Rotman

Master of Management Analytics

INTRO TO JMP – PART 3

Bootcamp

August 8, 2023 Prepared by Jay / <u>TDMDAL</u>



Plan

- Session 1
 - Workflow overview
 - Basic data manipulation
- Session 2
 - Join data tables
 - JMP graphing
- Session 3
 - Modelling
 - JMP Journal
 - JMP Scripting Language

Modeling in JMP

- Linear regression (done)
 - predict a continuous variable
- Logistic regression: predict categorical variable
 - supervised classification learning
 - e.g., binomial logistic regression: the categorical variable has binary outcomes (e.g., 0, 1)
- K-mean clustering: a method to partition observations (into clusters)
 - unsupervised classification learning
- Model selection
 - Validation Column and Model Comparison
- Outliers, missing values, and patterns

Note: the purpose of modelling isn't just prediction.

Binomial Logistic Regression

- let Y be the binary outcome variable
 - e.g. $\{0, 1\} = \{fail, success\}$

• Let
$$p = prob(Y = 1)$$
; $\frac{p}{1-p}$ is then the odds of being 1 (or success)

- Binary logistic regression models the logit-transformed probability as a linear relationship with the predictor variables
 - maximum likelihood estimation

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Binomial Logistic Regression (Demo): Data

- Preference for a brand of detergent (Ries and Smith, 1963)
 - Help > Sample Data Library
 - Detergent.jmp
- Survey Questions
 - 1. which brand do you prefer, x or m
 - 2. water softness
 - 3. previous user of m
 - 4. water temperature

强 Detergent - JMP Pro							_	×
File Edit Tables Rows Cols				View Window	w Help			
Detergent Locked File C:\Program Files\SAS		brand	softness	previous use	temperature	count		
Notes Ries and Smith (1963) Che ▶ Nominal Logistic Regression	1	x	soft	yes	high	19		^
Normal Logistic Regression	2	x	soft	yes	low	57		
	3	x	soft	no	high	29		
	4	x	soft	no	low	63		
Columns (5/0)	5	m	soft	yes	high	29		
🔥 brand	6	m	soft	yes	low	49		
previous use *	7	m	soft	no	high	27		
temperature \star	8	m	soft	no	low	53		
a count	9	x	med	yes	high	23		
	10	x	med	yes	low	47		
	11	x	med	no	high	33		
	12	x	med	no	low	66		
	13	m	med	yes	high	47		
	14	m	med	yes	low	55		
Rows	15	m	med	no	high	23		
All rows 24 Selected 0	16	m	med	no	low	50		
Excluded 0	17	x	hard	yes	high	24		
Hidden 0	18	x	hard	yes	low	37		
Labelled 0	19		hard	no	high	42		~
evaluations done								

Binomial Logistic Regression (Demo): Fit

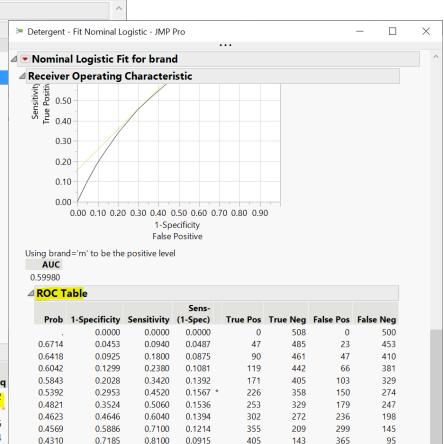
• Analyze > Fit Model

🏓 Fit Model - JMP Pro				—	\times
Model Specification					
Select Columns 5 Columns 5 Columns 5 Softness 5 previous use 5 temperature Count	Pick Role Variables Personali Y brand optional Weight optional numeric Freq count Validation optional By optional	vel:	m `	inal Logi Run ep dialo	~
	Construct Model Effects Add Cross Nest Macros ▼ Degree 2 Attributes ▼ Transform ▼ No Intercept Softness No Intercept Softness				

