variable(s) and response (output) variable(s).

As popular as it is, regression essentially is a relatively simple statistical technique to model the dependence of a variable (response or output variable) on one (or more) explanatory (input) variables. Once identified, this relationship between the variables can be formally represented as a linear/additive function/equation. As is the case with many other modeling techniques, regression aims to capture the functional relationship between and among the characteristics of the real world and describe this relationship with a mathematical model, which can then be used to discover and understand the complexities of reality—explore and explain relationships or forecast future occurrences.

Regression can be used for one of two purposes: hypothesis testing—investigating potential relationships between different variables—and prediction/forecasting—estimating values of a response variable based on one or more explanatory variables. These two uses are not mutually exclusive. The explanatory power of regression is also the foundation of its predictive ability. In hypothesis testing (theory building), regression analysis can reveal the existence/strength and the directions of relationships between a number of explanatory variables (often represented with x_i) and the response variable (often represented with y). In prediction, regression identifies additive mathematical relationships (in the form of an equation) between one or more explanatory variables and a response variable. Once determined, this equation can be used to forecast the values of the response variable for a given set of values of the explanatory variables.

CORRELATION VERSUS REGRESSION Because regression analysis originated from correlation studies, and because both methods attempt to describe the association between two (or more) variables, these two terms are often confused by professionals and even by scientists. **Correlation** makes no a priori assumption of whether one variable is dependent on the other(s) and is not concerned with the relationship between variables; instead it gives an estimate on the degree of association between the variables. On the other hand, regression attempts to describe the dependence of a response variable on one (or more) explanatory variables where it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect. Also, although correlation is interested in the low-level relationships between two variables, regression is concerned with the relationships between all explanatory variables and the response variable.

SIMPLE VERSUS MULTIPLE REGRESSION If the regression equation is built between one response variable and one explanatory variable, then it is called *simple regression*. For instance, the regression equation built to predict/explain the relationship between the height of a person (explanatory variable) and the weight of a person (response variable) is a good example of simple regression. Multiple regression is the extension of simple regression when the explanatory variables are more than one. For instance, in the previous example, if we were to include not only the height of the person but also other personal characteristics (e.g., BMI, gender, ethnicity) to predict the person's weight, then we would be performing multiple regression analysis. In both cases, the relationship between the response variable and the explanatory variable(s) is linear and additive in nature. If the relationships are not linear, then we might want to use one of many other nonlinear regression methods to better capture the relationships between the input and output variables.

How Do We Develop the Linear Regression Model?

To understand the relationship between two variables, the simplest thing that one can do is to draw a scatter plot where the *y*-axis represents the values of the response variable and the *x*-axis represents the values of the explanatory variable (see Figure 3.13). A scatter plot would show the changes in the response variable as a function of the changes in the explanatory variable. In the case shown in Figure 3.13, there seems to be a positive relationship between the two; as the explanatory variable values increase, so does the response variable.

Simple regression analysis aims to find a mathematical representation of this relationship. In reality, it tries to find the signature of a straight line passing through right between the plotted dots (representing the observation/historical data) in such a way that it minimizes the distance between the dots and the line (the predicted values on the

