CHAPTER 6

Logistic 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

Logistic regression can be thought of as a special case of linear regression when the predicted outcomes are categorical. If there are only two outcomes, the logistic regression is called binomial logistic regression which is the most popular one. If there are more than two outcomes, the logistic regression is called multinomial logistic regression. In this chapter, we are learning binomial logistic regression in Excel.

Binomial logistic regression can also be thought of as a special case of LDA, but its mechanism of achieving good separability between data groups is different from that of LDA.

The name of logistic regression comes from the fact that logistic function is used in the model. Equation (6-1) shows the logistic function in its general form (where a, b, and c are all positive numbers).

$$f(x) = \frac{c}{1 + a \cdot b^{-kx}} \tag{6-1}$$

The well-known sigmoid function is a special case of the logistic function.

$$s(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{6-2}$$

H. Zhou, *Learn Data Mining Through Excel*, https://doi.org/10.1007/978-1-4842-9771-1_6

Logistic regression is a statistical model, capable of estimating the probability of the occurrence of an event. Let 1 represent the case when the event happens and 0 represent the case when the event does not happen. Thus, P(1) is the probability of the event and `P(0) = 1- P(1)`. Note, the probability must be between 0 and 1 inclusively. Finally, `P(1)/(1-P(1))` is the odds for the event to happen.

Suppose the occurrence of the event is dependent on the variables $x_1, x_2, \cdots, x_n$. In logistic regression, it is believed that the log of the odds is a linear function of x_i, as depicted in Equation (6-3).

$$\ln\left(\frac{P(1)}{1-P(1)}\right) = m_1x_1 + m_2x_2 + \cdots + m_nx_n + b \tag{6-3}$$

Solving for P(1), we can get the logistic function as shown in Equation (6-4).

$$P(1) = \frac{1}{1 + \mathrm{e}^{-(m_1x_1 + m_2x_2 + \cdots + m_nx_n + b)}} \tag{6-4}$$

Logistic regression does not maximize the probability of P(1) by optimizing the coefficients $m_1, m_2, \cdots, m_n, b$ (here, let's take `b` as a coefficient; later we will know that b is also called "bias" in machine learning community). Instead, it tries to maximize the likelihood associated with each data sample. What then is likelihood? Simply speaking, likelihood is the probability of a sample to match the actual event outcome (either 1 or 0). Still confused? Let me use one example to explain what is likelihood.

Assume there are two data records A and B in the training dataset. A has the event happened, that is, its event outcome is 1, and B does not have the event happened, that is, its event outcome is 0. Through the logistic function, the P(1) for A and the P(1) for B are calculated as 0.8 and 0.6, respectively. Because A's actual event outcome is 1, A's likelihood is then its probability for the event to happen, that is, P(1) which is 0.8. On the other hand, because B's actual outcome is 0, its likelihood is its probability for the event not to happen, that is, its matching event outcome becomes 0. Thus, B's likelihood is P(0) instead of P(1), and `P(0) = 1 - P(1) = 0.4`.

Logistic regression tries to maximize the likelihoods of all samples as a whole. By maximizing the likelihoods, logistic regression achieves good separability between two groups. In implementation, it is common to use the logarithm of a likelihood, either natural logarithm or base 10 logarithm. The advantage of logarithm lies in the fact that it decreases rapidly when the predicted likelihood diverges from the actual event outcome, regardless if the actual event outcome is 1 or 0. Because when the likelihood is 0, the calculation of log(0) will result in an error. Thus, log(0) is treated as a predefined minimum number.

In implementation, logistic regression can also be achieved by minimizing the log loss of likelihoods. The log loss is the negation of the log of a likelihood. Thus, when the actual outcome is 1, the log loss is computed as -log(P(1)), and when the actual outcome is 0, the log loss is computed as -log(1-P(1)). Here, log(0) is treated as a predefined maximum number. Note, the log loss increases rapidly when the predicted likelihood diverges from the actual event outcome. Figure 6-1 shows the log loss when the event outcome is 1. Here, the log(0) is predefined as 0.000001.

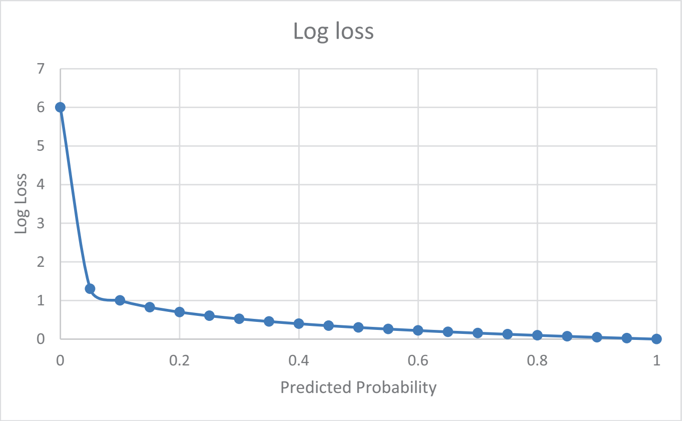


Figure 6-1. *The calculation of log loss when event = 1*

With this understanding, our approach of logistic regression is to use linear regression to obtain a set of coefficients first, then apply Solver to optimize these coefficients by either maximizing the likelihoods as a whole or minimizing the log loss as a whole.

