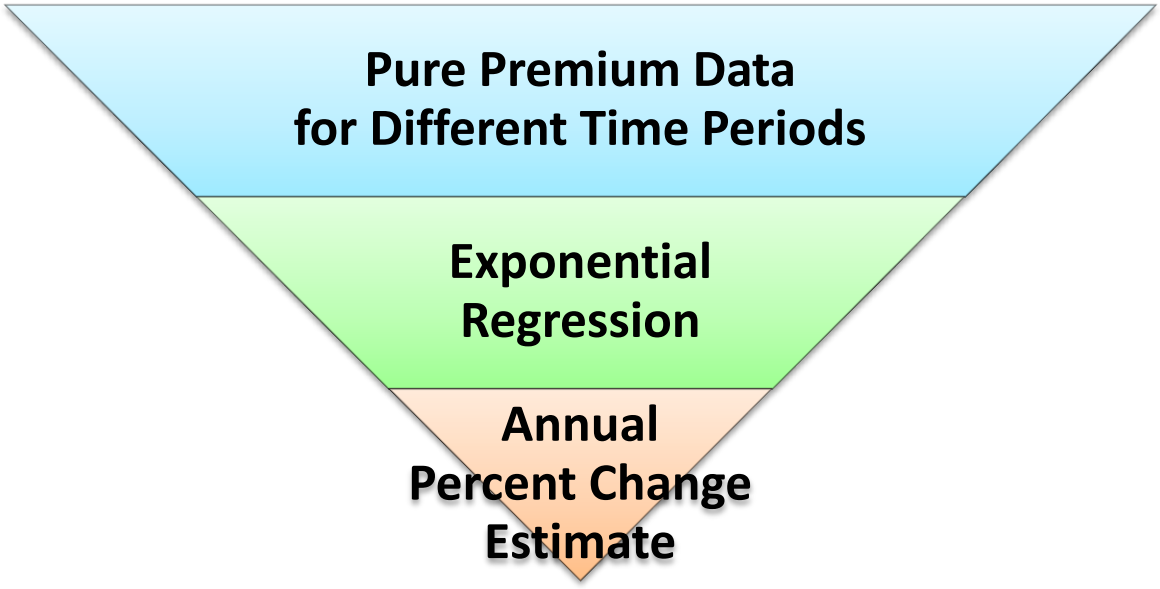
From Confusion to Understanding

By Hernan L. Medina, MS, CPCU

The objectives of this article are to introduce the reader to logistic regression and to provide an example of how logistic regression can be used to help maximize profit in a cross-sell campaign. Most insurance professionals are familiar with linear and exponential regression in the context of estimating trends in claim frequency, claim severity, or pure premium. Generally, the objective in using regression for trend analysis is to estimate an annual change. For example, using linear regression, you could estimate a dollar amount by which you would expect pure premium to change on an annual basis. Similarly, using exponential regression, you could estimate a percentage by which you could expect pure premium to change on an annual basis.

