# 7 Regression

Learning Objectives

* Understand the basic concept of correlation and regression
* Learn how to do a regression exercise
* Learn how to do a nonlinear regression exercise
* Know about logistic regression as a technique for classification
* Appreciate the advantages and disadvantages of regression

### INTRODUCTION

Regression is a well-known statistical technique to model the predictive relation- ship between several independent variables (DVs) and one dependent variable. The objective is to find the best-fitting curve for a dependent variable in a multi- dimensional space, with each independent variable being a dimension. The curve could be a straight line, or it could be a nonlinear curve. The quality of fit of the curve to the data can be measured by a coefficient of correlation (*r*), which is the square root of the amount of variance explained by the curve.

The key steps for regression are simple

1. List all the variables available for making the model.
2. Establish a Dependent Variable (DV) of interest.
3. Examine visual (if possible) relationships between variables of interest.
4. Find a way to predict DV using other variables.

#### Caselet: Data Driven Prediction Markets

*Traditional pollsters still seem to be using methodologies that worked well a decade or two ago. Nate Silver is a new breed of data-based political forecasters who are seeped in big data and advanced analytics. In the 2012 elections, he predicted that Obama would win the election with 291 electoral votes, compared to 247 for Mitt Romney, giving the President a 62% lead and reelection. He stunned the political forecasting world by correctly predicting the Presidential winner in all 50 states, including all nine swing states. He also correctly predicted the winner in 31 of the 33 US Senate races.*

*Nate Silver brings a different view to the world of forecasting political elections, viewing it as a scientific discipline. State the hypothesis scientifically, gather all available information, analyze the data and extract insights using sophisticated models and algorithms and finally, apply human judgment to interpret those in- sights. The results are likely to be much more grounded and successful. (Source: The Signal and the Noise: Why Most Predictions Fail but Some Don’t, by Nate Silver, 2012)*

1. *What is the impact of this story on traditional pollsters and commentators?*

### CORRELATIONS AND RELATIONSHIPS

Statistical relationships are about which elements of data hang together and which ones hang separately. It is about categorizing variables that have a relationship with one another and categorizing variables that are distinct and unrelated to other variables. It is about describing significant positive relationships and significant negative differences.

The first and foremost measure of the strength of a relationship is co-relation (or correlation). The strength of a correlation is a quantitative measure that is measured in a normalized range between 0 and 1. A correlation of 1 indicates a perfect relationship, where the two variables are in perfect sync. A correlation of 0 indicates that there is no relationship between the variables.

The relationship can be positive, or it can be an inverse relationship, that is, the variables may move together in the same direction or in the opposite direction. Therefore, a good measure of correlation is the correlation coefficient, which is the square root of correlation. This coefficient, called *r*

to +1. An *r* value of 0 signifies no relationship. An *r* value of 1 shows perfect relationship in the same direction, and an *r * - ship but moving in opposite directions.



Given two numeric variables *x* and *y*, the coefficient of correlation *r* is mathematically computed by the following equation. *x*– (called *x*-bar) is the mean of *x*, and *y*– (*y*-bar) is the mean of *y*.

[(*x* - *x* ][ *y* - *y* ]

[(*x* - *x* )2 ][( *y* - *y* )2]

(*x* - *x* )( *y* - *y* )

[(*x* - *x* )2 ][( *y* - *y* )2]

*r* =

*r* =

### VISUAL LOOK AT RELATIONSHIPS

A scatter plot (or scatter diagram) is a simple exercise for plotting all the data points between two variables on a two-dimensional graph. It provides a visual layout of all the data points placed in that two-dimensional space. The scatter plot can be useful for graphically intuiting the relationship between the two variables.

Figure 7.1 shows many possible patterns in scatter diagrams.

(a) Linear

(a) Linear

(c) Curvilinear

y y y

x

x

x

(d) Curvilinear (e) No Relationship (f) No Relationship

y y y

x

x

x

FIGURE 7.1 Scatter Plots showing Types of Relationships among Two Variables (*Source:* Groebner et al., 2013)

Chart (a) shows a very strong linear relationship between the variables *x* and *y*. This means the value of *y* increases proportionally with *x*. Chart (b) also shows a strong linear relationship between the variables *x* and *y*. Here, it is an inverse relationship. That means the value of *y* decreases proportionally with *x*.

Chart (c) shows a curvilinear relationship. It is an inverse relationship, which means that the value of *y* decreases proportionally with *x*. However, it seems a relatively well-defined relationship, like an arc of a circle, which can be represented by a simple quadratic equation (quadratic means the power of two, that is, using terms like *x*2 and *y*2). Chart (d) shows a positive curvilinear relationship. However, it does not seem to resemble a regular shape, and thus would not be a strong relationship. Charts (e) and (f) show no relationship, that means variables *x* and *y* are independent of each other.

