

Logistic regression

Venugopal Gopalakrishna-Remani (Venu), Ph.D.,
BVSc & A.H, MBA, PGDPM, ACUE

The University of Texas at

T**TYLER**TM

Your Success. Our Passion.

What is logistic regression?

Estimate (guess) the probability of an event given some previous data.

Works with binary data, event happens (1) or the event does not happen (0).

Outcome & independent Variables

- Two possible outcomes, "0" and "1" ("dead" vs. "alive" or "win" vs. "loss")

Prediction is based on what?

- Is the independent variable
- Predict a student pass or fail in an exam based on the number of hours spent studying.
- Number of hours studied become independent variable
- We can also consider his/her IQ and that becomes another dependent variable and so on

When should you use logistic regression?

- To predict the likelihood of an event to occur
- To understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.

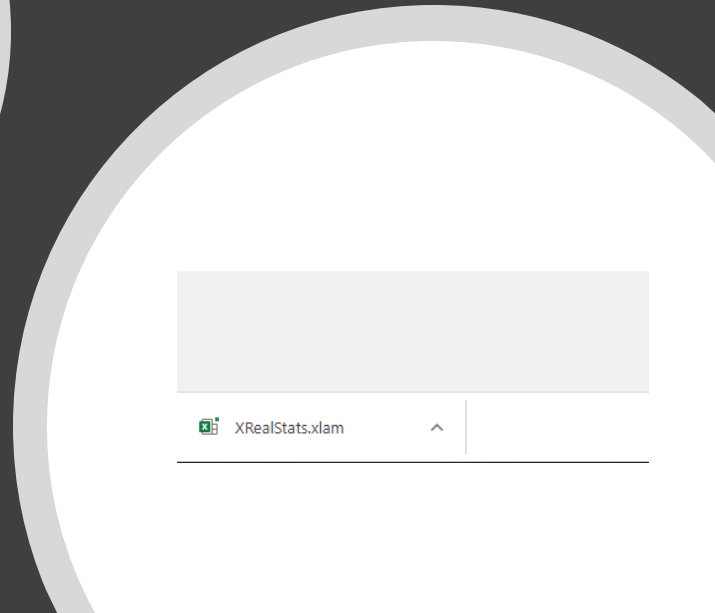
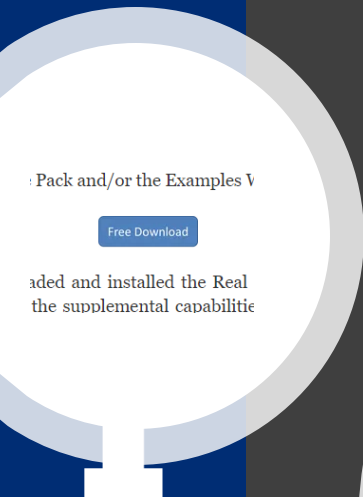
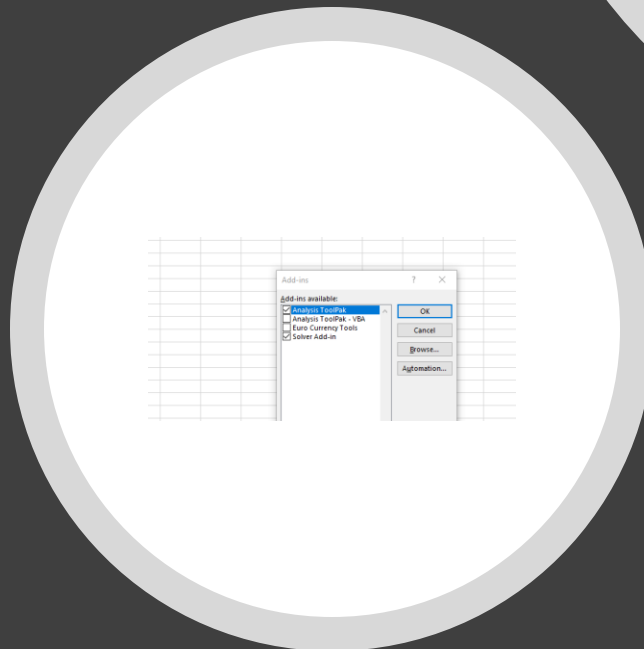
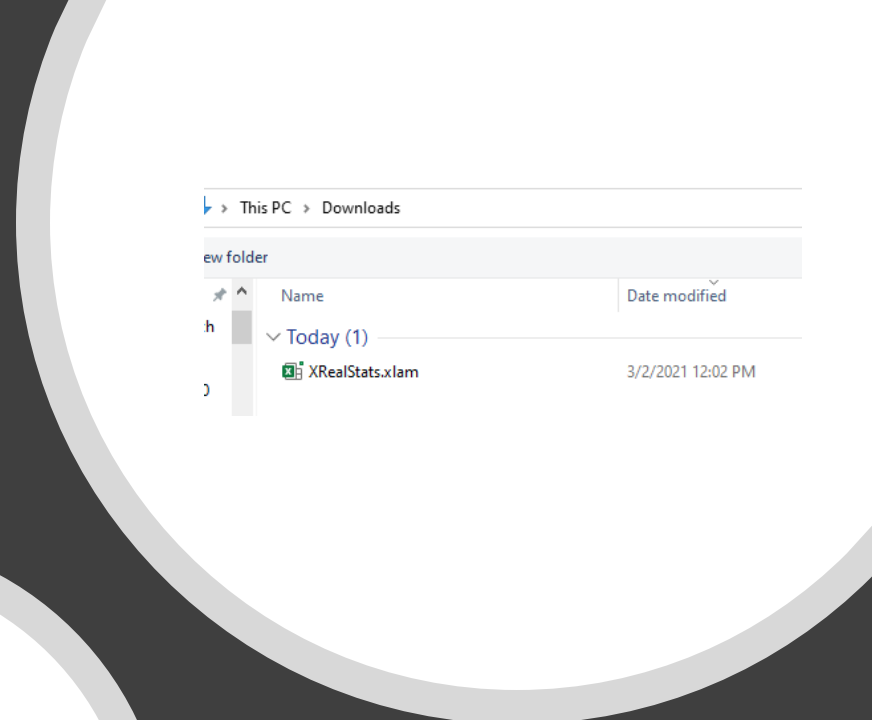