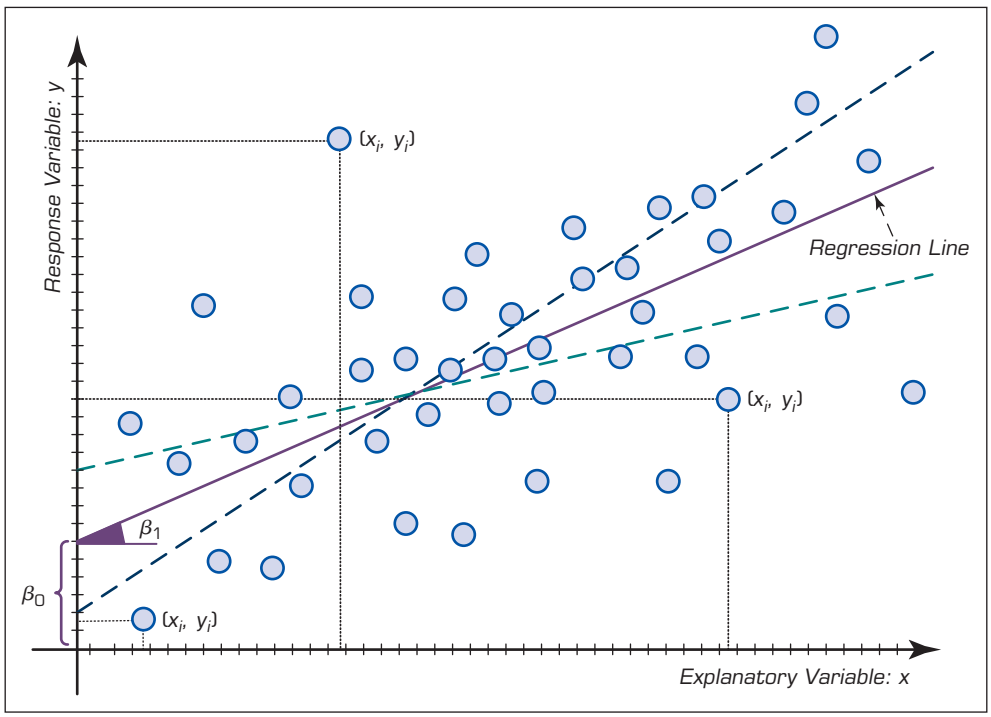

FIGURE 3.13 A Scatter Plot and a Linear Regression Line.

theoretical regression line). Even though there are several methods/algorithms proposed to identify the regression line, the one that is most commonly used is called the **ordinary least squares (OLS)** method. The OLS method aims to minimize the sum of squared residuals (squared vertical distances between the observation and the regression point) and leads to a mathematical expression for the estimated value of the regression line (which are known as β parameters). For simple **linear regression**, the aforementioned relationship between the response variable (y) and the explanatory variable(s) (x) can be shown as a simple equation as follows:

$$y = \beta_0 + \beta_1 x$$

In this equation, β_0 is called the intercept, and β_1 is called the slope. Once OLS determines the values of these two coefficients, the simple equation can be used to forecast the values of y for given values of x. The sign and the value of β_1 also reveal the direction and the strengths of relationship between the two variables.

If the model is of a multiple linear regression type, then there would be more coefficients to be determined, one for each additional explanatory variable. As the following formula shows, the additional explanatory variable would be multiplied with the new β_i coefficients and summed together to establish a linear additive representation of the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdot + \beta_n x_n$$

How Do We Know If the Model Is Good Enough?

Because of a variety of reasons, sometimes models as representations of the reality do not prove to be good. Regardless of the number of explanatory variables included, there is always a possibility of not having a good model, and therefore the linear regression model needs to be assessed for its fit (the degree to which it represents the response variable). In the simplest sense, a well-fitting regression model results in predicted values close to the observed data values. For the numerical assessment, three statistical measures are often used in evaluating the fit of a regression model: R^2(R − squared), the overall F-test, and the root mean square error (RMSE). All three of these measures are based on the sums of the square errors (how far the data are from the mean and how far the data are from the model's predicted values). Different combinations of these two values provide different information about how the regression model compares to the mean model.

Of the three, R^2 has the most useful and understandable meaning because of its intuitive scale. The value of R^2 ranges from 0 to 1 (corresponding to the amount of variability explained in percentage) with 0 indicating that the relationship and the prediction power of the proposed model is not good, and 1 indicating that the proposed model is a perfect fit that produces exact predictions (which is almost never the case). The good R^2 values would usually come close to one, and the closeness is a matter of the phenomenon being modeled—whereas an R^2 value of 0.3 for a linear regression model in social sciences can be considered good enough, an R^2 value of 0.7 in engineering might be considered as not a good enough fit. The improvement in the regression model can be achieved by adding more explanatory variables or using different data transformation techniques, which would result in comparative increases in an R^2 value. Figure 3.14 shows the process flow of developing regression models. As can be seen in the process flow, the model development task is followed by the model assessment task in which not only is the fit of the model assessed, but because of restrictive assumptions with which the linear models have to comply, the validity of the model also needs to be put under the microscope.