Learn Logistic Regression Through Excel

Data inside the two files Chapter6-1a.xlsx and Chapter6-2a.xlsx are the same. They simulate the following scenario:

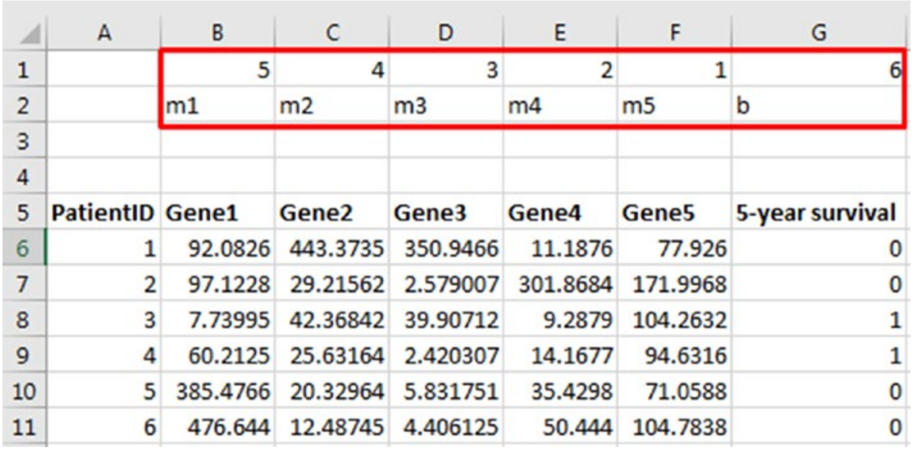
> *Suppose there are five genes whose expression levels can be used to predict the five-year survival probability of patients with a certain cancer disease. Value 1 indicates the patient did survive five years, and 0 means they did not.*

As stated earlier, in our logistic regression practices, we are going to apply Solver to optimize the model parameters (coefficients) by either maximizing the likelihoods as a whole or minimizing the log loss as a whole.

By Means of Maximizing Log Likelihoods

Please open the file Chapter6-1a.xlsx in which there are 87 samples and there are 4 blank rows on the top of the worksheet. These blank rows are left there on purpose.

Excel functions INDEX and LINEST will be used again to obtain the coefficients based on the least square method. As the first value returned by LINEST is for Gene 5, and the last value for coefficient b, we need to set up the top two blank rows as shown in Figure 6-2. Such a setup is for easy autofill later.



	A	B	C	D	E	F	G
1		5	4	3	2	1	6
2		m1	m2	m3	m4	m5	b
3							
4							
5	PatientID	Gene1	Gene2	Gene3	Gene4	Gene5	5-year survival
6	1	92.0826	443.3735	350.9466	11.1876	77.926	0
7	2	97.1228	29.21562	2.579007	301.8684	171.9968	0
8	3	7.73995	42.36842	39.90712	9.2879	104.2632	1
9	4	60.2125	25.63164	2.420307	14.1677	94.6316	1
10	5	385.4766	20.32964	5.831751	35.4298	71.0588	0
11	6	476.644	12.48745	4.406125	50.444	104.7838	0

Figure 6-2. *Set up the data for least square method coefficients*

Follow these instructions to exercise logistic regression in Excel:

1. Enter the following formula in both cells B3 and B4:

   ```
   =INDEX(LINEST($G$6:$G$92,$B$6:$F$92),1,B1)
   ```

   The reason to enter the same formula in two cells is that we want to keep one set of coefficients unchanged while let another set be modified by Solver.

Note, the “known_ys” parameter inside the LINEST function is cells `$G$6:$G$92`, which have only two values: 1 and 0. This indicates that if our original target values are not 1 and 0, they must be categorized into such.

2. Autofill B3 to G3, and B4 to G4.

3. Enter the text “m1x1+m2x2+...+b” in cell H5. Column H represents the right-side expression of Equation (6-3).

4. In cell H6, enter the following formula and autofill to cell H92:

   ```
   =SUMPRODUCT($B$4:$F$4,B6:F6)+$G$4
   ```

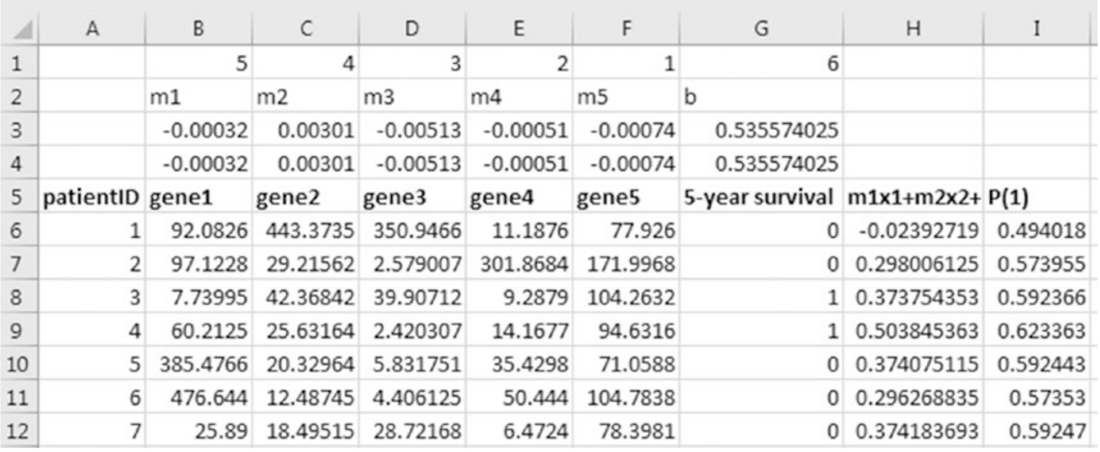
   The function SUMPRODUCT computes the product between two arrays. Note that absolute references are used for cell references B4, F4, and G4.

5. Enter the text “P(1)” in cell I5. Column I represents the five-year survival probability computed based on the values in column H.

