INTRO TO JMP – PART 3

Bootcamp

August 7, 2022   Prepared by Jay / TDMDAL

Master of Management Analytics

Rotman School of Management
UNIVERSITY OF TORONTO

Quick Review of Session 2

- Join data tables
  - Many ways of joining two tables
    - Inner join, left join, left join with exclusion, etc.

  - Combing data manipulation operations to achieve a task (high-value customers)
    - Point-and-click
      - good for getting quick result
      - not too convenient for long pipeline data transformation (Reproducibility is a big issue)
    - For data manipulation on large dataset, order of operation steps could affect performance
    - Hard: turn a problem into implementation steps (algorithm); Easy: learn tools/languages

- JMP graphing
  - Graph builder

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.
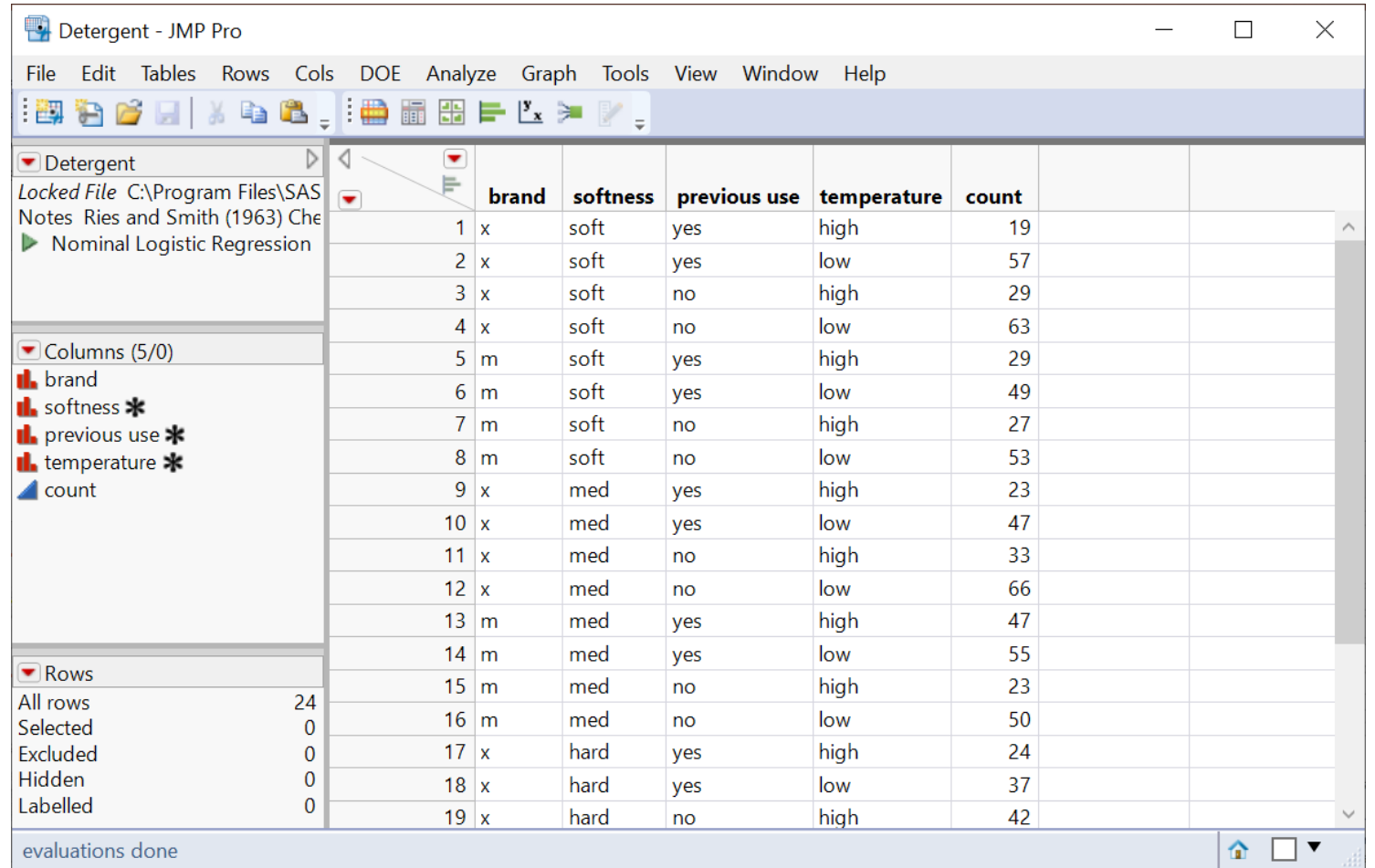
Binomial Logistic Regression