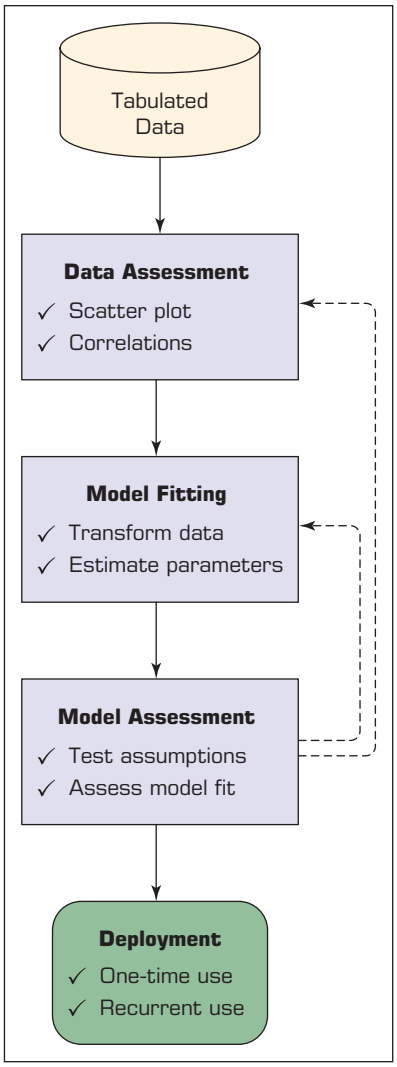

FIGURE 3.14 **A Process Flow for Developing Regression Models.**

What Are the Most Important Assumptions in Linear Regression?

Even though they are still the choice of many for data analyses (both for explanatory and for predictive modeling purposes), linear regression models suffer from several highly restrictive assumptions. The validity of the linear model built depends on its ability to comply with these assumptions. Here are the most commonly pronounced assumptions:

1. **Linearity.** This assumption states that the relationship between the response variable and the explanatory variables is linear. That is, the expected value of the response variable is a straight-line function of each explanatory variable while holding all other explanatory variables fixed. Also, the slope of the line does not depend on the values of the other variables. It also implies that the effects of different explanatory variables on the expected value of the response variable are additive in nature.
2. **Independence (of errors).** This assumption states that the errors of the response variable are uncorrelated with each other. This independence of the errors is weaker

than actual statistical independence, which is a stronger condition and is often not needed for linear regression analysis.

3. **Normality (of errors).** This assumption states that the errors of the response variable are normally distributed. That is, they are supposed to be totally random and should not represent any nonrandom patterns.
4. **Constant variance (of errors).** This assumption, also called *homoscedasticity*, states that the response variables have the same variance in their error regardless of the values of the explanatory variables. In practice, this assumption is invalid if the response variable varies over a wide enough range/scale.
5. **Multicollinearity.** This assumption states that the explanatory variables are not correlated (i.e., do not replicate the same but provide a different perspective of the information needed for the model). Multicollinearity can be triggered by having two or more perfectly correlated explanatory variables presented to the model (e.g., if the same explanatory variable is mistakenly included in the model twice, one with a slight transformation of the same variable). A correlation-based data assessment usually catches this error.

There are statistical techniques developed to identify the violation of these assumptions and techniques to mitigate them. The most important part for a modeler is to be aware of their existence and to put in place the means to assess the models to make sure that they are compliant with the assumptions they are built on.