Binomial Logistic Regression (Demo): Report

- Overall model fit is significant
- Softness doesn't seem to contribute too much
 - Likelihood ratio test
- ROC Table
 - <u>sensitivity</u>, <u>specificity</u>, etc.
 - Tools -> Help

j	- Fit Nominal Logi	stic - JMP P	ro				_	
 Nomina 	al Logistic Fit f	or brand			-			^
Effect S	ummary	> Deterger	nt - Fit Nominal Lo					
Source								
softness	*temperature		nal Logistic Fi					
softness*temperature 0.043 softness 0.021								er Operating
Remove	<u>Add</u> <u>Edit</u> 🗌 Fl	DR ('^' deno	tes effec	ts with c	ontaining effects a	above them)	Sensitivity ue Positiv	0.50
onvorged in	Gradient, 3 iteratio	206					Sens True P	
eq: count	Gladient, 5 iteratio	5115					°, ⊨ 0	.40
Iteratio	ns						0	0.30
	Nodel Test						0	0.20
			-1.1-					
Model	-LogLikelihood			•	Prob>ChiSo		0	0.10
Difference Full	16.41281 682.24780	11	32.	.82562	0.0006'		0	0.00
Reduced	698.66061							0.00 0.10 0.20
RSquare (L	J)	0.0235						and='m' to be the
AICc		1388.81					AUC	
BIC		1447.48					0.59980	
Observatio	ons (or Sum Wgts)	1008					⊿ <mark>ROC</mark>	Table
Fit Deta	ils						Pro	b 1-Specificity
Parame	ter Estimates							. 0.0000
	ikelihood Ratio	Toste					0.671	
Effort Li		1 1 2 3 1 3					0.641	
Effect Li								
		Nr	arm	DE	L-R ChiSquare	Proh ChiSa		
Source		Np	oarm 2	DF 2	ChiSquare	Prob>ChiSq	0.584	3 0.2028
Source softness		Nŗ	parm 2 1			Prob>ChiSq 0.9522 <.0001*	0.584 0.539 0.482	3 0.2028 2 0.2953 1 0.3524
Source softness previous u		Nŗ	2	2 1	ChiSquare 0.09804239	0.9522	0.584 0.539 0.482 0.462	3 0.2028 2 0.2953 1 0.3524 3 0.4646
Source softness previous u	ise previous use	Nŗ	2 1	2 1	ChiSquare 0.09804239 22.1316677 3.78609668	<mark>0.9522</mark> <.0001*	0.584 0.539 0.482	3 0.2028 2 0.2953 1 0.3524 3 0.4646 9 0.5886
Source softness previous u softness*p temperatu	ise previous use	Nŗ	2 1 2	2 1 2 1	ChiSquare 0.09804239 22.1316677 3.78609668	0.9522 <.0001* 0.1506	0.584 0.539 0.482 0.462 0.456	3 0.2028 2 0.2953 1 0.3524 3 0.4646 9 0.5886 0 0.7185
Source softness previous u softness*p temperatu softness*te previous u	ise previous use rre emperature ise*temperature		2 1 2 1 2 1	2 1 2 1 2 1	ChiSquare 0.09804239 22.1316677 3.78609668 3.63914017 0.19617686 2.26089203	0.9522 <.0001* 0.1506 0.0564	0.584 0.539 0.482 0.462 0.456 0.431 0.416 0.410	3 0.2028 2 0.2953 1 0.3524 3 0.4646 9 0.5886 0 0.7185 7 0.8012 7 0.8661
Source softness previous u softness*p temperatu softness*te previous u	ise srevious use ree emperature		2 1 2 1 2	2 1 2 1 2 1	ChiSquare 0.09804239 22.1316677 3.78609668 3.63914017 0.19617686	0.9522 <.0001* 0.1506 0.0564 <mark>0.906</mark> 6	0.584 0.539 0.482 0.462 0.456 0.431 0.416	3 0.2028 2 0.2953 1 0.3524 3 0.4646 9 0.5886 0 0.7185 7 0.8012 7 0.8661 8 1.0000



0.8100

0.8700

0.9160

1.0000

1.0000

0.0688

0.0499

0.0000

0.0000

435

458

500

500

101

68

0

0

407

440

508

508

65

42

0

0

🏠 🖽 🔲 ▼

Your Turn (Hands-on)

- Do the same analysis without the softness variable
- Save the analysis script in the data table
- Challenge
 - How to construct a table of *correct classification rate* at each probability cutoff?
 - What cutoff gives the best correct classification rate?