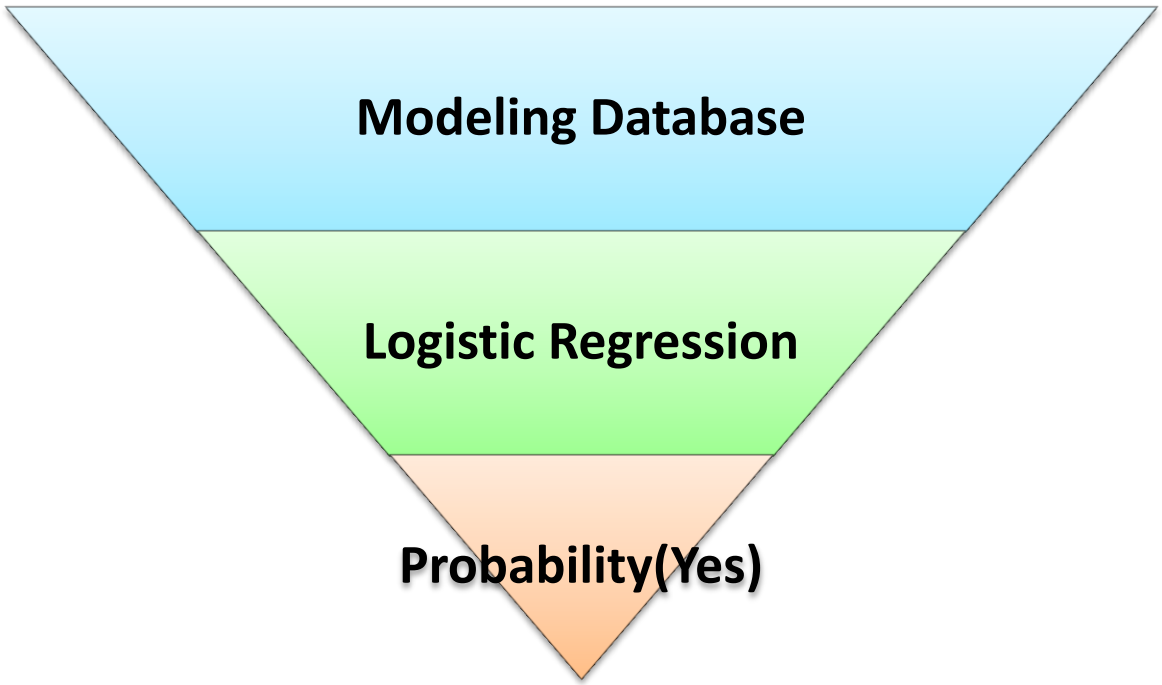


Logistic regression is another variation of the regression model, and it can be used to estimate the probability of an event. First, let us review the concept of probability. If you have a fair coin and you toss it, the result has an equal chance of being heads or tails. Heads is one out of two possible outcomes; thus, the probability of heads is one-half, or 50 percent. Similarly, fair dice have six equally likely results, all of which are equally possible. Rolling a five is one out of six possible outcomes; thus, the probability of rolling a five is one-sixth, or 16.67 percent.

Now suppose you offer to sell a homeowners policy to customers who currently have a personal auto policy with your company. What is the probability that any of these customers will accept your offer? One method for estimating this probability is to use logistic regression.

Let's assume your company has 5 million customers for Policy A, and it costs $1.00 to print a Policy B sales brochure and mail it with the Policy A renewal. Additionally, let's assume the average profit per Policy B customer is $50.00. If you spend $5 million mailing brochures, will you sell enough policies to make a profit?

Logistic regression is often used to assist in decisions such as these, which require a yes or no answer. One approach would be first to do a trial by mailing the sales brochure to a sufficiently large group of policyholders selected at random. A predictive modeler or data scientist can help select an appropriate sample. Once the trial marketing campaign is done, the analytic scientist can build a modeling database with everything you know about the Policy A customers, as well as their response: Did they buy a Policy B (yes or no)? Then the scientist can use logistic regression to see how all available information about those policyholders relates to whether they were interested in buying Policy B from your company.



Once an appropriate logistic regression model has been selected, it is necessary to assess how well it works. A best practice for doing this is to test the model on data that was not used to build it. For example, if you have 5 million policyholders and you sent the trial offer to 10 percent of them, your database would have 500 thousand records. Then you could use 300 thousand records (training sample) to build the model and hold out 200 thousand (test sample) to determine how well the model works.

If there is sufficient data, a three-way split may be used: a training sample to build models, a validation sample to select between competing models, and a holdout sample to test how well the final model works.

The model only produces a probability that a particular Policy A customer will buy a Policy B from your company. It does not tell you whether a specific person will buy. So, if the model tells you there is a 10 percent probability that a person will buy the new product, should you incur the expense of mailing that person an offer? What is the best way to make that decision? Fortunately, a person who knows about predictive modeling can use a holdout test sample to compare the probabilities against actual results for different scenarios and to help determine the optimal decision.

We cannot fit a table with thousands or millions of records into this article, so we will illustrate the process with 40 holdout cases, showing actual response and probability (based on a model built on other data) that the response would have been "Yes."