6. In cell I6, enter the following formula and autofill to cell I92:

   ```
   =1/(1+EXP(0-H6))
   ```

   This formula implements Equation (6-4). `(0-H6)` represents $-(m_1x_1 + m_2x_2 + \cdots + m_nx_n + b)$, while `EXP(0-H6)` stands for $e^{-(m_1x_1+m_2x_2+\cdots+m_nx_n+b)}$. Function EXP returns the power of the constant e (the base of the natural logarithm). By now, part of our worksheet should look like Figure 6-3.



	A	B	C	D	E	F	G	H	I
1		5	4	3	2	1	6		
2		m1	m2	m3	m4	m5	b		
3		-0.00032	0.00301	-0.00513	-0.00051	-0.00074	0.535574025		
4		-0.00032	0.00301	-0.00513	-0.00051	-0.00074	0.535574025		
5	patientID	gene1	gene2	gene3	gene4	gene5	5-year survival	m1x1+m2x2+	P(1)
6	1	92.0826	443.3735	350.9466	11.1876	77.926	0	-0.02392719	0.494018
7	2	97.1228	29.21562	2.579007	301.8684	171.9968	0	0.298006125	0.573955
8	3	7.73995	42.36842	39.90712	9.2879	104.2632	1	0.373754353	0.592366
9	4	60.2125	25.63164	2.420307	14.1677	94.6316	1	0.503845363	0.623363
10	5	385.4766	20.32964	5.831751	35.4298	71.0588	0	0.374075115	0.592443
11	6	476.644	12.48745	4.406125	50.444	104.7838	0	0.296268835	0.57353
12	7	25.89	18.49515	28.72168	6.4724	78.3981	0	0.374183693	0.59247

Figure 6-3. *P(1) computed*

At this point, it would be interesting to know how well the coefficients generated by the least square method can differentiate the two groups of patients: survived or did not survive five years.

7. Enter "Outcome" in cell J5 and the formula `=IF(I6>0.5,1,0)` in cell J6. This formula specifies if the probability in column I is greater than 0.5 (here, 0.5 is the cutoff), the predicted outcome of the patient is 1, otherwise 0.

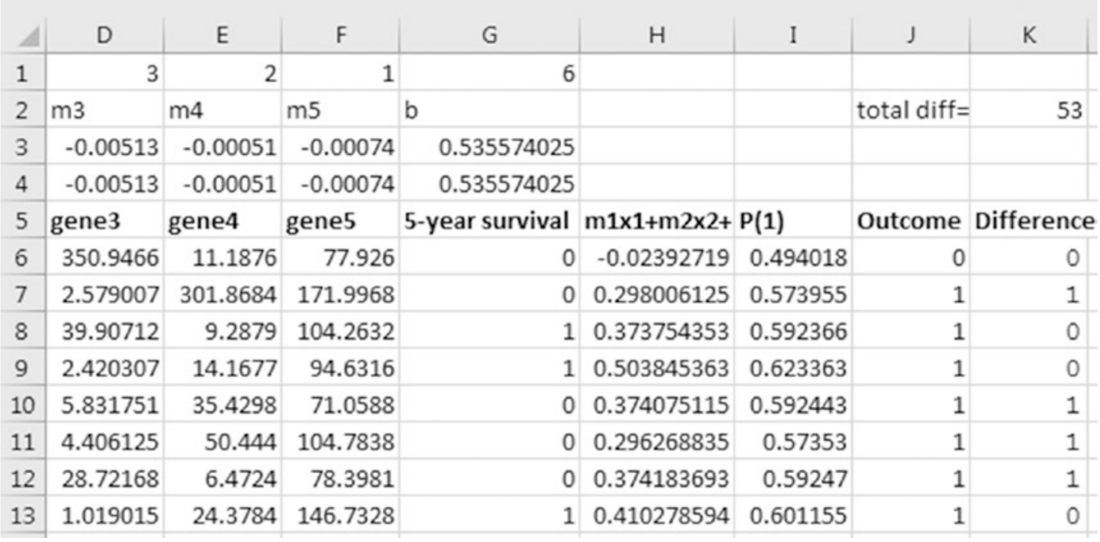
8. Autofill from cell J6 to cell J92.

9. Enter "Difference" in cell K5 and enter the formula `=IF(G6=J6,0,1)` in cell K6. This formula asserts that if the predicted outcome in column J does match the actual outcome in column G, 0 is returned; otherwise, 1 is returned.

10. Autofill from cell K6 to cell K92.

11. Enter "total diff=" in cell J2, and enter the formula `=SUM(K6:K92)` in cell K2. By counting how many 1s there are in the array K6:K92, we can tell how many predictions do not match the actual outcomes.

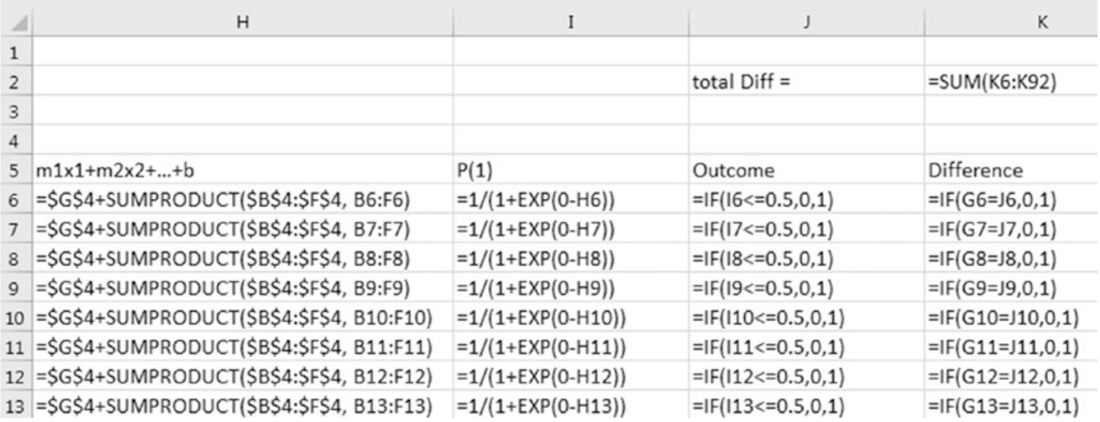
    Note that the value in cell K2 is 53. This indicates that 53 out of 87 samples are falsely classified. This is illustrated in Figure 6-4, which displays part of our current worksheet.