correct classification rate = $\frac{crac}{---}$

true positive + true negative

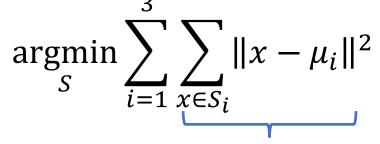
total # of predictions

K-Means Cluster Analysis

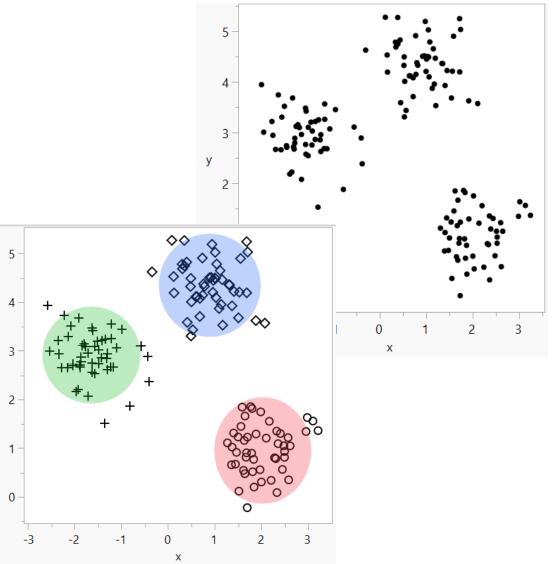
- A method to partition n observations into k clusters
 - such that total within-cluster sum of squares (between observations to cluster centroid) is minimized
- A **cluster** refers to a collection of data points aggregated together because of certain similarities
- Need to set k
 - There are methods to help you decide the value of k

K-Means Cluster Analysis: An Example

- Observations: 150 2-d points
- Set k = 3
 - partition each observation to one of the 3 clusters $S = \{S_1, S_2, S_3\}$
- K-means clustering algorithm finds
 3 clusters such that



Within-cluster sum of squares



K-mean Cluster Platform (Demo)

			🎋 K Means Cluster - JMP	Pro — 🗆	×
🔛 cluster - JMP Pro			Clusters rows based on num	neric variables into a specified number of clusters.	
File Edit Tables	Rows Cols DOE Anal	yze Graph Tools View Windo Distribution	Select Columns	Y, Columns 4 x	Ction OK
	1 2.6 2 0	Fit Y by X Tabulate Text Explorer	Columns Scaled Indivi	Weight optional numeric dually Freq By optional	Concol *☆ cluster - K Means Cluster □ × ···· ▲ Iterative Clustering ▲ Cluster Comparison
 Columns (2/0) x y 	3 4 0 5 2.6 6 1.5 7 1.7 8 2.3 9 -2.1 10 1.7	Fit Model Predictive Modeling Specialized Modeling Screening Multivariate Methods Clustering	Hierarchical Cluster	<pre> cluster - K Means Cluste</pre>	Method NCluster CCC Best K Means Cluster 3 20.5376 Optimal CCC Columns Scaled Individually Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Image: Control Panel Ima
Rows All rows 150 Selected 0 Excluded 0 Hidden 0 Labelled 0	16 -1.99912 17 -1.47804 18 1.87067	Quality and Process Reliability and Survival Consumer Research 714 2.71285741 153 3.2093591 766 0.77797407 443 2.76898682	Latent Class Analysis	Method K Means Cluster Number of Clusters Range of Clusters (Op Go Go Single Step	Biplot 3D Parallel Coord Plots Scatterplot Matrix
				 Use within-cluster std deviations Shift distances using sampling rates 	Simulate Clusters Save Cluster Distance

🟠 🖽 🔃 🔻

Your Turn (Hands-on)

- Import the country_risk.xlsx data (data/basics/country_risk.xlsx)
 - note that it's an Excel file and column header starts at row 2
- Perform a pair-wise correlation analysis across the following 5 variables
 - Corruption, Peace, Legal, GDP Growth, Population
 - Note that Corruption and Legal variables are highly correlated
 - hint: use the Multivariate platform
 - Menu: Analyze -> Multivariate Methods > Multivariate
- Perform a K-means cluster analysis
 - As a start, use Peace, Legal and GDP Growth as factors; and set k=3
 - Produce a scatterplot matrix
 - Can you label each cluster (high-risk, medium-risk, etc.)?
 - Hint: check Cluster Means

Model Selection

- Modeling for causal inference
 - Valid each model to check model assumptions are satisfied
 - e.g., analysis of residual in linear regression
 - Pick a common metric to compare across models and pick the best one
 - e.g., goodness of fit
- Modeling for prediction
 - Training, validation, and test data split
 - Training data: train/fit a model
 - Validation data: tune a model, and select the final best model based on a certain metric
 - Test data: obtain an unbiased performance measure of the final chosen model