Case Number	Actual Response	Probability (Yes)
1	No	1%
2	No	1%
3	No	1%
4	No	1%
5	No	1%
6	No	1%
7	No	1%
8	No	1%
9	No	1%
10	No	1%
11	No	1%
12	No	1%
13	No	1%
14	No	1%
15	No	1%
16	No	1%
17	No	1%
18	No	1%
19	No	1%
20	No	1%

Case Number	Actual Response	Probability (Yes)
21	No	10%
22	No	10%
23	No	10%
24	Yes	10%
25	No	10%
26	No	10%
27	No	10%
28	No	10%
29	No	10%
30	No	10%
31	No	20%
32	No	20%
33	Yes	20%
34	No	20%
35	No	20%
36	No	20%
37	No	20%
38	Yes	20%
39	No	20%
40	No	20%

With this information about the probability of a Yes response, we can test three decision scenarios, summarized as follows:

Scenario	Predicted Probability (Yes)	Predicted Response
Scenario 1	Less than 20%	No
	At least 20%	Yes
Scenario 2	Less than 10%	No
	At least 10%	Yes
Scenario 3	Less than 1%	No
	At least 1%	Yes

Case Number	Actual Response	Probability (Yes)	Scenario 1 20% Cutoff	Scenario 2 10% Cutoff	Scenario 3 1% Cutoff
1	No	1%	No	No	Yes
2	No	1%	No	No	Yes
3	No	1%	No	No	Yes
4	No	1%	No	No	Yes
5	No	1%	No	No	Yes
6	No	1%	No	No	Yes
7	No	1%	No	No	Yes
8	No	1%	No	No	Yes
9	No	1%	No	No	Yes
10	No	1%	No	No	Yes
11	No	1%	No	No	Yes
12	No	1%	No	No	Yes
13	No	1%	No	No	Yes
14	No	1%	No	No	Yes
15	No	1%	No	No	Yes
16	No	1%	No	No	Yes
17	No	1%	No	No	Yes
18	No	1%	No	No	Yes
19	No	1%	No	No	Yes
20	No	1%	No	No	Yes
21	No	10%	No	Yes	Yes
22	No	10%	No	Yes	Yes
23	No	10%	No	Yes	Yes
24	Yes	10%	No	Yes	Yes
25	No	10%	No	Yes	Yes
26	No	10%	No	Yes	Yes
27	No	10%	No	Yes	Yes
28	No	10%	No	Yes	Yes
29	No	10%	No	Yes	Yes
30	No	10%	No	Yes	Yes
31	No	20%	Yes	Yes	Yes

Case Number	Actual Response	Probability (Yes)	Scenario 1 20% Cutoff	Scenario 2 10% Cutoff	Scenario 3 1% Cutoff
32	No	20%	Yes	Yes	Yes
33	Yes	20%	Yes	Yes	Yes
34	No	20%	Yes	Yes	Yes
35	No	20%	Yes	Yes	Yes
36	No	20%	Yes	Yes	Yes
37	No	20%	Yes	Yes	Yes
38	Yes	20%	Yes	Yes	Yes
39	No	20%	Yes	Yes	Yes
40	No	20%	Yes	Yes	Yes

Note that the assumptions behind these scenarios lead to four possible types of outcomes:

1. **True Negative**: The actual response is No, and the predicted response is No.
2. **False Positive**: The actual response is No, and the predicted response is Yes.
3. **False Negative**: The actual response is Yes, and the predicted response is No.
4. **True Positive**: The actual response is Yes, and the predicted response is Yes.

Now we can produce a classification table for each scenario by counting the actual response versus the predicted response under each scenario. This classification table is often called a "confusion matrix" because it helps determine the amount of "confusion" in the model, meaning how often it is right versus how often it is wrong. As we shall see, the confusion matrix can also lead you to understand which scenario optimizes expected profit. The four possible outcomes described above relate to the classification table or confusion matrix as follows:

		Predicted		Total
Scenario 1		**No**	**Yes**	
Actual Response	**No**	True Negative	False Positive	Total Actual No
	Yes	False Negative	True Positive	Total Actual Yes
Total		Total Predicted No	Total Predicted Yes	Grand Total

We now proceed to calculate the number of cases in each category for each scenario. First of all, notice that of the 40 actual responses, 3 are Yes, and the other 37 are No. Thus, the Total column for all 3 scenarios should be the same: 37 No, 3 Yes, and 40 Total.

For Scenario 1, there are 30 predicted No responses; 29 of those are true negatives (the actual response is No), and 1 is a false negative (the actual response is Yes). These values are shown in the No column of the classification table. Similarly, there are 10 predicted Yes responses; 8 of those are false positives (the actual responses is No), and 2 are true positives (the actual response is Yes). These values are shown in the Yes column.

