Бинарная логистическая регрессия (7 шагов в Excel)

Целью этого примера бинарной логистической регрессии является построение уравнения, которое рассчитает вероятность того, что производственная машина в настоящее время будет в состоянии произвести продукцию, соответствующую требуемым спецификациям, в зависимости от возраста машины (в месяцах) и среднего числа рабочих смен, в которые машина работает в течение каждой недели своего жизненного цикла.

Данные были собраны на 20 подобных машинах следующим образом:

1) Производит ли машина продукцию, отвечающую требованиям, по меньшей мере, в 99% случаев (1 = машина соответствует своему назначению (спецификациям) - производит отвечающую требованиям продукцию в течении 99% времени, 0 = машина не соответствует назначению - она не производит отвечающую требованиям продукцию в течении 99% времени);

2) Возраст машины в месяцах;

3) Среднее количество рабочих смен, в течении которых машина эксплуатировалась каждую неделю в течение своего жизненного цикла.

Y	X_1	X_2
Machine Meets Spec?	Machine Age (Months)	Average Number of Shifts/Week
1	57	4
0	73	5
1	22	5
0	59	4
1	15	4
1	36	2
0	68	5
0	49	5
0	27	7
1	59	3
1	10	6
0	78	8
1	22	6
1	36	4
0	57	7
0	73	8
1	38	5
0	71	7
0	35	4
1	44	5

Логистическая регрессия в Excel

Логистическая регрессия Шаг 1 - Сортировка данных

Цель сортировки данных - сделать данные более наглядными. Используя инструмент сортировки данных Excel, выполните первичную сортировку по зависимой переменной. В нашем случае зависимая переменная – Y (соответствие машин своему назначению). Выполните подчиненные сортировки (вторичные, третичные и т.д.), вторичную - в

соответствии с возрастом машин, а третичную - в соответствии со средним числом рабочих смен в неделю. Результаты приведены ниже:

Y	X_1	X_2
Machine Meets Spec?	Machine Age (Months)	Average Number of Shifts/Week
0	78	8
0	73	8
0	73	5
0	71	7
0	68	5
0	59	4
0	57	7
0	49	5
0	35	4
0	27	7
1	59	3
1	57	4
1	44	5
1	38	5
1	36	4
1	36	2
1	22	6
1	22	5
1	15	4
1	10	6

Из сортировки данных становится очевидно: машины, которые не выпускали продукцию, соответствующую требованиям, либо относятся к более старым машинам, либо отработали больше рабочих смен в неделю.

Логистическая регрессия Шаг 2 - Вычислить логит для каждой строки данных

Учитывая следующие независимые переменные (регрессоры), X_1, X_2, ..., X_k, логит равен следующему:

$$L = b0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

Если объясняющими переменными являются Age и Average Number of Shifts, логит L, выглядит следующим образом:

$$Logit = b_0 + b_1 * Age + b_2 * Average\ number\ of\ shifts$$

Excel Solver (поиск решения) оптимизирует коэффициенты регрессии b_0, b_1 и b_2, чтобы построить уравнение, которое будет точно прогнозировать вероятность производства машиной требуемой продукции с учетом возраста машины и среднего числа рабочих смен в неделю.

Коэффициенты b_0, b_1 и b_2 произвольно устанавливаются на 0,1 до запуска Solver. Рекомендуется сначала установить коэффициенты, чтобы полученный логит был значительно ниже 20 для каждой строки записи. Логиты, которые превышают 20, вызывают экстремальные значения на последующих этапах логистической регрессии. Коэффициенты b_0, b_1 и b_2 были произвольно установлены в значение от 0,1 до первоначального и производят достаточно небольшие логиты, как показано ниже.

