

**DIY ADVANCED
MARKET RESEARCH
ANALYSIS
WORKSHOP**

|  |  |
| --- | --- |
| 9:30 | Registration |
| 9:50 | **Session 1: Correspondence Analysis** |
| 11:00 | Morning tea |
| 11:30 | **Session 2: Max-Diff** |
| 12:40 | Lunch |
| 1:40 | **Session 3: Driver Analysis** |
| 2:50 | Afternoon tea |
| 3:20 | **Session 4: Segmentation** |
| 4:30 | Drinks |



# Session 1: Correspondence analysis

All materials used today, as well as additional notes, can be downloaded from:
<http://wiki.q-researchsoftware.com/wiki/DIY_Advanced_Analysis>

The following blog posts from [www.displayr.com](http://www.displayr.com) provide a detailed discussion of the topics covered today:

* Use correspondence analysis to find patterns in large tables
* How correspondence analysis works (a simple explanation)
* How to interpret correspondence analysis plots (it probably isn’t the way you think)
* When to use, and not use, correspondence analysis
* Using correspondence analysis to compare sub-groups and understand trends
* Correspondence analysis versus multiple correspondence analysis: which to use and when?
* Easily add logos to a correspondence analysis map in Q
* Easily add images to a correspondence analysis map in Displayr
* Easily add images to a correspondence analysis plot in R

## Brand tracking data set

http://wiki.q-researchsoftware.com/images/0/04/Cola\_Tracking\_-\_January\_to\_December.sav

## When to use correspondence analysis

When we have a table with:

* At least two rows
* At least two columns
* No missing values
* No negatives
* Data on the same scale: Does the table cease to make sense if it is sorted by any of its rows or columns?
* No uninteresting outliers

## Animal – data



## Residuals – Data



## How to interpret correspondence analysis

* Compare between row labels based on distances (if *row principal* or *principal normalization*).
* Compare columns labels based on distances (if *column principal or principal normalization).*
* To compare a row label to a column label:
	+ Look at the length of the line connecting the row label to the origin. Longer lines indicate that the row label is highly associated with some of the column labels (i.e., it has at least one high residual).
	+ Look at the length of the label connecting the column label to the origin. Longer lines again indicate a high association between the column label and one or more row labels.
	+ Look at the angle formed between these two lines. Very small angles indicate association. 90 degree angles indicate no relationship. Angles near 180 degrees indicate negative associations (if *row principal, column principal,* or *symmetrical (1/2)*)*.*
* Always check conclusions using the raw data
* The lower the variance explained, the more we need to check the raw data



## Multiple correspondence analysis frequency tables



## Time series sales data

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| sales.data.png |  |  |  |  |  |  |  |  |  |  |  |  |

## Carbonated soft drink band associations

****

## Technology 2012

****

## Technology 2017



# Session 2: Max-Diff

All materials used today, as well as additional notes, can be downloaded from:
<http://wiki.q-researchsoftware.com/wiki/DIY_Advanced_Analysis>

The following blog posts from [www.displayr.com](http://www.displayr.com) provide a detailed discussion of the topics covered today:

* A beginner’s guide to max-diff
* How to create a max-diff experimental design in Q
* How to create a max-diff experimental design in Displayr
* How to create a max-diff experimental design in R
* How max-diff analysis works (simplish, but not for dummies)
* How to analyze max-diff data in Displayr
* How to analyze max-diff data in Q
* How to analyze max-diff data in R
* Using cross-validation to measure max-diff performance

## The brands

Apple, Microsoft, IBM, Google, Intel, Samsung, Sony, Dell, Yahoo, Nokia

## Data set

<http://wiki.q-researchsoftware.com/images/f/f1/Technology_2017.sav>

## The design

http://wiki.q-researchsoftware.com/images/7/78/Technology\_MaxDiff\_Design.csv

# Session 3: Driver analysis

All materials used today, as well as additional notes, can be downloaded from:
<http://wiki.q-researchsoftware.com/wiki/DIY_Advanced_Analysis>

Additionally:

