CHAPTER 2

Linear Regression

Please download the sample Excel files from https://github.com/hhohho/Learn-Data-Mining-through-Excel-2 for this chapter's exercises.

General Understanding

Linear regression is a predictive model in which training data is employed to construct a linear model to make predictions on the scoring data. When we talk about a linear model, we mean that the relationship between the target (dependent variable) and the attribute(s) (independent variables) is linear. It is a convention to use the terminologies "independent variable" and "dependent variable" in regression analysis. Therefore, in this chapter, we will replace attribute with independent variable and substitute dependent variable for target.

There might be one or more independent variables in linear regression analysis. When there is only one independent variable, the linear model is expressed by the commonly seen linear function $y = mx + b$, where y is the dependent variable, m is the slope of the line, and b is the y-intercept. In most cases, there is more than one independent variable; therefore, the linear model is represented as $y = m_1x_1 + m_2x_2 + \ldots + m_nx_n + b$, where there are n independent variables and m_i is the coefficient associated with a specific independent variable x_i, for $i = 1 \ldots n$. To construct such a linear model, we need to find the values of m_i and b, based on the known y and x_i values in the training dataset.

Let's start a scenario to learn what linear regression can do. In a southern beach town, Tommy, the manager of a supermarket is thinking about predicting ice cream sales based on the weather forecast. He has collected some data that relate weekly averaged daily high temperatures to ice cream sales during a summer season. The training data are presented in Table 2-1.

H. Zhou, *Learn Data Mining Through Excel*, https://doi.org/10.1007/978-1-4842-9771-1_2

Table 2-1. *Ice cream sale vs. temperature*

Temperature (F)	Ice Cream Sale (Thousands of Dollars)
91	89.8
87	90.2
86	81.1
88	83.0
92.8	90.9
95.2	119.0
93.3	94.9
97.7	132.4

Enter the preceding data into an Excel worksheet. Observe that in this scenario, the Temperature is the only independent variable, and the Ice Cream Sale is the dependent variable. Based on the given data, Tommy draws a scatter plot. The chart looks like Figure 2-1.

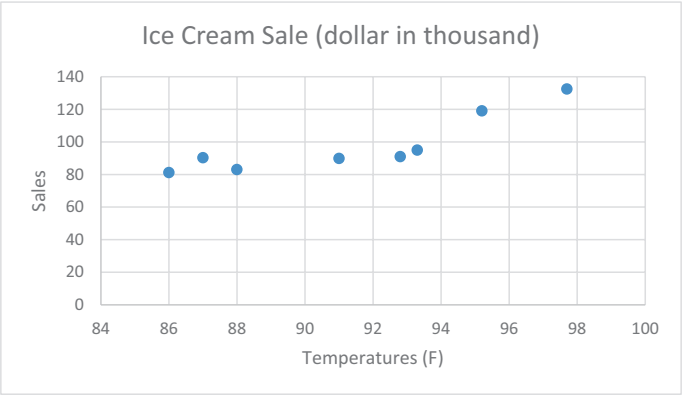

Figure 2-1. *Temperatures vs. ice cream sales*

Based on the chart in Figure 2-1, if the average of the daily high temperatures of next week is predicted to be 88.8 Fahrenheit degrees, Tommy finds it difficult to predict the sales since there is no dot directly matching to the temperature 88.8. Right-click on a dot; a small menu should show up. On the menu, click Add Trendline as shown in Figure 2-2.