Logistic Regression

Logistic regression is a very popular, statistically sound, probability-based classification algorithm that employs supervised **learning**. It was developed in the 1940s as a complement to linear regression and linear discriminant analysis methods. It has been used extensively in numerous disciplines, including the medical and social sciences fields. Logistic regression is similar to linear regression in that it also aims to regress to a mathematical function that explains the relationship between the response variable and the explanatory variables using a sample of past observations (training data). Logistic regression differs from linear regression with one major point: its output (response variable) is a class as opposed to a numerical variable. That is, whereas linear regression is used to estimate a continuous numerical variable, logistic regression is used to classify a categorical variable. Even though the original form of logistic regression was developed for a binary output variable (e.g., 1/0, yes/no, pass/fail, accept/reject), the present-day modified version is capable of predicting multiclass output variables (i.e., multinomial logistic regression). If there is only one predictor variable and one predicted variable, the method is called *simple logistic regression* (similar to calling linear regression models with only one independent variable *simple linear regression*).

In predictive analytics, logistic regression models are used to develop probabilistic models between one or more explanatory/predictor variables (which can be a mix of both continuous and categorical in nature) and a class/response variable (which can be binomial/binary or multinomial/multiclass). Unlike ordinary linear regression, logistic regression is used for predicting categorical (often binary) outcomes of the response variable—treating the response variable as the outcome of a Bernoulli trial. Therefore, logistic regression takes the natural logarithm of the odds of the response variable to create a continuous criterion as a transformed version of the response variable. Thus, the logit transformation is referred to as the *link function* in logistic regression—even though the response variable in logistic regression is categorical or binomial, the logit is the continuous criterion on which linear regression is conducted. Figure 3.15 shows a logistic

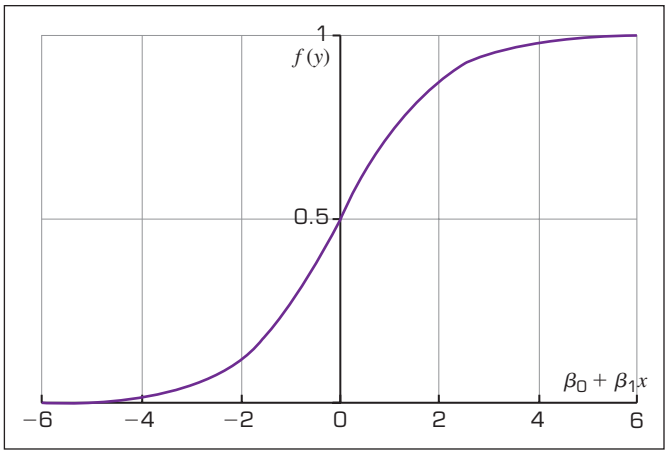

FIGURE 3.15 **The Logistic Function.**

regression function where the odds are represented in the *x-axis* (a linear function of the independent variables), whereas the probabilistic outcome is shown in the *y-axis* (i.e., response variable values change between 0 and 1).

The logistic function, $f(y)$ in Figure 3.15 is the core of logistic regression, which can take values only between 0 and 1. The following equation is a simple mathematical representation of this function:

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The logistic regression coefficients (the βs) are usually estimated using the maximum likelihood estimation method. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so an iterative process must be used instead. This process begins with a tentative starting solution, then revises the parameters slightly to see if the solution can be improved, and repeats this iterative revision until no improvement can be achieved or is very minimal, at which point the process is said to have completed/converged.

Sports analytics—use of data and statistical/analytics techniques to better manage sports teams/organizations—has been gaining tremendous popularity. Use of data-driven analytics techniques has become mainstream for not only professional teams but also college and amateur sports. Application Case 3.4 is an example of how existing and readily available public data sources can be used to predict college football bowl game outcomes using both classification and regression-type prediction models.

Time-Series Forecasting

Sometimes the variable that we are interested in (i.e., the response variable) might not have distinctly identifiable explanatory variables, or there might be too many of them in a highly complex relationship. In such cases, if the data are available in a desired format, a prediction model, the so-called time series, can be developed. A time series is a sequence of data points of the variable of interest, measured and represented at successive points in time spaced at uniform time intervals. Examples of time series include monthly rain volumes in a geographic area, the daily closing value of the stock market indexes, and