	D	E	F	G	H	I	J	K
1	3	2	1	6				
2	m3	m4	m5	b			total diff=	53
3	-0.00513	-0.00051	-0.00074	0.535574025				
4	-0.00513	-0.00051	-0.00074	0.535574025				
5	gene3	gene4	gene5	5-year survival	m1x1+m2x2+	P(1)	Outcome	Difference
6	350.9466	11.1876	77.926	0	-0.02392719	0.494018	0	0
7	2.579007	301.8684	171.9968	0	0.298006125	0.573955	1	1
8	39.90712	9.2879	104.2632	1	0.373754353	0.592366	1	0
9	2.420307	14.1677	94.6316	1	0.503845363	0.623363	1	0
10	5.831751	35.4298	71.0588	0	0.374075115	0.592443	1	1
11	4.406125	50.444	104.7838	0	0.296268835	0.57353	1	1
12	28.72168	6.4724	78.3981	0	0.374183693	0.59247	1	1
13	1.019015	24.3784	146.7328	1	0.410278594	0.601155	1	0

Figure 6-4. *Examine the temporary classification result*

If your worksheet has a result different from what is shown in Figure 6-4, double check the formulas, especially those in cells H6:K6. Figure 6-5 displays some formulas entered so far. Make sure that yours are the same as those in Figure 6-5.



	H	I	J	K
1				
2			total Diff =	=SUM(K6:K92)
3				
4				
5	m1x1+m2x2+...+b	P(1)	Outcome	Difference
6	=$G$4+SUMPRODUCT($B$4:$F$4, B6:F6)	=1/(1+EXP(0-H6))	=IF(I6<=0.5,0,1)	=IF(G6=J6,0,1)
7	=$G$4+SUMPRODUCT($B$4:$F$4, B7:F7)	=1/(1+EXP(0-H7))	=IF(I7<=0.5,0,1)	=IF(G7=J7,0,1)
8	=$G$4+SUMPRODUCT($B$4:$F$4, B8:F8)	=1/(1+EXP(0-H8))	=IF(I8<=0.5,0,1)	=IF(G8=J8,0,1)
9	=$G$4+SUMPRODUCT($B$4:$F$4, B9:F9)	=1/(1+EXP(0-H9))	=IF(I9<=0.5,0,1)	=IF(G9=J9,0,1)
10	=$G$4+SUMPRODUCT($B$4:$F$4, B10:F10)	=1/(1+EXP(0-H10))	=IF(I10<=0.5,0,1)	=IF(G10=J10,0,1)
11	=$G$4+SUMPRODUCT($B$4:$F$4, B11:F11)	=1/(1+EXP(0-H11))	=IF(I11<=0.5,0,1)	=IF(G11=J11,0,1)
12	=$G$4+SUMPRODUCT($B$4:$F$4, B12:F12)	=1/(1+EXP(0-H12))	=IF(I12<=0.5,0,1)	=IF(G12=J12,0,1)
13	=$G$4+SUMPRODUCT($B$4:$F$4, B13:F13)	=1/(1+EXP(0-H13))	=IF(I13<=0.5,0,1)	=IF(G13=J13,0,1)

Figure 6-5. *Examine the formulas entered so far*

12. Continue by entering “Likelihood” in cell L5.

13. Enter `=IF(G6=1,I6,1-I6)` in cell L6. This formula asserts that if the actual outcome is 1, then the likelihood is I6; otherwise, the likelihood is `1-I6`.

14. Autofill from cell L6 to cell L92.

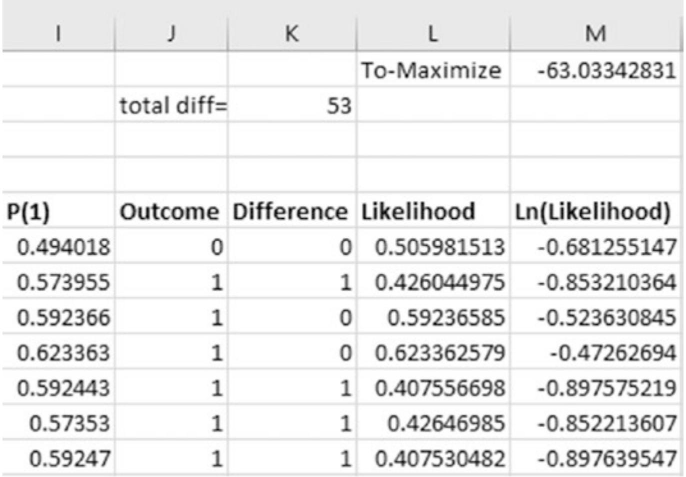
15. We need to treat all the likelihood values as a whole. There can be different ways to do it, such as summing all the values in L6:L92, or computing the product of L6:L92. What I am introducing you here is the sum of the logarithm of each likelihood value. Enter the text "Ln(Likelihood)" in cell M5.