Charts (a) and (b) are good candidates that model a simple linear regression model (the terms regression model and regression equation can be used inter- changeably). Chart (c) too could be modeled with a little more complex, quadratic regression equation. Chart (d) might require an even higher order polynomial regression equation to represent the data. Charts (e) and (f) have no relation- ship, thus, they cannot be modeled together, by regression or using any other modeling tool.

### Regression Exercise

The regression model is described as a linear equation that follows. *y* is the dependent variable, that is, the variable being predicted. *x* is the independent variable, or the predictor variable. There could be many predictor variables (such as *x*1, *x*2, …) in a regression equation. However, there can be only one dependent variable (*y*) in the regression equation.

*y* = b0 + b1*x* + e

Where b0 and b1 are the constant, and the co-efficient for the *x* variable; and e

is the random error variable.

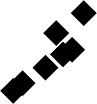
A simple example of a regression equation would be to predict a house price from the size of the house. Dataset 7.1 shows a sample house prices data:

Dataset 7.1

|  |  |
| --- | --- |
| House Price | Size (Sq |
| ($) | ft) |
| 229,500 | 1850 |
| 273,300 | 2190 |
| 247,000 | 2100 |
| 195,100 | 1930 |
| 261,000 | 2300 |
| 179,700 | 1710 |
| 168,500 | 1550 |
| 234,400 | 1920 |
| 168,800 | 1840 |
| 180,400 | 1720 |
| 156,200 | 1660 |
| 288,350 | 2405 |
| 186,750 | 1525 |
| 202,100 | 2030 |
| 256,800 | 2240 |

The two dimensions (one predictor and one outcome variable) of the data can be plotted on a scatter diagram. A scatter plot with a best-fitting line looks like the graph that follows (Figure 7.2).

$350,000



$300,000

$250,000

House Price

$200,000

$150,000

$100,000

$50,000

$0

0 1000 2000 3000

Size (Sq ft)

House Price

Linear (House Price)

FIGURE 7.2 Scatter Plot and Regression Equation between House Price and House Size

Visually, one can see a positive correlation between house price and size (Sq ft). However, the relationship is not perfect. Running a regression model between the two variables produces the following output (truncated).

Regression Statistics

*R* 0.891

*r*2 0.794

Coefficients

Intercept 54191

Size (Sq ft) 139.48

It shows the coefficient of correlation to be 0.891. *r*2, the measure of total variance explained by the equation, is 0.794 or 79%. That means the two variables are moderately and positively correlated. Regression coefficients help create the following equation for predicting house prices.

### House Price ($) = 139.48 \* Size (Sq ft) – 54191

This equation explains only 79% of the variance in house prices. Suppose other predictor variables are made available, such as the number of rooms in the house. It might help improve the regression model.

The house data will now look like as shown in Dataset 7.2 given below.

Dataset 7.2

|  |  |  |
| --- | --- | --- |
| House Price | Size (Sq | No. of Rooms |
| ($) | ft) |  |
| 229,500 | 1850 | 4 |
| 273,300 | 2190 | 5 |
| 247,000 | 2100 | 4 |
| 195,100 | 1930 | 3 |
| 261,000 | 2300 | 4 |
| 179,700 | 1710 | 2 |
| 168,500 | 1550 | 2 |
| 234,400 | 1920 | 4 |
| 168,800 | 1840 | 2 |
| 180,400 | 1720 | 2 |
| 156,200 | 1660 | 2 |
| 288,350 | 2405 | 5 |
| 186,750 | 1525 | 3 |
| 202,100 | 2030 | 2 |
| 256,800 | 2240 | 4 |

While it is possible to make a three-dimensional scatter plot, one can alternatively examine the correlation matrix among the variables.

|  |  |  |  |
| --- | --- | --- | --- |
|  | House Price | Size (Sq ft) | No. of Rooms |
| House Price | 1 |  |  |
| Size (Sq ft) | 0.891 | 1 |  |
| Rooms | 0.944 | 0.748 | 1 |

It shows that the house price has a strong correlation with number of rooms (0.944) as well. Thus, it is likely that adding this variable to the regression model will add to the strength of the model.

Running a regression model between these three variables produces the following output (truncated).