- 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

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

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

Binomial Logistic Regression (Demo): Fit

- Analyze > Fit Model

Binomial Logistic Regression (Demo): Report

- Overall model fit is significant

- Softness doesn't seem to contribute too much
  - Likelihood ratio test

- ROC Table
  - sensitivity, specificity, etc.
  - Tools -> Help
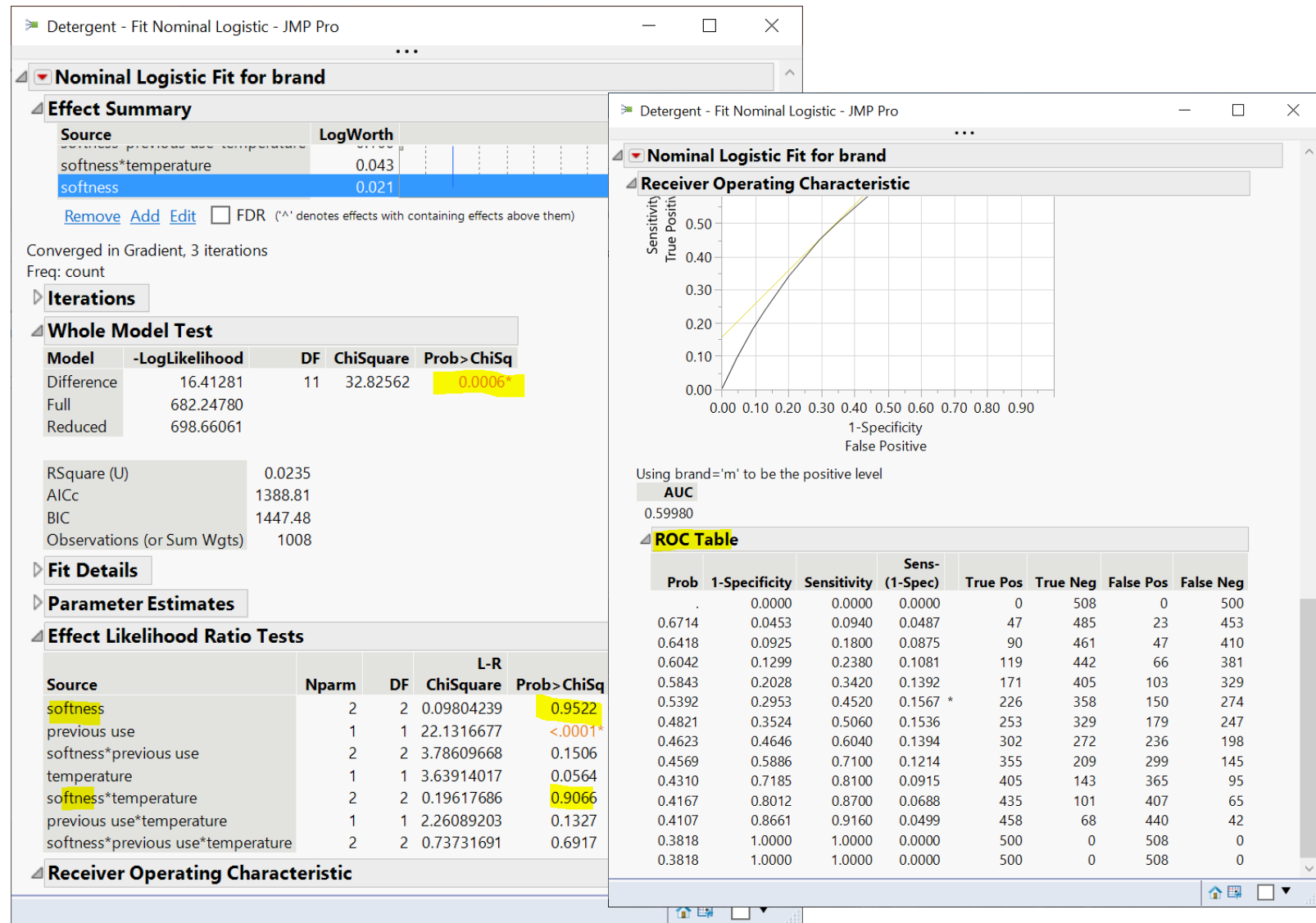
Your Turn (Hands-on)