		Predicted		Total
Scenario 1		**No**	**Yes**	

Actual Response	**No**	29	8	37
	Yes	1	2	3
Total		30	10	40

For Scenario 2, there are 20 predicted No responses, and these are all true negatives because the actual response is No. Additionally, there are 20 predicted Yes responses; 17 are false positives (the actual response is No), and 3 are true positives (the actual response is Yes).

		Predicted		**Total**
Scenario 2		**No**	**Yes**	
Actual Response	**No**	20	17	37
	Yes	0	3	3
Total		20	20	40

Finally, for Scenario 3, there are 40 predicted Yes responses; 37 are false positives, and 3 are true positives.

		Predicted		**Total**
Scenario 3		**No**	**Yes**	
Actual Response	**No**	0	37	37
	Yes	0	3	3
Total		0	40	40

Continuing with the assumptions stated at the top of this article — a cost of $1.00 to print and mail the sales brochure with the Policy A renewal and an expected profit per Policy B of $50.00 — we can determine the net profit for the four possible types of outcomes.

1. **True Negative**: If both the actual and predicted responses are No, then there is no mailing cost and no Policy B profit. Thus, the expected profit is $0.00.
2. **False Positive**: If the actual response is No and the predicted response is Yes, then there is a $1.00 printing and mailing cost but no Policy B profit, so the expected profit is negative $1.00.
3. **False Negative**: If the actual response is Yes and the predicted response is No, then there is no mailing cost and no insurance profit, so the total expected profit is $0. Additional profit could have been made if this case had been correctly identified, but achieving that profit would require a different decision scenario or a different predictive model.
4. **True Positive**: If the actual response is Yes and the predicted response is Yes, then profit is insurance profit minus mailing cost, or $50.00 – $1.00 = $49.00.

To summarize, when the predicted response is No, the total expected profit is zero, because there are no mailing costs and no Policy B profit. When the predicted response is Yes, you either lose $1.00 (false positive) or gain $49.00 (true positive). With this information we can calculate an expected profit matrix for the three scenarios.

		Predicted		Total
Scenario 1		**No**	**Yes**	
Actual Response	**No**	$0	8 × −$1 = −$8	–$8
	Yes	$0	2 × $49 = $98	$98
Total		$0	$90	$90

		Predicted		Total
Scenario 2		**No**	**Yes**	
Actual Response	**No**	$0	17 × −$1 = −$17	−$17
	Yes	$0	3 × $49 = $147	$157
Total		$0	$130	$130

		Predicted		Total
Scenario 3		**No**	**Yes**	
Actual Response	**No**	$0	37 × −$1 = −$37	−$37
	Yes	$0	3 × $49 = $147	$147
Total		$0	$110	$110

From the three tables above, we can draw the following conclusions:

- Choosing too high a probability threshold (Scenario 1) can leave some potential profit unrealized. The profit for this scenario is about 31 percent less than the profit for Scenario 2.
- Scenario 2 produces the highest profit in this example.
- Mailing an offer to every Policy A customer (Scenario 3) is not the best choice. But this is generally the case in real-world situations. If the percentage of people accepting the offer is low, then at some point the mailing costs outweigh the potential revenue. The profit for this scenario is about 15 percent less than the profit for Scenario 2.

This example illustrates how the "confusion matrix" can lead to understanding which scenario helps maximize profit. In general, predictive modeling applications often involve hundreds or thousands of scenarios. So you need good data and data scientists with the appropriate skills as well as the right software and computing equipment.

References

SAS Institute Inc. 2014. SAS/STAT® 13.2 User's Guide. Cary, NC: SAS Institute Inc.
SAS Institute Inc. 2012. Predictive Modeling Using Logistic Regression Course Notes. Cary, NC: SAS Institute Inc.
De Jong, Piet and Heller, Gillian Z. 2008. Generalized Linear Models for Insurance Data. New York, NY: Cambridge University Press.

About the Author

Hernan L. Medina is director, Analytical Data Management, at ISO Insurance Programs and Analytic Services. He has a master's degree in mathematics from New York University. Additionally, he is a SAS® Certified Base Programmer and SAS® Certified Statistical Business Analyst. Mr. Medina is also a Chartered Property Casualty Underwriter (CPCU), and he holds the AIM, API, AU, and ARC designations.