16. Enter the formula =IF(L6=0, -1000000, LN(L6)) in cell M6, then autofill from cell M6 to cell M92. Here, the predefined minimum is -1000000 when the likelihood is 0.

17. In cell L1, enter "To-Maximize".

18. Enter the formula `=SUM(M6:M92)` in cell M1. Cell M1 stores the value to be maximized by Solver.

Part of our worksheet should look like Figure 6-6.



I	J	K	L	M
			To-Maximize	-63.03342831
	total diff=	53		
P(1)	Outcome	Difference	Likelihood	Ln(Likelihood)
0.494018	0	0	0.505981513	-0.681255147
0.573955	1	1	0.426044975	-0.853210364
0.592366	1	0	0.59236585	-0.523630845
0.623363	1	0	0.623362579	-0.47262694
0.592443	1	1	0.407556698	-0.897575219
0.57353	1	1	0.42646985	-0.852213607
0.59247	1	1	0.407530482	-0.897639547

Figure 6-6. *Likelihood computed*

19. The next step is to maximize the value inside cell M1 by using Solver. Select cell M1 ➤ click the main tab Data ➤ click Solver.

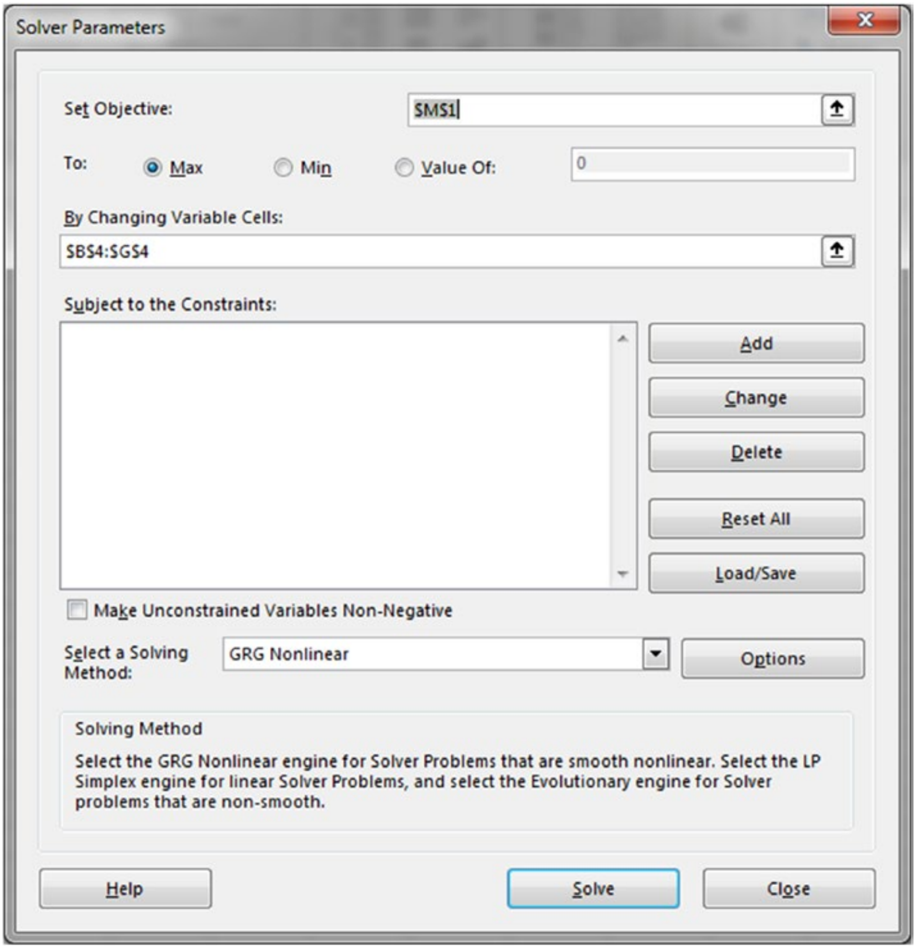
20. A menu shows up for us to select proper cells and set up requirements. Follow Figure 6-7 to

    a. Select cell $M$1 for "Set Objective".

    b. Select $B$4:$G$4 for "By Changing Variable Cells".

    c. Choose "Max" as shown in Figure 6-7.

    d. Select the solving method "GRG Nonlinear". Observe that the check box for "Make Unconstrained Variables Non-Negative" is not checked.



Solver Parameters

Set Objective: $M$1

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $B$4:$G$4

Subject to the Constraints:

□ Make Unconstrained Variables Non-Negative

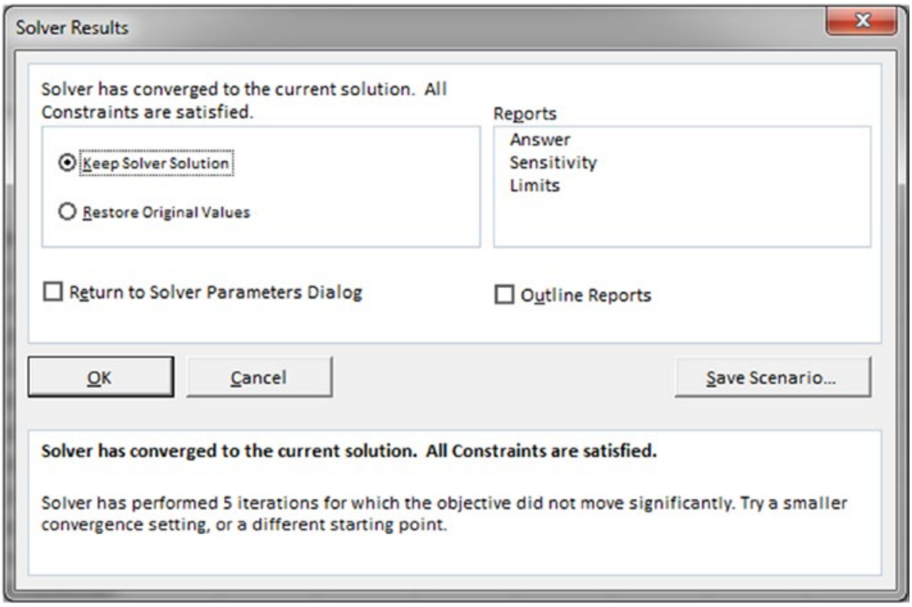
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 6-7. *Using Solver to maximize the likelihood*