Logistic regression

- With 14 points, 4 rebounds and 5 assists, will Dr.V will make it to the awesome team?
- WE WILL SOLVE THIS AT THIS ONLINE WORKSHOP:
- Pic Credit:
<https://www.shutterstock.com/image-photo/back-view-basketball-player-holding-against->

PROCESS 1: INSTALLING SOFTWARE

- Go to Real Statistics .com (<https://www.real-statistics.com/>)
- Click on Free Download
- Download Real Statistics resource pack
- Click on the download and install it
- Go to Excel Home > Options > Add ins >Browse
- Browse >downloads> XrealStats.Xlam
- Once added in> addins



Pack and/or the Examples V

Free Download

aded and installed the Real
the supplemental capabilitie

Data set

The screenshot shows the Kaggle website interface. The browser address bar displays 'kaggle.com/c/titanic/data'. The top navigation bar includes a search bar, 'Sign In', and 'Register' buttons. A left sidebar contains navigation links: Home, Compete, Data, Code, Communities, Courses, and More. The main content area features a banner for the 'Titanic - Machine Learning from Disaster' competition, with the text 'Start here! Predict survival on the Titanic and get familiar with ML basics'. Below the banner, there are tabs for 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard', and 'Rules', with 'Data' selected. A 'Join Competition' button is visible. The 'Data Description' section is partially visible, starting with an 'Overview' heading and the text 'The data has been split into two groups:'. A bulleted list follows: 'training set (train.csv)' and 'test set (test.csv)'. Below the list, there is a paragraph explaining that the training set is used to build machine learning models and that the outcome is the 'ground truth' for each passenger, with features like gender and class provided. A link to 'feature engineering' is also present.

kaggle.com/c/titanic/data

Apps Maps Gmail YouTube Five D Cattle Comp... Trait anger and cyb... Explaining the impa... (4) Average Varianc... Characteristics of 3... 5 മിനുട്ട് കൊണ്ട്... WileyPLUS Login Pa...