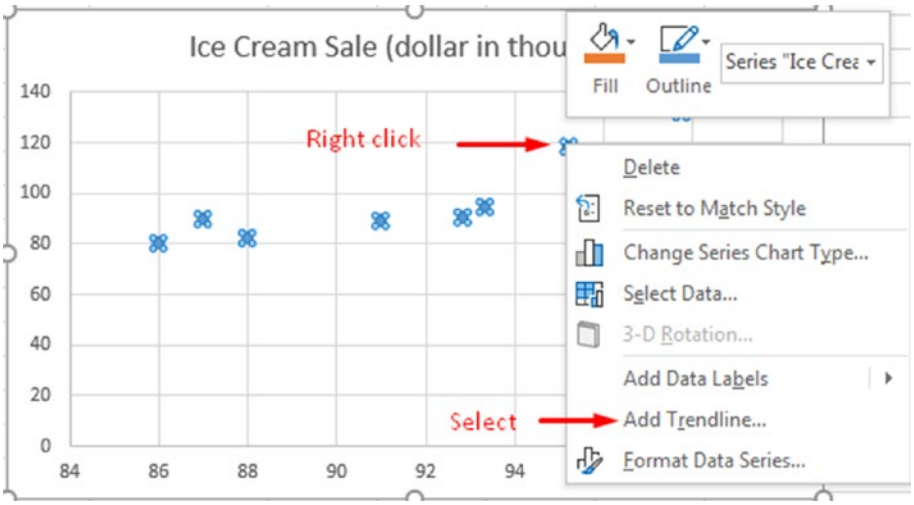

Figure 2-2. *Add Trendline*

A trendline shows up in the chart as illustrated by Figure 2-3. Be aware that by default, the line in Figure 2-3 is not a solid red line and the equation won't show up. It would be a good exercise to figure out how to format the trendline to be a solid red line and how to display the equation on the chart. To display the equation on the chart, right-click on the trendline in the chart ➤ select Format Trendline… ➤ check "Display Equation on chart".

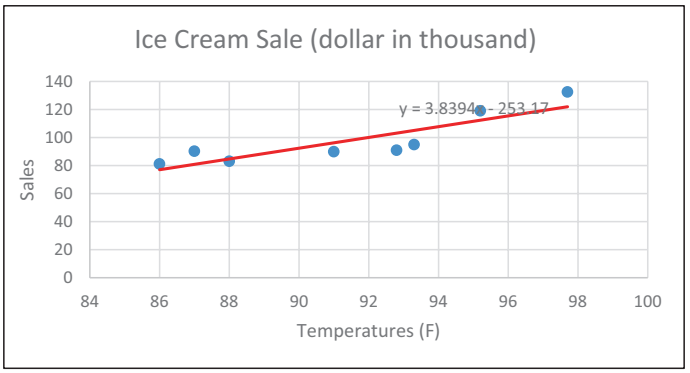

Figure 2-3. *Using trendline to predict*

With this linear line, Tommy can then estimate that when the temperature = 88.8, the sales would be about $89k. Using the equation $y = 3.8394x - 253.17$, where `m` = 3.8394 and `b` = -253.17, Tommy can predict his ice cream sales more precisely:

```
3.8394 × 88.8 - 253.17 = $87.8k.
```

The linear equation here is the linear data mining model for this specific linear regression study. We understand that for this type of linear equation `y = mx + b`, `m` and `b` are the two determinants. Once `m` and `b` are found, the model is constructed and finalized. Certainly, for this simple case, the model is represented by `m` and `b`, that is, the parameter set composed of `m` and `b` is the model. The model construction process is to find `m` and `b`.

How are `m` and `b` obtained? From Figure 2-3, we can tell that the trendline does not pass through every point, it only follows the trend. There could be multiple distinct trendlines if several people are drawing it manually and independently. How does Excel come up with this specific trendline?

To find a specific linear equation in the form of `y = mx + b`, the least square method is employed. Let me explain what the least square method is.

Observe the point corresponding to temperature = 91, the dot is not on the trendline. This indicates that there is an error (we can understand it as difference) between the trendline (the predicted value) and the actual data. In fact, all the data points in Figure 2-3 have small errors. The sum of the squares of the errors can be represented as

$$E = \sum_{i=1}^{n} \left(y_i - \left(mx_i + b \right) \right)^2 \tag{2-1}$$

where y_i is the actual ice cream sale corresponding to the temperature value x_i in the training dataset. The goal is to minimize the sum of the errors. For this purpose, we can take the partial derivatives:

$$\frac{\partial E}{\partial x} = 0 \text{ and } \frac{\partial E}{\partial y} = 0 \tag{2-2}$$

By solving the preceding two partial derivative equations, we can obtain the values of `m` and `b` as

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n\left(\sum_{i=1}^{n} x_i^{\,2} \right) - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{2-3}$$

$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{2-4}$$

Learn Linear Regression Through Excel

In Equations (2-3) and (2-4), x_i is the known averaged daily high temperature and y_i is the actual ice cream sale corresponding to x_i. They are from the training dataset shown in Table 2-1.