21. Click the button Solve. Another menu shows up on which make sure that “Keep Solver Solution” is chosen as shown in Figure 6-8.



Solver Results

Solver has converged to the current solution. All Constraints are satisfied.

● Keep Solver Solution

○ Restore Original Values

□ Return to Solver Parameters Dialog

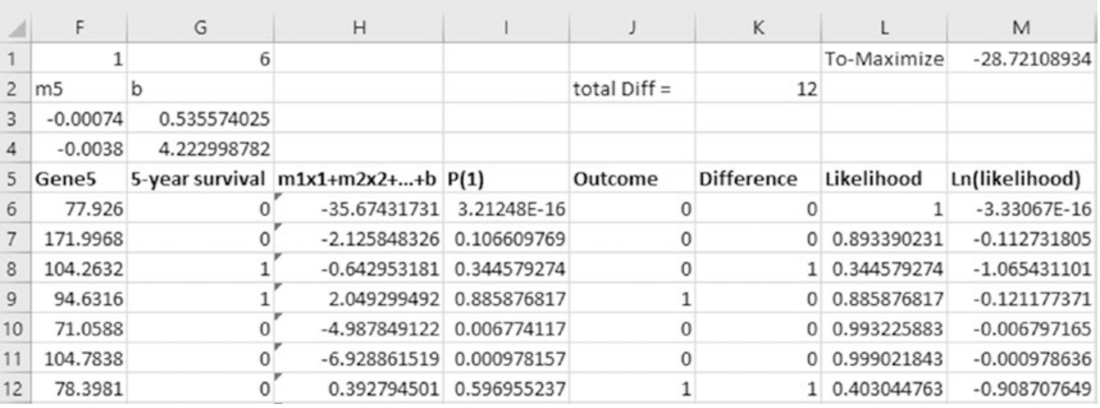
□ Outline Reports

Solver has converged to the current solution. All Constraints are satisfied.

Solver has performed 5 iterations for which the objective did not move significantly. Try a smaller convergence setting, or a different starting point.

Figure 6-8. *Keep Solver Solution*

The number of mistakenly classified samples is reduced to 12, as shown in Figure 6-9 (cell K2). Observe that the values inside cells B4:G4 are changed, that is, optimized by Solver to maximize the value inside cell M1. They are different from values inside B3:G3.



	F	G	H	I	J	K	L	M
1	1	6					To-Maximize	-28.72108934
2	m5	b			total Diff =	12		
3	-0.00074	0.535574025						
4	-0.0038	4.222998782						
5	Gene5	5-year survival	m1x1+m2x2+...+b	P(1)	Outcome	Difference	Likelihood	Ln(likelihood)
6	77.926	0	-35.67431731	3.21248E-16	0	0	1	-3.33067E-16
7	171.9968	0	-2.125848326	0.106609769	0	0	0.893390231	-0.112731805
8	104.2632	1	-0.642953181	0.344579274	0	1	0.344579274	-1.065431101
9	94.6316	1	2.049299492	0.885876817	1	0	0.885876817	-0.121177371
10	71.0588	0	-4.987849122	0.006774117	0	0	0.993225883	-0.006797165
11	104.7838	0	-6.928861519	0.000978157	0	0	0.999021843	-0.000978636
12	78.3981	0	0.392794501	0.596955237	1	1	0.403044763	-0.908707649

Figure 6-9. *The result of logistic regression by means of maximizing log likelihoods*

Recall at step 15, I mentioned that we can compute the value in cell M1 differently. For example, we can enter the formula `=SUM(L6:L92) or =PRODUCT(L6:L92)` in cell M1, or substitute LOG function for LN function in M6:M92, and then use Solver to maximize the value in M1 by optimizing the coefficients in cells B4:G4. I would like to challenge you to try these formulas and compare your results with the method used in this book. You can also try other methods, too.

The aforementioned process is stored in the file chapter6-1b.xlsx. Chapter6-1b.xlsx also includes some scoring data in A1:J101. Please take a look so that you are confident how to apply the constructed logistic regression model to scoring data. Again, the model is just a set of parameters. In this specific case, the set of parameters are in B4:G4.