Search Sign In Register

Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 20,890 teams · Ongoing

Overview **Data** Code Discussion Leaderboard Rules [Join Competition](#)

Data Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use [feature engineering](#) to create new features.



Sample Size

- **Sample Size:**
- Equation is $10k/q$ where k = the number of independent variables and q = the smaller of the percentage of cases with $y = 0$ or $y = 1$, with a minimum of 100.
- For Example 1, $k = 2$ and $q = 200/500 = .40$, and so $10k/q = 50$. A minimum sample of size 100 is recommended.



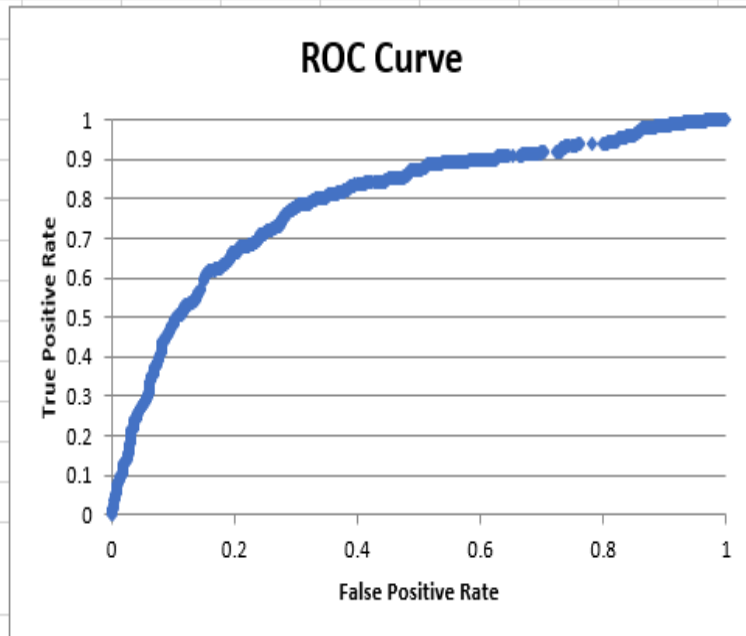
Process

- Excel
- > Add-ins
- >Real Statistics> Data Analytic tools
- >Reg>Logit and Probit regression
- >Select Input Range to Fill
- > Select Output Range –New
- >OK

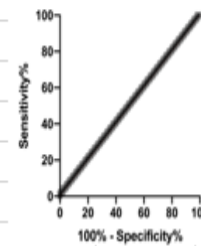


R-Sq (N)	0.296389
Hosmer	1138.052
df	1070
p-value	0.07279
alpha	0.05
sig	no

The Hosmer test of the goodness of fit suggests the model is a good fit to the data as $p=0.072$ ($<.05$) is not significant. However the chi-squared statistic on which it is based is very dependent on sample size and it suggests that the model is explaining more of the variance in the outcome and the chi-square is highly significant (chi-square=296.0685, df=7, $p<.000$)

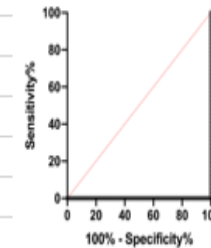


ROC curve with no predictive power:



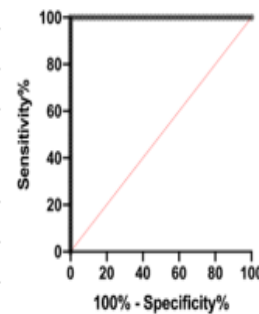
The fit model predicts outcome no better than flipping a coin

Worst-case ROC curve:



"successes" (or 1s) were more commonly predicted to be "failures" (or 0s) than what would be expected by random chance

Best-case ROC curve



If you have this curve, then you probably don't need statistics, since it is trivial to discriminate between the 0s and 1s



24 : X ✓ fx =3.68118 + 0.11283*(14) -(0.395684*(4) + 0.679539*(5))