Validation Column and Model Comparison

File Edit Tables Ro		Ana	<mark>lyze</mark> Graph Tools View W	indow	Help			_										
Companies_mma Source Bivariate			Distribution Fit Y by X Tabulate Text Explorer	-	ation Pro	ed Formula Profits (\$M) 11.889426293 527.65004251	Pred Formu	172.5	ts (\$M) 2 50108908 18414941	~								
	3 8		Fit Model	-	ation ation	157.5415523 672.730148			el Comparison performance		P Pro s models using predictio	n formula c	olumns.			_		×
Columns (10/0) Columns (10/0) Sales (\$M) Profits (\$M) # Employ Assets sales/emp Sales/emp Validation Pred Forofits (\$M) = Pred Forfits (\$M) =	5 12 6 54 7 6 8 5 9 18 10 4 11 13 12 28 13 95 14 8 15 83 16 3 17 34	418	Predictive Modeling • Specialized Modeling • Screening • Multivariate Methods • Clustering • Quality and Process • Reliability and Survival • Consumer Research • 784 0.283557636 small 6756.7 0.1947209302 medium	★ 課 ふ の ※ べ で	Naive Ba Support Model C	Tree t Neighbors			ilumns ; (\$M) ts (\$M) uploy ts /emp ation Formula Profit Formula Profit	ts (\$M)		Y, Prec Gro Wei	ijctors up ght (olumns into F Pred Form Pred Form optional Validation optional nume optional nume optional nume optional nume optional nume optional nume	ula Profits (S ula Profits (S eric eric	\$M) 2	Ca Rer Re	OK OK ancel move ecall telp
Rows	18 56 19 42	000	4500 0.1063571429 big 8324.8 0.1402304038 big		Formula				⊿ ▼ Model		nparison		•••					
All rows32Selected0Excluded0Hidden0Labelled0	20 31 21 8 22 8	404 3527 3578	5611.1 0.0942332187 medium 1791.7 0.1405300809 small 624.3 0.1154698065 small	Traini Traini Valida	ng	247.89836171 50.383042221 27.076010199		522. 178. 179.	Predict Measur Validatio Training Training	res of on Pre	of Fit for Profits (\$N redictor red Formula Profits (\$M) red Formula Profits (\$M)	Creator Fit Least	Squares		0.5482	RASE 319.22 443.97	213.24	Freq 24 24
evaluations done									Validation Validation	n Pre	red Formula Profits (\$M) red Formula Profits (\$M)	Fit Least	Squares	s	0.0437	1213.1 608.71	602.01	8

https://www.jmp.com/support/help/en/17.1/#page/jmp/example-of-model-comparison.shtml#