- 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?

$$correct\ classification\ rate = \frac{true\ positive + true\ negative}{total\ \#\ of\ predictions}$$

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

$$\underset{S}{\text{argmin}} \sum_{i=1}^{3} \underbrace{\sum_{x \in S_i} \|x - \mu_i\|^2}_{\text{Within-cluster sum of squares}}$$
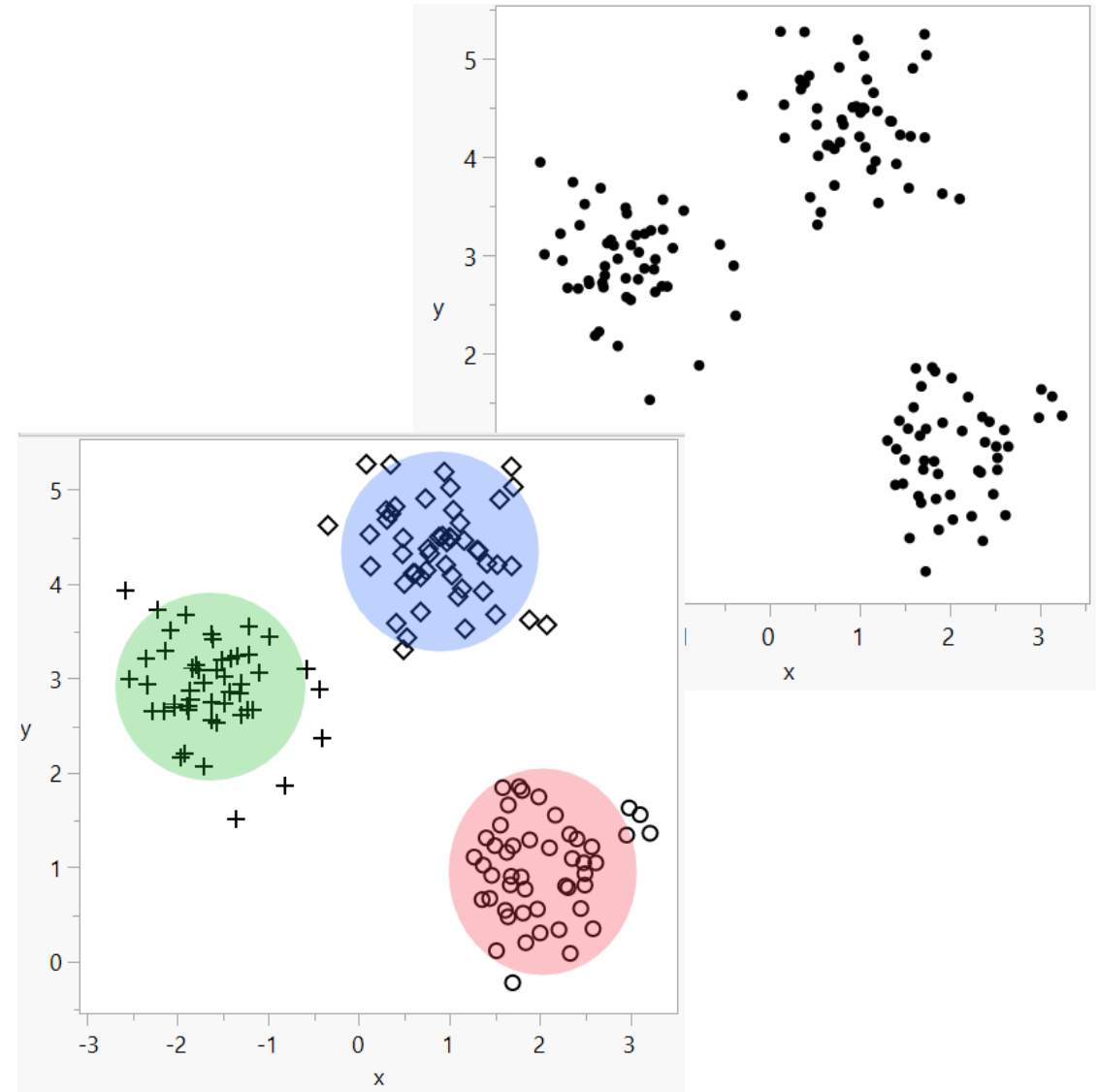
K-mean Cluster Platform (Demo)