Индивидуальный логит строится для каждой из 20 строк данных на основе начальных настроек переменных решения (коэффициентов) следующим образом:

	A	B	C	D	E	F	G
1			Solver Decision Variables				
2		b_0 =	0.1				
3		b_1 =	0.1				
4		b_2 =	0.1				
5							
6					L = Logit		
7	Y	X_1	X_2		$L = b_0 + b_1*X_1 + b_2*X_2$		
8	Machine Meets Spec?	Machine Age (Months)	Average Number of Shifts/Week		L		
9	0	78	8		8.7	=$C$2+$C$3*B9+$C$4*C9	
10	0	73	8		8.2		
11	0	73	5		7.9		
12	0	71	7		7.9		
13	0	68	5		7.4		
14	0	59	4		6.4		
15	0	57	7		6.5		
16	0	49	5		5.5		
17	0	35	4		4		
18	0	27	7		3.5		
19	1	59	3		6.3		
20	1	57	4		6.2		
21	1	44	5		5		
22	1	38	5		4.4		
23	1	36	4		4.1		
24	1	36	2		3.9		
25	1	22	6		2.9		
26	1	22	5		2.8		
27	1	15	4		2		
28	1	10	6		1.7		
29							

Логистическая регрессия Шаг 3 - Расчет e^L для каждой записи данных

Число e является базой натурального логарифма. Оно приблизительно равно 2.71828163 и является пределом $(1 + 1/n)^n$, при $n \to \infty$. e^L должно быть рассчитано для каждой записи данных. Этот шаг будет показан на рисунке ниже.

Логистическая регрессия Шаг 4 - Рассчитайте P (X) для каждой записи данных

P(X) - вероятность события X. Событие X возникает, когда машина производит продукцию, соответствующую требованиям. P(X) - это вероятность того, что машина будет производить продукцию, соответствующую требованиям.

$P(X) = e^L / (1 + e^L)$

$L = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_k * X_k$

Вычисление e^L и P(X) для каждой из строк выполняется следующим образом:

	D	E	F	G	H
5					
6		e =	2.71828183		
7		C9 =	=$F$6^E9		
8		L	e^L	$P(X) = e^L/(1 + e^L)$	
9		8.7	6002.912247	0.999833442	=F9/(1+F9)
10		8.2	3640.950324	0.999725422	
11		7.9	2697.28234	0.999629394	
12		7.9	2697.28234	0.999629394	
13		7.4	1635.984437	0.999389121	
14		6.4	601.8450401	0.998341199	
15		6.5	665.1416355	0.998498818	
16		5.5	244.691933	0.995929862	
17		4	54.59815016	0.98201379	
18		3.5	33.11545202	0.970687769	
19		6.3	544.5719121	0.998167061	
20		6.2	492.7490428	0.99797468	
21		5	148.4131595	0.993307149	
22		4.4	81.45086887	0.987871565	
23		4.1	60.34028774	0.983697501	
24		3.9	49.40244921	0.980159694	
25		2.9	18.1741454	0.947846437	
26		2.8	16.4446468	0.942675824	
27		2	7.389056107	0.880797078	
28		1.7	5.473947397	0.845534735	

e^L также можно вычислить в Excel как exp (L).

Логистическая регрессия Шаг 5 - Рассчитать LL, функцию логарифмического правдоподобия

Условная вероятность $Pr(Y_i = y_i \mid X1_i, X2_i, ... Xk_i)$ - вероятность того, что предсказанная зависимая переменная y_i равна фактическому наблюдаемому значению Y_i с учетом значений независимых переменных $X1_i, X2_i, ... Xk_i$.

Условная вероятность $Pr(Y_i = y_i \mid X1_i, X2_i, ... Xk_i)$ будем далее для удобства обозначать сокращенно $Pr(Y = y \mid X)$.

Условная вероятность $Pr(Y = y \mid X)$ вычисляется по следующей формуле:

$$Pr (Y = y \mid X) = P (X)^Y * (1-P (X))^{(1-Y)}$$

Если взять натуральный логарифм с обеих сторон уравнения, получим следующее:

$$\ln (Pr (Y = y \mid X)) = y * \ln (P (X)) * (1-y) * \ln (1-P (X))$$

Логарифмическая функция правдоподобия, LL, является суммой членов ln (Pr (Y = y | X)) для всех строк данных по следующей формуле:

$$LL = \Sigma Y_i * \ln(P(X_i)) + (1 - Y_i) \ln((1-P(X_i)))$$

Вычисление LL выполняется следующим образом:

	A	B	C
6			
7	Y	X_1	X_2
8	Machine Meets Spec?	Machine Age (Months)	Average Number of Shifts/Week
9	0	78	8
10	0	73	8
11	0	73	5
12	0	71	7
13	0	68	5
14	0	59	4
15	0	57	7
16	0	49	5
17	0	35	4
18	0	27	7
19	1	59	3
20	1	57	4
21	1	44	5
22	1	38	5
23	1	36	4
24	1	36	2
25	1	22	6
26	1	22	5
27	1	15	4
28	1	10	6

	D	E	F	G	H	I	J	K	L
6		e =	2.71828183		$Pr[P(X_i)=Y_i \mid X_1,..,X_k] = p^Y(1-p)^{(1-Y)}$				
7									
8		L	e^L	P(X)	$\ln(Pr[P(X_i)=Y_i \mid X_1,..,X_k])$	= y*ln(p) + (1-y)*ln(1-p)			
9		8.7	6002.912247	0.999833442	-8.700166577	=A9*LN(G9)+(1-A9)*LN(1-G9)			
10		8.2	3640.950324	0.999725422	-8.200274621	=A10*LN(G10)+(1-A10)*LN(1-G10)			
11		7.9	2697.28234	0.999629394	-7.900370679				
12		7.9	2697.28234	0.999629394	-7.900370679				
13		7.4	1635.984437	0.999389121	-7.40061107				
14		6.4	601.8450401	0.998341199	-6.401660182				
15		6.5	665.1416355	0.998498818	-6.501502314				
16		5.5	244.691933	0.995929862	-5.504078446				
17		4	54.59815016	0.98201379	-4.01814993				
18		3.5	33.11545202	0.970687769	-3.52975042				
19		6.3	544.5719121	0.998167061	-0.001834621				
20		6.2	492.7490428	0.99797468	-0.002027374				
21		5	148.4131595	0.993307149	-0.006715348				
22		4.4	81.45086887	0.987871565	-0.012202585				
23		4.1	60.34028774	0.983697501	-0.016436847				
24		3.9	49.40244921	0.980159694	-0.020039767				
25		2.9	18.1741454	0.947846437	-0.053562776				
26		2.8	16.4446468	0.942675824	-0.059032826				
27		2	7.389056107	0.880797078	-0.126928011	=A27*LN(G27)+(1-A27)*LN(1-G27)			
28		1.7	5.473947397	0.845534735	-0.167786029	=A28*LN(G28)+(1-A28)*LN(1-G28)			
29									
30			LL = Log-Likelihood =		-66.5235011	=SUM(H9:H28)			

Логистическая регрессия Шаг 6 - Используйте Excel Solver (поиск решения0 для вычисления MLL, функции максимального логарифмического правдоподобия

Цель логистической регрессии - найти коэффициенты логита (b_0, b_1, b_2, ..., b_k), которые максимизируют LL, функцию логарифмического правдоподобия в ячейке H30, для создания MLL, функции максимального логарифмического правдоподобия.