* The PowerPoint deck on the wiki: **Session 3: Driver Analysis** contains detailed notes and instructions for replicating analyses.
* A video of similar content can be downloaded from:
<http://wiki.q-researchsoftware.com/wiki/DIY_Driver_Analysis>
* The following blog posts from [www.displayr.com](http://www.displayr.com) provide a detailed discussion of some aspects of the content:
	+ 5 ways to visualize relative importance scores from key driver analysis
	+ 4 reasons to compute importance using Relative Weights rather than Shapley Regression
	+ The difference between Shapley Regression and Relative Weights

## Data Sets

<http://wiki.q-researchsoftware.com/images/c/cf/Stacked_Technology.sav>

<http://wiki.q-researchsoftware.com/images/6/69/Stacked_Cola_Brand_Associations.sav>

## Basic process for a driver analysis

1. Import *stacked data*
2. Start with a linear regression model
3. Check the assumptions

|  |  |  |
| --- | --- | --- |
| Assumption to be checked | How to check | Options |
| 1. There is no multicollinearity/correlations between predictors (if using GLMs)
 | VIF, correlations, inspect coefficients  | Shapley, Relative Importance Analysis |
| 1. There are 15 or fewer predictors and we have multicollinearity/correlations between predictors
 | Count them | Relative Importance Analysis |
| 1. The outcome variable is monotonically increasing
 | Create a table and use judgment | Make it numeric (merge, recode, missing values) |
| 1. The outcome variable is numeric
 | Judgement | Change to an appropriate **Type** (usually binary or orderd logit, or quasi-Poisson if counts) |
| 1. The predictor variables are numeric or binary
 | Create a table and use judgment | Make them binary or numeric (merge, recode, missing values) |
| 1. People do not differ in their needs/wants (segmentation)
 | 1. Latent class analysis
2. Run models by segment
 | 1. Presented segmented analysis
2. Ignore the problem (99.9% of the time)
 |
| 1. The causal model is plausible
 |  “Common sense” | Cross fingers |
| 1. The correlations between the predictors and the outcome variable are sensible
 |  Compute correlations | Remove problematic variables |
| 1. The signs from coefficients from a traditional linear regression make sense (NB: where there are high correlations from predictor variables, these signs may not make sense)
 | Q, Displayr, and flipRegression::Regression append the regression coefficients to the driver analysis scores.  | Make them absolute, explain to client, remove from model |
| 1. The predictor variables have no missing values
 |  Create tables and look at the sample sizes. | * Bespoke model
* Multiple imputation
* Cross fingers
 |
| 1. There are no outliers/influential data points
 | * Hat/influence scores
* Standardized residuals
* Cook’s distance
 | * Inspect and filter if appropriate
* Robust methods
* Cross fingers
 |
| 1. There is no serial correlation (aka autocorrelation)
 |  Durbin-Watson test | * Bespoke model
* Be cautious about stat tests
 |
| 1. The residuals have constant variance (i.e., no heteroscedasticity in a model with a linear outcome variable)
 | Breusch-Pagan test | Use a more appropriate model (e.g., *ordered logit*)Use *robust standard errors* |

## Cola brand

### Outcome variable



### Predictor variables



## Technology – Outcome variable and predictors



# Session 4: Segmentation

All materials used today, as well as additional notes, can be downloaded from:
<http://wiki.q-researchsoftware.com/wiki/DIY_Advanced_Analysis>

Additionally:

* The PowerPoint deck on the wiki: **Session 4: Segmentation** contains detailed notes and instructions for replicating analyses.
* A video of similar content can be downloaded from:
<http://wiki.q-researchsoftware.com/wiki/DIY_Driver_Analysis>
* The following blog posts from [www.displayr.com](http://www.displayr.com) provide a detailed discussion of some aspects of the content:
	+ 5 Ways to Deal with Missing Data in Cluster Analysis
	+ Assigning respondents to clusters/segments in new data files in Q
	+ Assigning respondents to clusters/segments in new data files in Displayr

## What consultants think clients want



## US General Social Survey – attitudes to US institutions

### Confidence in



### Importance