Outliers, Missing Values, and Patterns

📲 companies_mma - JN	1P Pro													
File Edit Tables Ro	ows Cols DOE Ana	<mark>Ilyze</mark> Graph Tools View W	/indo	w Help				🖉 companies_mma - Explore Outliers of # Employ - JMP Pro — 🗌 🗙						
i 🛤 🐑 💕 🖬 🐰		Distribution						⊿ 💌 Explore Outliers						
	💽 companies_mma 🕨 🔍 💽 🛛 💆 Fit Y by X							⊿ Commands						
 Source Fit Y by# Employ 	• F	Tabulate);	# Employ 50816		sales/emp 0.1327731423	size	Quantile Range Outliers Values farther than some quantile ranges from the tail quantile						
Data Filter	5 Cc	Text Explorer	4		2743.9	0.437852171	3	Robust Fit Outliers Given robust center and scale estimates, values far from center with respect to scale						
	6 Ph 7 Cc	Fit Model	7	54100		0.1741589649		Multivariate Robust Outliers Given a robust centers and covariance, measure Mahalanobis						
	8 Cc		_ 5	9500 5000	468.1	0.3027473684		distance						
	9 Cc	Predictive Modeling Specialized Modeling	4 4			0.14180		Multivariate k-Nearest Neighbor Outliers Outliers far from the kth nearest neighbors						
	10 Cc			nall 🛛		nall	⊿ Quantile Range Outliers							
Columns (7/0)	11 Cc	Multivariate Methods				nall	Outliers are values Q times the interquantile range past the lower and upper quantiles.							
🔥 Туре	12 Ph				Explore <mark>Missing Values</mark>		edium	Tail Quantile 0.05 Select columns and choose an action.						
Sales (\$M)	13 Cc	Quality and Process Reliability and Survival		Explore Patterns			g nall	Q Select Columns and choose an action.						
<pre>A Profits (\$M) A # Employ</pre>	14 Cc		0	Respon	se Screer	Restrict search to integers Exclude Rows Color Rows								
Assets sales/emp +	15 Ph 16 Ph				s Screenir	-	g nall	Show only columns with outliers Add to Missing Value Codes						
📥 size 🕂	17 Ph					5	edium	Rescan Change to Missing						
	18 Соттра			Predictor Screening		iing	g	Close						
	19 Pharma	aceutical 5903.7 68	1 🛒	Associa	ation Ana	lysis	g	10% 90% Low High Number of Outliers						
	20 Compu			Durana		Fueleses	edium	Column Quantile Quantile Threshold Threshold Outliers (Count)						
	21 Pharma		6	Process History Explorer		nall	# Employ 4795.6 82860 -229398 317053 1 383220							
	22 Compu		0.9	8578			small							
	23 Pharma		1.3	21300		0.1522535211								
	24 Compu		0.3	2900		0.4766551724								
	25 Compu 26 Compu		7.7 0.8	9100 10200		0.1114285714								
	26 Compu	iter 1769.2 6	0.8	10200	1269.1	0.1734509804	small							

https://www.jmp.com/support/help/en/17.1/#page/jmp/launch-the-explore-outliers-platform.shtml

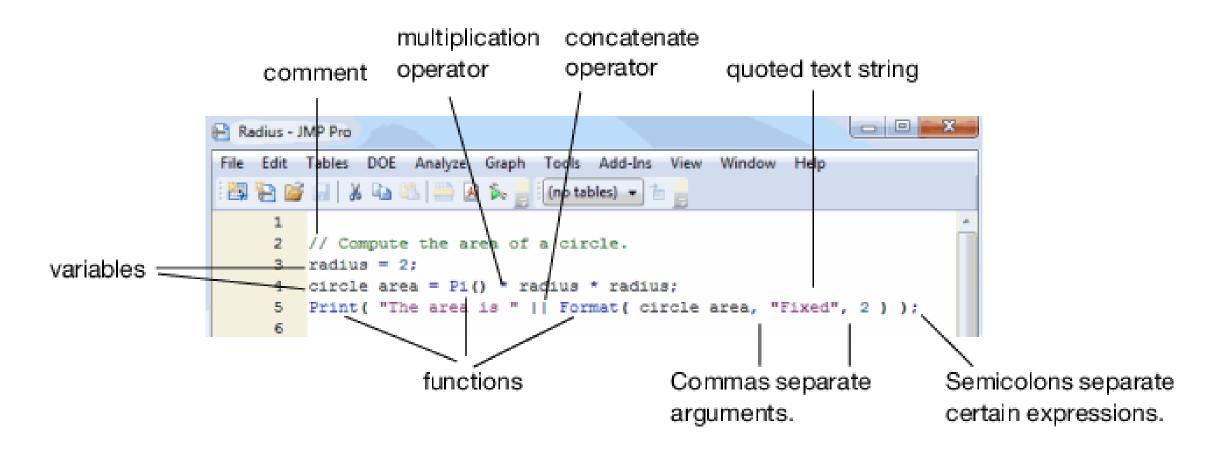
Plan

- Session 1
 - Workflow overview
 - Basic data manipulation
- Session 2
 - Join data tables
 - JMP graphing
- Session 3
 - Modelling
 - JMP Journal
 - JMP Scripting Language

JMP Journal – Communicate Your Results

- Create a JMP journal when you want to present your results
- A JMP journal combine two kind of presentations
 - Static: embed output of JMP (graphs and reports), fixed at a moment in time
 - Dynamic: built from outlines containing text and buttons (links) that organize data tables and reports
- Getting-started resources
 - Dmitry's video about JMP Journal on Quercus (6 mins)
 - <u>Creating, Using and Sharing JMP Journals</u> (1 hour 11 mins)

JMP Script Language (JSL)



https://www.jmp.com/support/help/en/17.1/#page/jmp/jsl-building-blocks.shtml#