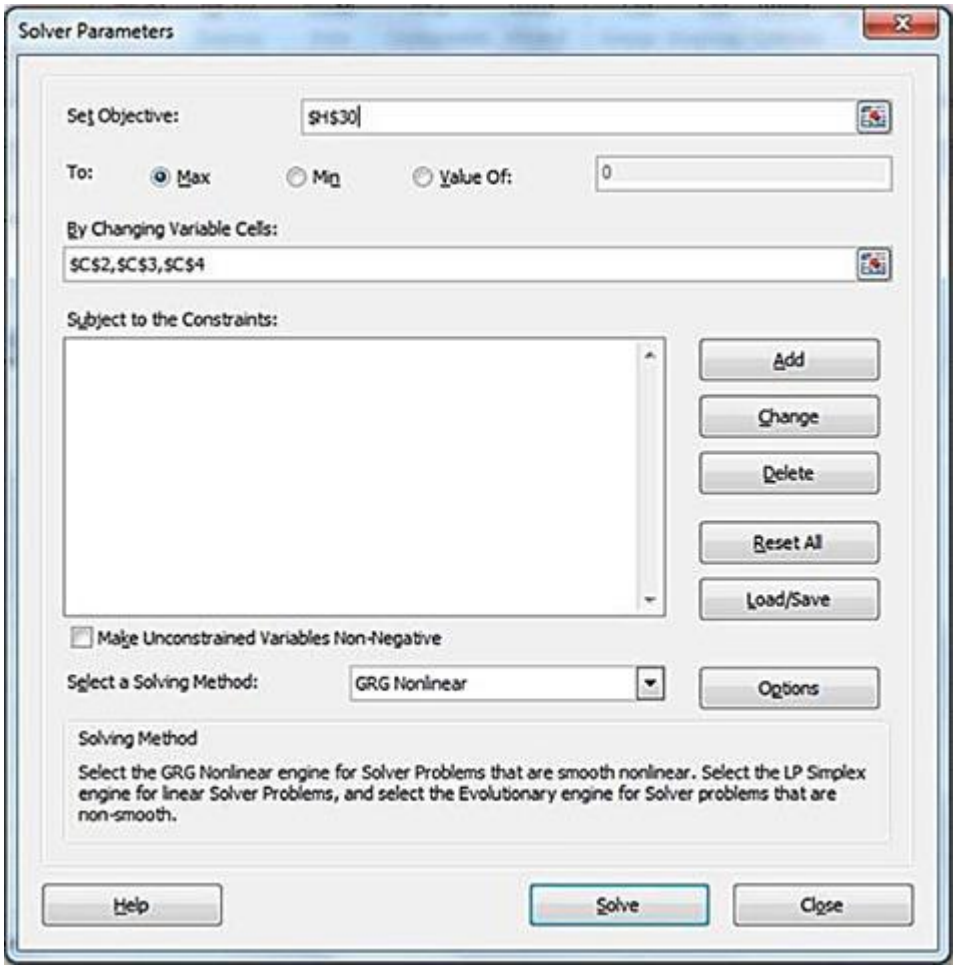
Функциональность Excel Solver довольно проста: Excel Solver настраивает числовые значения в определенных ячейках, чтобы максимизировать или минимизировать значение в другой отдельной ячейке.

Ячейка, которую Solver пытается максимизировать или минимизировать, называется решением задачи (Solver Objective). Это LL в ячейке H30.

Ячейки, значения которых корректируются Solver, называются переменными решения. Переменные решения Solver содержатся в ячейках C2, C3 и C4. Они содержат b_0, b_1, b_2, ..., bk, коэффициенты для логита. Эти ячейки будут скорректированы, чтобы максимизировать LL, который находится в ячейке H30.

Excel Solver (поиск решения) - это надстройка, которая включена в большинство пакетов Excel. Решатель чаще всего активируется пользователем вручную, прежде чем его можно будет использовать в первый раз. Для разных версий Excel требуется разный способ активации для Solver. После активации Solver обычно находится на вкладке «Данные» версий Excel с 2007 года, которые используют навигацию по ленте. Excel 2003 предоставляет ссылку на Solver в раскрывающемся меню в разделе «Инструменты».

Эти переменные (Decision variables) и задача (Objective) вводятся в диалоговое окно Solver следующим образом:



Убедитесь, что вы не установили флажок “Make Unconstrained Variables Non-Negative”.

Метод нелинейного решения GRG Excel Solver

Метод нелинейного решения GRG следует выбирать, если любое из уравнений с переменными решения или ограничениями является нелинейным и гладким (непрерывным, непрерывным, т.е. без разрывов). GRG означает обобщенный сокращенный градиент и является старым, проверенным, надежным методом решения нелинейных задач.

Уравнения для вычисления Objective (максимизация LK) включают вычисление e^L, P (X) и Pr(Y = y | X). Каждое из этих трех уравнений является нелинейным и гладким. Уравнение является «гладким», если это уравнение и производная этого уравнения не имеют разрывов (непрерывны). Поэтому следует выбрать метод нелинейного решения GRG.

Один из способов определить, является ли уравнение или функция негладкой (график имеет точку, указывающую, что производная является разрывной) или прерывистой (график уравнения резко меняет значения в определенных точках) - построить график уравнения в его ожидаемом диапазоне значений.

Solver должен быть запущен несколько раз для обеспечения оптимального решения.