Open the file Chapter2-1a.xlsx. Enter the texts and formulas in columns C and D as shown in Figure 2-4 to compute the values of `m` and `b`. The formulas in Figure 2-4 are based on Equations (2-3) and (2-4). Observe that in Figure 2-4, temperature is renamed as `x` and ice cream sale is renamed as `y` so that the setup follows Equations (2-3) and (2-4) more closely.

	A	B	C	D
1	x	y		
2	91	89.8		
3	87	90.2	SUM(X)	=SUM(A:A)
4	86	81.1	SUM(Y)	=SUM(B:B)
5	88	83	SUM(XY)	=SUMPRODUCT(A:A, B:B)
6	92.8	90.9	SUM(X^2)	=SUMPRODUCT(A:A,A:A)
7	95.2	119	n	=COUNT(A:A)
8	93.3	94.9	m	=(D7*D5-D3*D4)/(D7*D6-D3^2)
9	97.7	132.4	b	=(D6*D4-D3*D5)/(D7*D6-D3^2)

Figure 2-4. *Computing m and b via least square approximation*

In Figure 2-4, A:A and B:B are used to represent all cells in column A and column B, respectively. Thus, =SUM(A:A) and =SUM(B:B) will sum up all cells in column A and column B, respectively. Also, an Excel function SUMPRODUCT is used to simplify the computation between two arrays. An important understanding of SUMPRODUCT is that the two arrays must be of the same type and length. Here, the same type means that the two arrays can be both a column and a row. With both `m` and `b` computed, Tommy can accomplish his predictions as shown in Figure 2-5 (enter the formula `=D$8*C12 + D$9` inside cell D12 and autofill from D12 to D14).

	A	B	C	D	E
1	x	y			
2	91	89.8			
3	87	90.2	SUM(X)	731	
4	86	81.1	SUM(Y)	781.3	
5	88	83	SUM(XY)	71851.77	
6	92.8	90.9	SUM(X^2)	66915.06	
7	95.2	119	n	8	
8	93.3	94.9	m	3.83943386	
9	97.7	132.4	b	-253.165769	
10					
11			Scoring Data	Prediction	
12			88.8	87.8	
13			96.9	118.9	
14			94.7	110.4	
15					

Figure 2-5. *Prediction based on a linear regression model*

A much more efficient way to compute `m` and `b` is to make use of the two functions SLOPE and INTERCEPT for this specific example. Follow these instructions to practice the use of the two functions:

1. Input the `x` and `y` values as shown Figure 2-6. Note: input `x` and `y` values only. Columns C and D are empty.
2. Enter `m` in cell C2 and `b` in cell D2.
3. Enter the formula =SLOPE(B2:B9, A2:A9) in cell C3.

	A	B	C	D	E
1	X	Y			
2	91	89.8			
3	87	90.2			
4	86	81.1			
5	88	83			
6	92.8	90.9			
7	95.2	119			
8	93.3	94.9			
9	97.7	132.4			
10					

Figure 2-6. *Prepare data to use functions SLOPE and INTERCEPT*

4. Enter the formula =INTERCEPT(B2:B9, A2:A9) in cell D3.

The result should look like Figure 2-7.

	A	B	C	D	E
1	x	y			
2	91	89.8	m	b	
3	87	90.2	3.8394	-253.2	
4	86	81.1			
5	88	83			
6	92.8	90.9			
7	95.2	119			
8	93.3	94.9			
9	97.7	132.4			

Figure 2-7. *The result obtained by the SLOPE and INTERCEPT functions*

Another way to obtain the values of `m` and `b` is through the Data Analysis tool of Excel. This is a popular approach to take by many people, though I dislike it. The disadvantage of using Data Analysis tool for regression is that once the coefficients and intercept value are obtained, they won't update along with the changes in the training data.