Regression Statistics

*r* 0.984

*r*2 0.968

|  |  |
| --- | --- |
|  | Coefficients |
| Intercept | 12923 |
| Size (Sq ft) | 65.60 |
| Rooms | 23613 |

It shows that the coefficient of correlation of this regression model is 0.984. *R*2, the total variance explained by the equation, is 0.968 or 97%. That means the variables are positively and very strongly correlated. Adding a new relevant variable has helped improve the strength of the regression model.

Using the regression coefficients helps create the following equation for predicting house prices.

### House Price ($) = 65.6 \* Size (Sq ft) + 23613 \* Rooms + 12924

This equation shows a 97% goodness of fit with the data, which is very good for business and economic data. There is always some random variation in naturally occurring business data, and it is not desirable to over fit the model to the data.

This predictive equation should be used for future transactions. Given a situation as below, it will be possible to predict the price of the house with 2000 Sq ft and 3 rooms.

House Price Size (Sq

ft)

#No. of

Rooms

?? 2000 3

### House Price ($) = 65.6 \* 2000 (Sq ft) + 23613 \* 3 + 12924 = $214,963

The predicted values should be compared with the actual values to see how close the model is able to predict the actual value. As new data points become avail- able, there are opportunities to fine-tune and improve the model.

### NON-LINEAR REGRESSION EXERCISE

The relationship between the variables may also be curvilinear. For example, given past data from electricity consumption (kWh) and temperature (K), the objective is to predict the electrical consumption from the temperature value. Dataset 7.3 shows a dozen past observations.

|  |  |
| --- | --- |
| Dataset 7.3 |  |
| Kwatts | Temp (F) |
| 12530 | 46.8 |
| 10800 | 52.1 |
| 10180 | 55.1 |
| 9730 | 59.2 |
| 9750 | 61.9 |
| 10230 | 66.2 |
| 11160 | 69.9 |
| 13910 | 76.8 |
| 15690 | 79.3 |
| 15110 | 79.7 |
| 17020 | 80.2 |
| 17880 | 83.3 |

In two dimensions (one predictor and one outcome variable), data can be plotted on a scatter diagram. A scatter plot with a best-fitting line looks like the graph on next page (Figure 7.3).

It is visually clear that the first line does not fit the data well. The relationship between temperature and Kwatts follows a curvilinear model, where it hits bottom at a certain value of temperature. The regression model confirms the relationship since R is only 0.77 and Rsquare is also only 60%. Thus, only 60% of the variance is explained.

This regression model can be enhanced by introducing a nonlinear variable (such as a quadratic variable Temp2) in the equation. The second line is the relation- ship between kWh and Temp2. The scatter plot shows that energy consumption has a strong linear relationship with Temp2. Computing the regression model after adding the Temp2 variable leads to the following results

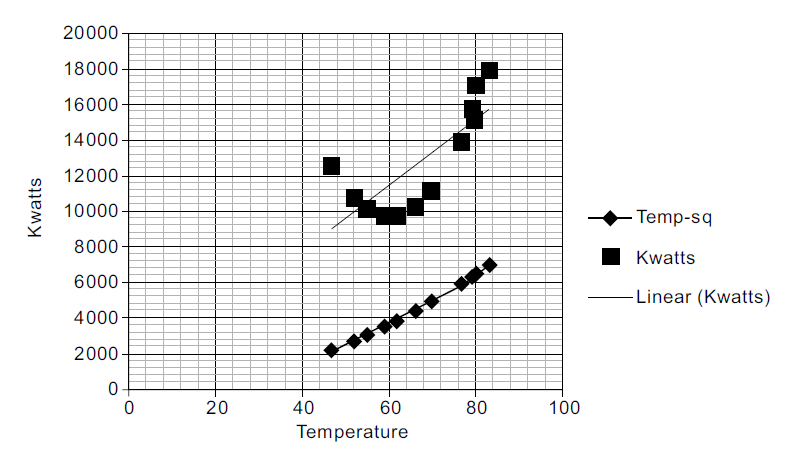


FIGURE 7.3 Scatter Plots showing Regression between (a) Kwatts and Temperature, and

(b) Kwatts and Temperature Square

Regression Statistics

*R* 0.992

*r*2 0.984

Coefficients

Intercept 67245

Temp (F) 1911

Temp Sq 15.87

It shows that the coefficient of correlation of the regression model is now 0.99. *R*2, the total variance explained by the equation is 0.985 or 98.5%. That means the variables are very strongly and positively correlated. The regression coefficients help create the following equation

### Energy Consumption (Kwatts) = 15.87 \* Temp2 –1911 \* Temp + 67245

This equation shows a 98.5% fit which is very good for business and economic con- texts. Now one can predict the Kwatts value when the temperature is 72 degree.