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

- 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

Outliers, Missing Values, and Patterns

https://www.jmp.com/support/help/en/15.2/#page/jmp/modeling-utilities.shtml#
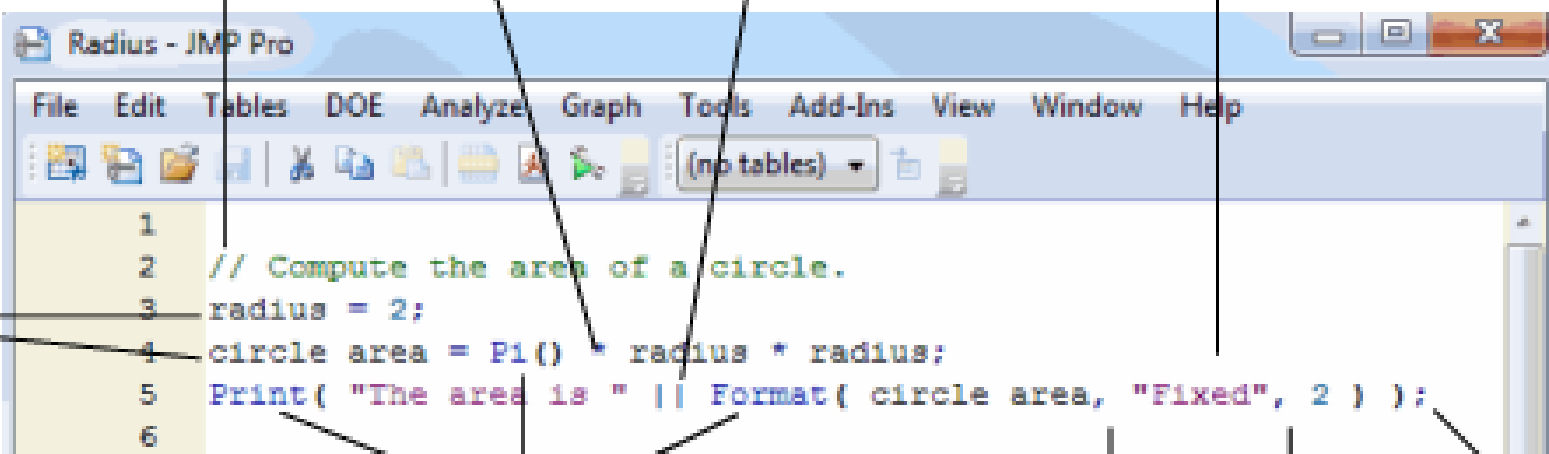
Plan

- 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)
  - [Creating, Using and Sharing JMP Journals](#) (1 hour 11 mins)

JMP Script Language (JSL)

https://www.jmp.com/support/help/en/15.2/#page/jmp/jsl-building-blocks.shtml#

What's Next & Others

- JMP
  - Barely touched its surface (especially on modeling and experimental design)
  - Continue to learn JMP on your own
  - Use it as a stepping stone to learn concepts and other tools/languages

- A tip on preparing technical interview questions
  - Create a tech interview question list on your own; start now, and do it every day

- Tips on studying
  - Understand instead of memorizing
  - Learn academic, practical and soft skills
  - Manage your time well, and study hard, but don't be a perfectionist
  - Often take a moment to step back to see the big picture, and take it easy
  - Enjoy your learning journey