By default, our Excel does not have Data Analysis tool available. To make it available, we must enable the built-in add-in Analysis ToolPak in our Excel. I will show you how to enable Analysis ToolPak and Solver in our Excel later when we need to use Solver. As of right now, we do not need to use the Data Analysis tool.

The results of the preceding learning processes are stored in the file Chapter2-1b.xlsx.

Learn Multiple Linear Regression Through Excel

Tommy has successfully predicted his ice cream sales, but he wants to improve his prediction even more. After some studies, he collected additional information such as the number of tourists (obtained from hotels) and the number of sunny days in a week. His updated training dataset is presented in Table 2-2.

Table 2-2. *Ice cream sale vs. temperature, tourists, and sunny days*

Temperature (F)	Tourists	Sunny Days	Ice Cream Sale (dollar in thousand)
91	998	4	89.8
87	1256	7	90.2
86	791	6	81.1
88	705	5	83
92.8	1089	3	90.9
95.2	1135	6	119
93.3	1076	4	94.9
97.7	1198	7	132.4

There are three independent variables in Table 2-2: temperature, tourists, and sunny days. When there is more than one independent variable, the linear regression is specifically called multiple linear regression. Let's denote temperature as x_1, tourists as x_2, sunny days as x_3, the y-intercept as b, and the sales as y. Our multiple linear regression equation is $y = m_1x_1 + m_2x_2 + m_3x_3 + b$. For such multiple linear regression, we need to make use of LINEST function to obtain the coefficients m_i and b. Follow these instructions to practice multiple linear regression through Excel:

1. Enter the data of Table 2-2 into cells A1:D9 in a new Excel worksheet, or open the file Chapter2-2a.xlsx.

The model construction process is to find the optimal values of m_1, m_2, m_3, and b. We are going to make use of the array function LINEST. Be advised that LINEST returns an array of values. The returned values are in the order of m_n, m_{n-1}, …, m_1, and b, the reverse order to the way in which the formula is expressed. This necessitates the use of the INDEX function to fetch individual values from the returned array. The INDEX function is designed to work with a matrix or table. It requires a row index and a column index to locate an element in a table. As LINEST returns an array which has only one row, the row index is always 1, while the column indexes are 1, 2, 3, and 4, respectively.

2. Enter the texts "Sunny Days", "Tourists", "Temperature", and "Y-intercept" in cells A11:A14.

3. Enter numbers 1, 2, 3, and 4 in cells B11:B14. Our worksheet looks like Figure 2-8. I will explain why we need to enter these numbers soon.

	A	B	C	D
1	Temperature (F)	Tourists	Sunny Days	Ice Cream Sale (dollar in thousand)
2	91	998	4	89.8
3	87	1256	7	90.2
4	86	791	6	81.1
5	88	705	5	83
6	92.8	1089	3	90.9
7	95.2	1135	6	119
8	93.3	1076	4	94.9
9	97.7	1198	7	132.4
10				
11	Sunny Days	1		
12	Tourists	2		
13	Temperature	3		
14	Y-intercept	4		

Figure 2-8. *Set up the table to find coefficients and y-intercept*

4. Enter the following formula in cell C11 and hit Enter:

   ```
   =INDEX(LINEST(D$2:D$9,A$2:C$9,TRUE,TRUE),1,B11)
   ```

   In this formula, the first input argument in the function LINEST is D2:D9, the dependent variable values. The second input argument is A2:C9 which contains values for x_1, x_2, and x_3. The returned array from LINEST is fed to the function INDEX as an input. Since B11 = 1, the function INDEX fetches the first element in the returned array, which is m_3. Note again, LINEST returns an array that arranges the coefficients in the order of m_n, m_{n-1}, m_{n-2}, ..., m_1, b. In this specific example, the data in the returned array are m_3, m_2, m_1, and b for independent variables Sunny Days, Tourists, Temperature, and Y-intercept.

   Refer to Figure 2-9.