Energy consumption = (15.87 \* 72 \* 72) – (1911 \* 72) + 67245 = 11923 Kwatts

### LOGISTIC REGRESSION

Regression models traditionally work with continuous numeric value data for dependent and independent variables. Logistic regression models can, however, work with dependent variables that have categorical values, such as whether a loan is approved or not. Logistic regression measures the relationship between a categorical dependent variable and one or more independent variables. For example, logistic regression might be used to predict whether a patient has a given disease (e.g., diabetes), based on observed characteristics of the patient (age, gender, body mass index, results of blood tests, etc.).

Logistical regression models use probability scores as the predicted values of the dependent variables. Logistic regression takes the natural logarithm of the prob- ability of the dependent variable being a case (referred to as the logit function), and creates a continuous criterion as a transformed version of the dependent variable. Thus, the logit transformation is used in logistic regression as the dependent variable. The net effect is that although the dependent variable in logistic regression is binomial (or categorical, i.e., has only two possible values), the logit is the continuous function upon which linear regression is conducted. Here is the general logistic function with independent variable on the horizontal axis and the logit dependent variable on the vertical axis (Figure 7.4).

1

0.5

0

–6 –4 –2 0 2 4 6

FIGURE 7.4 General Logit Function

All popular data mining platforms provide support for regular multiple regression models, as well as options for Logistic Regression.

### ADVANTAGES AND DISADVANTAGES OF REGRESSION MODELS

Regression models are very popular because they offer many advantages. Few are as follows

* Regression models are easy to understand as they are built upon basic statistical principles such as correlation and least square error.
* Regression models provide simple algebraic equations that are easy to understand and use.
* The strength (or the goodness of fit) of the regression model is measured in terms of the correlation coefficients, and other related statistical parameters that are well understood.
* Regression models can match and beat the predictive power of other modeling techniques.
* Regression models can include all the variables that one wants to include in the model.
* Regression modeling tools are pervasive. They are found in statistical packages as well as data mining packages. MS-Excel spreadsheets can also provide simple regression modeling capabilities.

Regression models can however prove inadequate under many circumstances.

* Regression models cannot cover for poor data quality issues. If the data is not prepared well to remove missing values or is not well-behaved in terms of a normal distribution, the validity of the model suffers.
* Regression models suffer from collinearity problems (meaning strong linear correlations among some independent variables). If the independent variables have strong correlations among themselves, then they will eat into each other’s predictive power and the regression coefficients will lose their ruggedness. Regression models will not automatically choose between highly collinear variables, although some packages attempt to do that.
* Regression models can be unwieldy and unreliable if a large number of variables are included in the model. All variables entered into the model will be reflected in the regression equation, irrespective of their contribution to the predictive power of the model. There is no concept of automatic pruning of the regression model.
* Regression models do not automatically take care of nonlinearity. The user needs to imagine the kind of additional terms that might be needed to be added to the regression model to improve its fit.
* Regression models work only with numeric data and not with categorical variables. There are ways to deal with categorical variables though by creating multiple new variables with a yes or no value.

## Conclusion

Regression models are simple, versatile, visual/graphical tools with high predictive ability. They include nonlinear as well as binary predictions. Regression models should be used in conjunction with other data mining techniques to confirm the findings.

## Questions

1. What is a regression model?
2. What is a scatter plot? How does it help?
3. Compare and contrast decision trees with regression models.
4. Using the data given in Dataset 7.4 as shown below, create a regression model to predict the Test2 from Test1 score. Then predict the score for the one who got a 46 in Test1.

|  |  |
| --- | --- |
| Dataset 7.4 |  |
| Test1 | Test2 |
| 59 | 56 |
| 52 | 63 |
| 44 | 55 |
| 51 | 50 |
| 42 | 66 |
| 42 | 48 |
| 41 | 58 |
| 45 | 36 |
| 27 | 13 |
| 63 | 50 |
| 54 | 81 |
| 44 | 56 |
| 50 | 64 |
| 47 | 50 |

## True/False

1. Regression is an artificial intelligence technique.
2. In regression, a dependent variable is predicted using many independent variables.
3. Correlation coefficient (*R*) can take only positive values from zero to 1.
4. The best-fitting regression line can be straight or a curved one.
5. Regression model can automatically adjust the model to take into account any nonlinear relationship.
6. There can be only one dependent variable in one regression equation.
7. Regression models can be used for time-series analysis.
8. Regression models traditionally work with continuous numeric data.
9. Regression model is uniquely determined by the data.
10. Regression modeling tools are found in almost all statistical as well as data mining packages.