Когда Solver запускает алгоритм GRG, он выбирает отправную точку для своих вычислений. Каждый раз, когда вновь выполняется алгоритм Solver GRG, он выбирает несколько иную стартовую точку. Вот почему разные ответы часто появляются после каждого запуска метода нелинейного решения GRG. Solver должен быть запущен несколько раз, пока Objective (LK) не будет максимизирован. Это должно обеспечить наилучшие локально-оптимальные значения переменных принятия решений (b_0, b_1, b_2, ..., b_k).

Метод нелинейного решения GRG гарантирует получение локально-оптимальных решений, но не глобально-оптимальных решений. Метод нелинейного решения GRG даст глобально оптимальное решение, если все функции на пути к Objective и всем ограничениям являются выпуклыми. Если какая-либо из функций или ограничений не является выпуклой, метод нелинейного решения GRG может найти только локально-оптимальные решения.

Функция является выпуклой, если она имеет только один пик вверх или вниз. Выпуклая функция всегда может быть решена в глобально-оптимальном решении. Функция невыпукла, если она имеет более одного пика или имеет разрывы. Невыпуклые решения часто можно найти только для локально-оптимальных решений.

Функция e^L с $L = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_k * X_k$ может быть невыпукла, поскольку регрессоры X_1, X_2, ..., X_k могут быть нелинейными. Поэтому метод нелинейного решения GRG гарантированно найдет локально-оптимальное решение.

Как увеличить вероятность, что Solver найдет глобально-оптимальное решение ?

Существует три способа:

Во-первых, многократно запускать Solver с использованием разных наборов значений для переменных. Эта опция позволяет вам выбирать начальные наборы переменных на основе понимания общей проблемы и часто является наилучшим способом достижения наиболее желательного решения.

Во-вторых, нужно выбрать «Use Multistart». Это запускает GRG Solver несколько раз и случайным образом выбирает другой набор начальных значений для переменных в ходе каждого прогона. Затем Solver представляет лучшие из найденных локально-оптимальных решений.

Третий способ - установить ограничения в диалоговом окне Solver, которые заставят Solver попробовать новый набор значений. Ограничения - это ограничения, помещенные вручную в Decision Variables. Ограничения могут быть полезны, если переменные должны быть ограничены определенным диапазоном значений. Оптимальное решение на глобальном уровне вряд ли можно найти, применив ограничения, но более реалистичное решение может быть получено путем ограничения переменных до вероятных значений.

Интерпретация результатов решения Excel

Запуск решения приводит к следующим результатам:

	B	C
1		**Solver Decision Variables**
2	b_0 =	12.48285608
3	b_1 =	-0.117031374
4	b_2 =	-1.469140055

	D	E	F	G	H	I	J	K	L
6		e =	2.71828183		$Pr[P(X_i) = Y_i \mid X_1,..,X_k] = p^Y(1-p)^{(1-Y)}$				
7									
8		L	e^L	P(X)	$\ln(Pr[P(X_i) = Y_i \mid X_1,..,X_k])$	= y*ln(p) + (1-y)*ln(1-p)			
9		-8.39871	0.000225157	0.000225107	-0.000225132	=A9*LN(G9)+(1-A9)*LN(1-G9)			
10		-7.81355	0.000404219	0.000404055	-0.000404137	=A10*LN(G10)+(1-A10)*LN(1-G10)			
11		-3.40613	0.033169168	0.032104295	-0.03263094				
12		-6.11035	0.00221977	0.002214853	-0.00221731				
13		-2.82098	0.059547698	0.056201055	-0.057842117				
14		-0.29856	0.741889316	0.42591071	-0.554970338				
15		-4.47191	0.011425442	0.011296376	-0.011360665				
16		-0.59738	0.550250566	0.354942987	-0.438416573				
17		2.510198	12.30736379	0.924853636	-2.58831755				
18		-0.96097	0.382521121	0.276683745	-0.323908731				
19		1.170585	3.223877539	0.763250712	-0.270168715				
20		-0.06449	0.937543176	0.483882469	-0.725913234				
21		-0.01222	0.987849758	0.496943873	-0.69927819				
22		0.689964	1.993642936	0.665958826	-0.406527433				
23		2.393166	10.9481051	0.91630472	-0.087406306				
24		5.331447	206.7368035	0.995186216	-0.004825407				
25		1.093326	2.984181538	0.749007421	-0.289006388				
26		2.562466	12.96775088	0.928406512	-0.07428559				
27		4.850825	127.8458511	0.992238788	-0.007791487	=A27*LN(G27)+(1-A27)*LN(1-G27)			
28		2.497702	12.15453083	0.923980565	-0.079064241	=A28*LN(G28)+(1-A28)*LN(1-G28)			
29		Solver Objective Is To Maximize							
30		MLL_m = Max Log-Likelihood		Full Model =	-6.654560484	=SUM(H9:H28)			