Logistic Regression Classificat

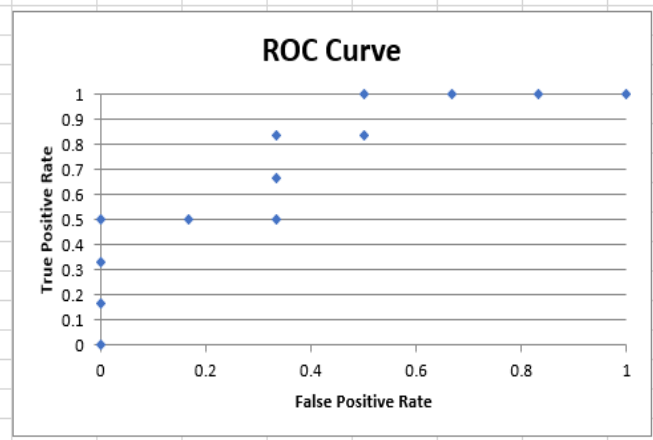
pts	reb	assists	Success	Failure	Total	p-Obs	p-Pred	Suc-Pred	Fail-Pred	LL	% Correct	HL Stat
12	3	6	0	1	1	0	0.557043	0.557043	0.442957	-0.81428	0	1.257558
12	9	9	1	0	1	1	0.990453	0.990453	0.009547	-0.00959	100	0.009639
13	4	4	1	0	1	1	0.300052	0.300052	0.699948	-1.2038	0	2.332758
13	4	6	0	1	1	0	0.625281	0.625281	0.374719	-0.98158	0	1.66867
14	4	4	0	1	1	0	0.276902	0.276902	0.723098	-0.32421	100	0.382939
14	4	5	1	0	1	1	0.43037	0.43037	0.56963	-0.84311	0	1.32358
17	2	2	0	1	1	0	0.030804	0.030804	0.969196	-0.03129	100	0.031783
17	6	5	1	0	1	1	0.543029	0.543029	0.456971	-0.61059	100	0.841523
21	5	7	1	0	1	1	0.66477	0.66477	0.33523	-0.40831	100	0.50428
21	9	3	0	1	1	0	0.389171	0.389171	0.610829	-0.49294	100	0.637119
24	4	5	0	1	1	0	0.196451	0.196451	0.803549	-0.21872	100	0.24448
24	11	11	1	0	1	1	0.995672	0.995672	0.004328	-0.00434	100	0.004346
			6	6	12			6	6	-5.94277	66.66667	9.238676

Coeff	LL0	-8.31777
	LL1	-5.94277
	Chi-Sq	4.75
	df	3
	p-value	0.191046
	alpha	0.05
	sig	no
	R-Sq (L)	0.285533
	R-Sq (CS)	0.326881
	R-Sq (N)	0.435841
	Hosmer	9.238676
	df	10
	p-value	0.509612
	alpha	0.05
	sig	no

Covariance Matrix				Converge
20.24499	-0.36894	-0.77876	-1.99424	-6.4E-17
-0.36894	0.046172	-0.06067	-0.01828	1.78E-16
-0.77876	-0.06067	0.252716	0.109848	-3E-16
-1.99424	-0.01828	0.109848	0.352062	-2.7E-16

	coeff b	s.e.	Wald	p-value	exp(b)	lower	upper
Intercept	-3.68118	4.499443	0.669353	0.413277	0.025193		
pts	-0.11283	0.214878	0.27571	0.599527	0.893304	0.586267	1.361142
reb	0.395684	0.502709	0.619531	0.431221	1.485399	0.554545	3.978776
assists	0.679539	0.593349	1.311623	0.252101	1.972968	0.616681	6.312178