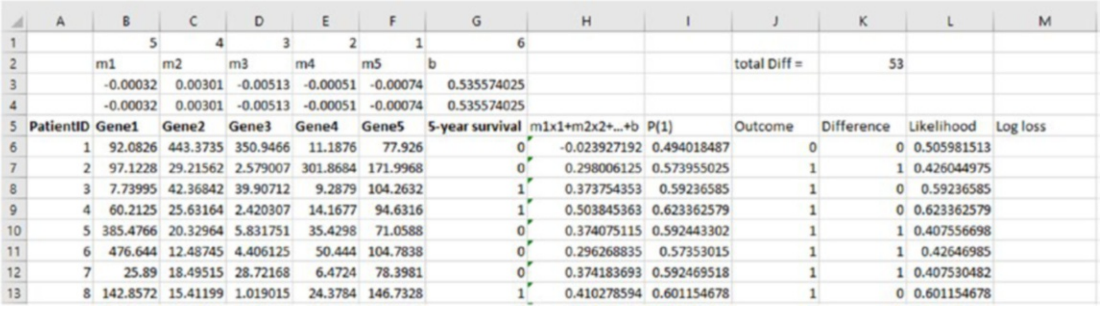
By Means of Minimizing Log Losses

"Loss function" is a phrase appears very often in machine learning. What is it? Well, we have used it before in linear regression. In linear regression, the least square method is used to minimize the prediction error, commonly called mean square error or quadratic loss, which is the sum of the squares of the differences between predictions and actual values (see Equation 2-1). Here, Equation 2-1 is a bona fide loss function. Thus, we can tell that a loss function is just a mathematical function that quantifies the error between the predicted and actual values in a machine learning method. Indeed, we can understand "loss" as "error" in this sense.

Generally speaking, there are two types of losses in machine learning: regression loss and classification loss. Equation 2-1 is a type of regression loss. The "total diff", that is, the number of mistakenly classified samples in cell K2 which we used earlier in step 11, is a type of classification loss.

A loss function usually works with an optimization function which helps a machine learning model to minimize the loss. In this book, we won't explain optimization function, but we have been using one optimization function employed by the GRG Nonlinear algorithm of Excel Solver to maximize a target value. However, when using a loss function, we want to minimize the target value.

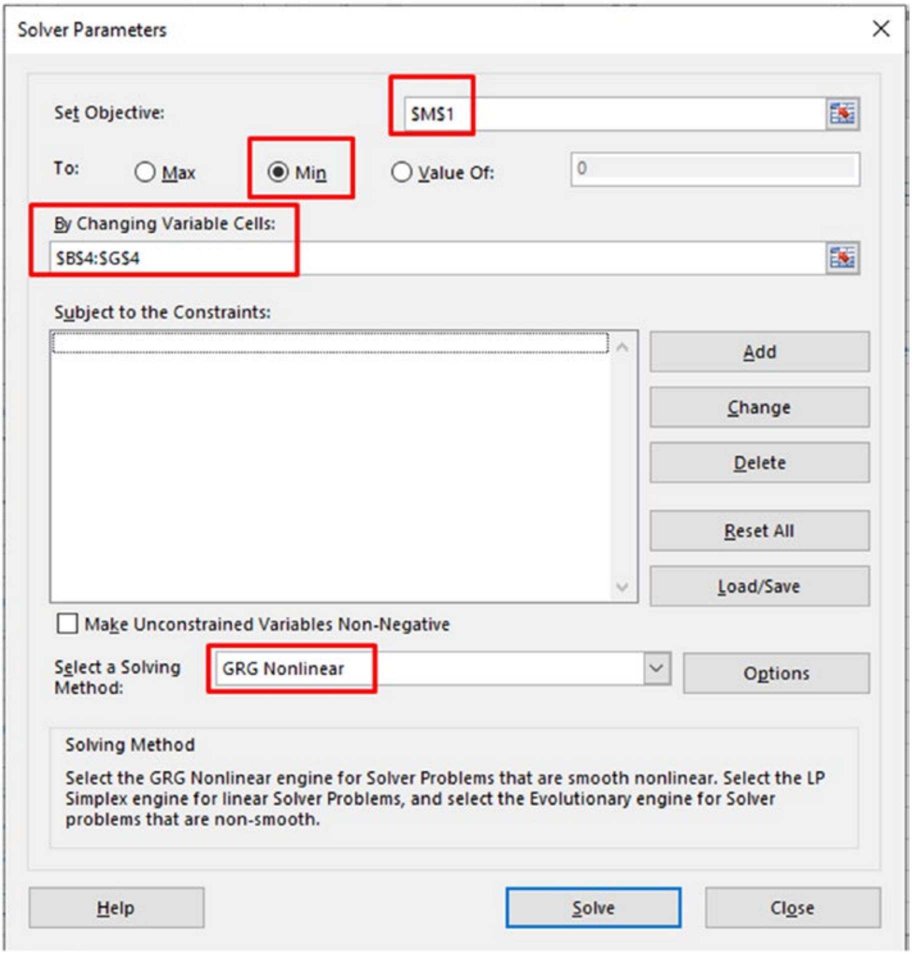
Let's open the file Chapter6-2a.xlsx, you shall notice that the work in this worksheet is near completion as shown in Figure 6-10. Yes, as explained before, log loss is the negation of log of likelihood. So, we do not need to repeat the steps up to the calculation of likelihoods but restart from step 16 to compute the log loss for each likelihood.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		5	4	3	2	1	6						
2		m1	m2	m3	m4	m5	b			total Diff =	53		
3		-0.00032	0.00301	-0.00513	-0.00051	-0.00074	0.535574025						
4		-0.00032	0.00301	-0.00513	-0.00051	-0.00074	0.535574025						
5	PatientID	Gene1	Gene2	Gene3	Gene4	Gene5	5-year survival	m1x1+m2x2+...+b	P(1)	Outcome	Difference	Likelihood	Log loss
6	1	92.0826	443.3735	350.9466	11.1876	77.926	0	-0.023927192	0.494018487	0	0	0.505981513	
7	2	97.1228	29.21562	2.579007	301.8684	171.9968	0	0.298006125	0.573955025	1	1	0.426044975	
8	3	7.73995	42.36842	39.90712	9.2879	104.2632	1	0.373754353	0.59236585	1	0	0.59236585	
9	4	60.2125	25.63164	2.420307	14.1677	94.6316	1	0.503845363	0.623362579	1	0	0.623362579	
10	5	385.4766	20.32964	5.831751	35.4298	71.0588	0	0.374075115	0.592443302	1	1	0.407556698	
11	6	476.644	12.48745	4.406125	50.444	104.7838	0	0.296268835	0.57353015	1	1	0.42646985	
12	7	25.89	18.49515	28.72168	6.4724	78.3981	0	0.374183693	0.592469518	1	1	0.407530482	
13	8	142.8572	15.41199	1.019015	24.3784	146.7328	1	0.410278594	0.601154678	1	0	0.601154678	