	A	B	C	D
1	Temperature (F)	Tourists	Sunny Days	Ice Cream Sale (dollar in thousand)
2	91	998	4	89.8
3	87	1256	7	90.2
4	86	791	6	81.1
5	88	705	5	83
6	92.8	1089	3	90.9
7	95.2	1135	6	119
8	93.3	1076	4	94.9
9	97.7	1198	7	132.4
10				
11	Sunny Days	1	=INDEX(LINEST(D$2:D$9,A$2:C$9, TRUE, TRUE), 1, B11)	
12	Tourists	2		
13	Temperature	3		
14	Y-intercept	4		

Figure 2-9. *Using INDEX and LINEST functions to obtain coefficients*

5. Autofill from cell C11 to cell C14. Pay attention to the use of B11, B12, B13, and B14.

   When we autofill from cell C11 to C14, the preceding formula automatically becomes `=INDEX(LINEST(D$2:D$9,A$2:C$9,TRUE,TRUE),1,B12)` in cell C12. Because B12 = 2, the formula in C12 correctly fetches the coefficient for Tourists (m_2). The same logic applies to the formulas in C13 and C14. By pre-entering numbers in cells B11:B14, once we have the initial formula entered correctly, we can then obtain all needed values by autofill.

   This is a common technique. We certainly do not want to enter the formula multiple times. Note, the function LINEST is repeatedly called for every coefficient.

6. Our last step is to examine how well the new model is. Enter the text "Predicted" in cell E1.

7. Enter the following formula in cell E2 and then autofill to cell E9:

   ```
   =A2*C$13+B2*C$12+C2*C$11+C$14
   ```

   As mentioned before, the parameter set, that is, the coefficients and the Y-intercept, represent the linear regression model. We want to examine how good the model is; therefore, we need to generate the predicted sales in column E.

8. To examine the quality of the preceding linear regression model, we can further compute the errors based on Equation (2-1). Enter the text "Error" in cell F1, and the formula `=POWER(D2-E2,2)` in cell F2, then autofill from cell F2 to F9.

9. Enter the text "Sum of Errors" in cell D11 and the formula `=SUM(F2:F9)` in cell E11. Our worksheet looks like Figure 2-10.

	A	B	C	D	E	F
1	Temperature	Tourists	Sunny Days	Ice Cream Sale (dollar in thousand)	Predicted	Error
2	91	998	4	89.8	88.7706	1.059665
3	87	1256	7	90.2	90.390726	0.036376
4	86	791	6	81.1	81.085358	0.000214
5	88	705	5	83	83.195593	0.038256
6	92.8	1089	3	90.9	89.847272	1.108235
7	95.2	1135	6	119	117.19569	3.25555
8	93.3	1076	4	94.9	97.80896	8.462051
9	97.7	1198	7	132.4	133.0058	0.367
10						
11	Sunny Days	1	5.95647	Sum of Errors	14.327347	
12	Tourists	2	-0.0013			
13	Temperature	3	3.97541			
14	Y-intercept	4	-295.47			

Figure 2-10. *A model of multiple linear regression*

Recall that the coefficients obtained through the function LINEST guarantee that the error value in cell E11 is minimized. You can manually modify the coefficients to examine if you can obtain a smaller error value. I will leave this as an exercise for you to practice.

The coefficient for the independent variable "Tourists" is negative. This simply means when there are more tourists, the ice cream sales will be smaller. However, in this specific example, this coefficient is close to zero, indicating that the number of tourists does not have much impact on the ice cream sales.

The complete work can be found in the file Chapter2-2b.xlsx.

Reinforcement Exercises

Practice is the key to success. It is a good idea to open Chapter2-HW.xlsx to practice linear regression by yourself. Chapter2-HW-withAnswers.xlsx provides one way of solutions to reference.

Review Points

We have reached the end of Chapter 2. For this chapter, please review the following concepts and Excel skills:

1. Linear regression model
2. Least square method
3. Multiple linear regression
4. Functions SUM, COUNT, SLOPE, INTERCEPT, LINEST, SUMPRODUCT, INDEX, and POWER