MLL, максимальное логарифмическое правдоподобие равно -.654560484, при коэффициентах, скорректированных как Decision Variables для Solver:

b_0 = 12,48285608

b_1 = -0.117031374

b_2 = -1,469140055

Логистическая регрессия Шаг 7 - Проверка результата решения путем запуска сценариев

Подтвердите вывод, выполнив несколько сценариев с помощью результатов Solver. В каждом сценарии будет использоваться другое изменение входных переменных X_1, X_2, .., X_k для создания выходов, которые должны соответствовать исходному набору данных.

Вид начальных данных показал, продукция, несоответствующая требованиям, с большей вероятностью производится на старых машинах и/или на машинах, которые работали чаще.

Следующие три сценария были выполнены следующим образом:

Сценарий 1

Возраст машины = 40 месяцев

Среднее число рабочих смен = 7

P (X) = Вероятность продукции, соответствующей требованиям = 8%

			Solver Decision Variables
X_1 = Machine Age (Months) =	40	b_0 =	12.4828561
X_2 = Average # Weekly Shifts =	7	b_1 =	-0.1170314
$L = b_0 + b_1 * X_1 + b_2 * X_2$	-2.482379273	b_2 =	-1.4691401
e^L =	0.0835		
$P(x) = e^L / (1+e^L)$ =	8%		

Сценарий 2
Возраст машины = 40 месяцев
Среднее число рабочих смен = 4
P (X) = Вероятность продукции, соответствующей требованиям = 87%

			Solver Decision Variables
X_1 = Machine Age (Months) =	40	b_0 =	12.4828561
X_2 = Average # Weekly Shifts =	4	b_1 =	-0.1170314
$L = b_0 + b_1 * X_1 + b_2 * X_2$	1.925040893	b_2 =	-1.4691401
e^L =	6.8554		
$P(x) = e^L / (1+e^L)$ =	87%		

Сценарий 3
Возраст машины = 12 месяцев
Среднее число рабочих смен = 7
P (X) = Вероятность продукции, соответствующей требованиям = 69%

			Solver Decision Variables
X_1 = Machine Age (Months) =	12	b_0 =	12.4828561
X_2 = Average # Weekly Shifts =	7	b_1 =	-0.1170314
$L = b_0 + b_1 * X_1 + b_2 * X_2$	0.794499203	b_2 =	-1.4691401
e^L =	2.2133		
$P(x) = e^L / (1+e^L)$ =	69%		

Результаты этих трех сценариев согласуются с данными, очевидными в первоначальном отсортированном наборе данных, что несоответствующая продукция, скорее всего, будет производиться более старыми машинами и / или машинами, которые работали чаще:

Y	X_1	X_2
Machine Meets Spec?	Machine Age (Months)	Average Number of Shifts/Week
0	78	8
0	73	8
0	73	5
0	71	7
0	68	5
0	59	4
0	57	7
0	49	5
0	35	4
0	27	7
1	59	3
1	57	4
1	44	5
1	38	5
1	36	4
1	36	2
1	22	6
1	22	5
1	15	4
1	10	6

Задание

Проанализировать влияние событий от двух факторов, построить функцию, проанализировать сценарии

Возможные варианты, можно придумать свое

1)С1900 по сегодняшний день США уровень доверие президента в процентах(и его возраст) и развязывание войны
2) изменение Цены на нефть / солнечная активность/ колебание температуры
3) ВВП растет/ падает от количества санкций и продукции на экспорт
4) влияние на продолжительность жизни выхода на пенсию и минимального потребительского бюджета