Figure 6-10. *The starting point of computing the log loss*

1. Enter the formula `=IF(L6=0, 1000000, -LN(L6))` in cell M6, then autofill from cell M6 to cell M92. Here, the predefined maximum loss is 1000000 when the likelihood is 0.
2. In cell L1, enter "To-Minimize".
3. Enter the formula `=SUM(M6:M92)` in cell M1. Cell M1 stores the value to be minimized by Solver.
4. The next step is to minimize the value inside cell M1 by using Solver. Select cell M1 ➤ click the main tab Data ➤ click Solver.
5. A menu shows up for us to select proper cells and set up requirements. Follow Figure 6-11 to
    a. Select cell $M$1 for "Set Objective".
    b. Select $B$4:$G$4 for "By Changing Variable Cells".
    c. Choose "Min" as shown in Figure 6-11.
    d. Select the solving method "GRG Nonlinear".



Set Objective: $M$1

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$B$4:$G$4

Subject to the Constraints:

□ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

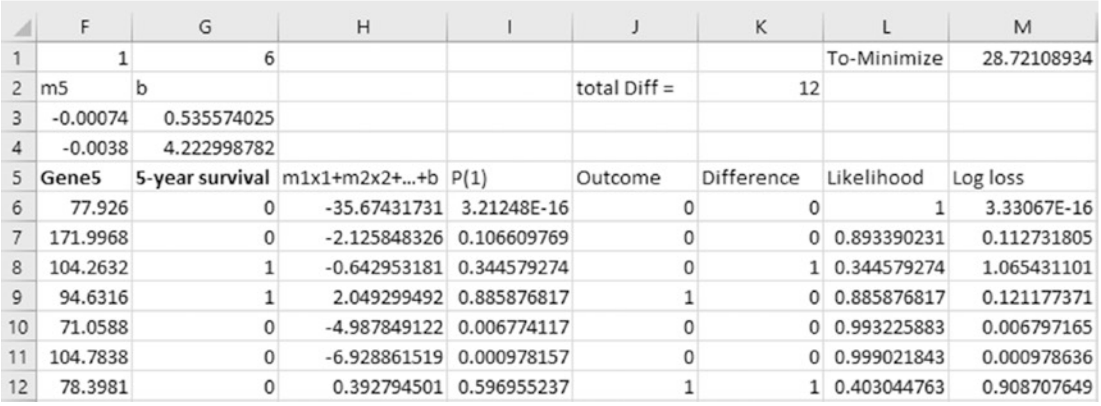
Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 6-11. *Using Solver to minimize the log loss*

6. Click the button Solve. Another menu shows up on which make sure that “Keep Solver Solution” is chosen.

Part of our worksheet should look like Figure 6-12.



	F	G	H	I	J	K	L	M
1	1	6					To-Minimize	28.72108934
2	m5	b			total Diff =	12		
3	-0.00074	0.535574025						
4	-0.0038	4.222998782						
5	Gene5	5-year survival	m1x1+m2x2+...+b	P(1)	Outcome	Difference	Likelihood	Log loss
6	77.926	0	-35.67431731	3.21248E-16	0	0	1	3.33067E-16
7	171.9968	0	-2.125848326	0.106609769	0	0	0.893390231	0.112731805
8	104.2632	1	-0.642953181	0.344579274	0	1	0.344579274	1.065431101
9	94.6316	1	2.049299492	0.885876817	1	0	0.885876817	0.121177371
10	71.0588	0	-4.987849122	0.006774117	0	0	0.993225883	0.006797165
11	104.7838	0	-6.928861519	0.000978157	0	0	0.999021843	0.000978636
12	78.3981	0	0.392794501	0.596955237	1	1	0.403044763	0.908707649

***Figure 6-12.** The result of logistic regression by means of minimizing log losses*

The completed result can be found in Chapter6-2b.xlsx. You shall notice that the result generated by log loss is the same as that by likelihood.

This summarizes another chapter. Certainly, Excel is fairly capable of carrying out logistic regression analysis.

Reinforcement Exercises

The reinforcement exercises make use of the Heart Disease dataset downloaded from UCI Machine Learning Repository at `http://archive.ics.uci.edu/dataset/45/heart+disease` (more details can be found at: `https://doi.org/10.24432/C52P4X`).

- Chapter6-HW-1.xlsx is for practicing the logistic regression process. Chapter6-HW-1-withAnswers.xlsx presents two solutions: one by maximizing log likelihood and one by minimizing log loss.
- Chapter6-HW-2-cross-validation.xlsx provides an additional exercise for practicing cross-validation as we have just learned cross-validation.
- Chapter6-HW-2-cross-validation-withAnswers.xlsx presents a reference solution for the cross-validation exercise.

Review Points

1. The mechanism of logistic regression
2. Logistic function, sigmoid function, and odds
3. The assumption that log of odds is a linear function
4. The concept of likelihood and log loss
5. Excel functions IF, INDEX, LINEST, SUMPRODUCT, EXP, and LN
6. The use of Solver