0.280369
0.569637

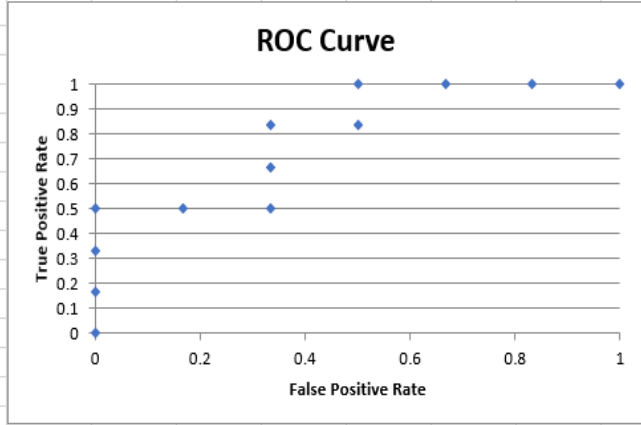


O25 =EXP(O24)/(1+EXP(O24))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA								
1	Logistic Regression																																		Classifica
2																																			
3	<i>pts</i>	<i>reb</i>	<i>assists</i>	<i>Success</i>	<i>Failure</i>	<i>Total</i>	<i>p-Obs</i>	<i>p-Pred</i>	<i>Suc-Pred</i>	<i>Fail-Pred</i>	<i>LL</i>	<i>% Correct</i>	<i>HL Stat</i>		Coeff	LL0	-8.31777			Covariance Matrix						Converge									
4	12	3	6	0	1	1	0	0.557043	0.557043	0.442957	-0.81428	0	1.257558			LL1	-5.94277			20.24499	-0.36894	-0.77876	-1.99424		-6.4E-17							Suc-Pred			
5	12	9	9	1	0	1	1	0.990453	0.990453	0.009547	-0.00959	100	0.009639		-3.68118					-0.36894	0.046172	-0.06067	-0.01828		1.78E-16							Fail-Pred			
6	13	4	4	1	0	1	1	0.300052	0.300052	0.699948	-1.2038	0	2.332758		-0.11283		Chi-Sq	4.75							-3E-16										
7	13	4	6	0	1	1	0	0.625281	0.625281	0.374719	-0.98158	0	1.66867		0.395684		df	3							-2.7E-16										
8	14	4	4	0	1	1	0	0.276902	0.276902	0.723098	-0.32421	100	0.382939		0.679539		p-value	0.191046															Accuracy		
9	14	4	5	1	0	1	1	0.43037	0.43037	0.56963	-0.84311	0	1.32358				alpha	0.05																	
10	17	2	2	0	1	1	0	0.030804	0.030804	0.969196	-0.03129	100	0.031783				sig	no																Cutoff	
11	17	6	5	1	0	1	1	0.543029	0.543029	0.456971	-0.61059	100	0.841523																						
12	21	5	7	1	0	1	1	0.66477	0.66477	0.33523	-0.40831	100	0.50428				R-Sq (L)	0.285533																	
13	21	9	3	0	1	1	0	0.389171	0.389171	0.610829	-0.49294	100	0.637119				R-Sq (CS)	0.326881																	
14	24	4	5	0	1	1	0	0.196451	0.196451	0.803549	-0.21872	100	0.24448				R-Sq (N)	0.435841																	
15	24	11	11	1	0	1	1	0.995672	0.995672	0.004328	-0.00434	100	0.004346																						
16				6	6	12												Hosmer	9.238676																
17																		df	10																

	<i>coeff b</i>	<i>s.e.</i>	<i>Wald</i>	<i>p-value</i>	<i>exp(b)</i>	<i>lower</i>	<i>upper</i>
19	Intercept	-3.68118	4.499443	0.669353	0.413277	0.025193	
20	pts	-0.11283	0.214878	0.27571	0.599527	0.893304	0.586267 1.361142
21	reb	0.395684	0.502709	0.619531	0.431221	1.485399	0.554545 3.978776
22	assists	0.679539	0.593349	1.311623	0.252101	1.972968	0.616681 6.312178

0.280369
0.569637





References


- You can download the realstats from <https://www.real-statistics.com/free-download/real-statistics-resource-pack/>
 - Basketball data from <https://www.statology.org/logistic-regression-excel/>
- 

Photo credits

- Pic Credit:
<https://www.shutterstock.com/image-photo/back-view-basketball-player-holding-against->