RESEARCHSOFTWARE
A DIVISION OF DISPLAYR

TIM BOCK PRESENTS

DIY Advanced Analysis

Session 3: Driver Analysis

Overview

- Objectives of (key) driver analysis
- Overview of techniques
- Assumptions that need to be checked when doing QA for driver analysis
- Visualization

The basic objective of (key) driver analysis

The basic objective: work out the relative importance of a series of *predictor variables* in predicting an *outcome variable*. For example:

- NPS: comfort vs customer service vs price.
- Customer satisfaction: wait time vs staff friendliness vs comfort.
- Brand preference: modernity vs friendliness vs youthfulness.

What driver analysis is not: predictive analysis (e.g., predicting sales, customer churn). Although, you can use driver analysis to make strategic predictions (e.g., if I improve, say, *fun,* then preference will increase.)

Basic process for driver analysis

- Import *stacked* data
- Start with a linear regression model
- Check the assumptions

What the data looks like

1 outcome variable

Predictor variables
(Typically there will be more than 3.)

This data shows 7 observations

Likelihood to recommend	This brand is *fun*	This brand is *exciting*	This brand is *youthful*
6	1	1	1
9	0	1	0
7	0	0	0
6	1	1	1
9	0	1	0
7	0	0	1
7	0	0	0

Case study 1: Cola brand attitude

Outcome variable(s)	**34 Predictor variable(s)**	*If the brand was a person, what would its personality be?*	
Hate/Dislike/Neither/ Like/Love/Don't know: • Coke Zero • Coke • Diet Coke • Diet Pepsi • Pepsi Max • Pepsi	Brand associations: • Beautiful • Carefree • Charming • Confident • Down-to-earth • Feminine • Fun • Health-conscious • Hip • Honest • Humorous	• Imaginative • Individualistic • Innocent • Intelligent • Masculine • Older • Open to new experiences • Outdoorsy • Rebellious • Reckless • Reliable	• Sexy • Sleepy • Tough • Traditional • Trying to be cool • Unconventional • Up-to-date • Upper-class • Urban • Weight-conscious • Wholesome • Youthful

Case study 2 (time permitting): Technology

Outcome variable(s)	Predictor variable(s)
Likelihood to recommend: • Apple • Microsoft • IBM • Google • Intel • Hewlett-Packard • Sony • Dell • Yahoo • Nokia • Samsung • LG • Panasonic	Brand associations: • Fun • Worth what you pay for • Innovative • Good customer service • Stylish • Easy-to-use • High quality • High performance • Low prices

The data (stacked)

From: one row per respondent

To: one row per brand per respondent

	Likelihood to recommend			This brand is *fun*			This brand is *exciting*		
ID	Apple	Microsoft	IBM	Apple	Microsoft	IBM	Apple	Microsoft	IBM
1	6	9	7	1	0	0	1	1	0
2	8	7	7	1	0	0	1	0	0
3	0	9	8	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0

ID	Brand	Likelihood to recommend	This brand is *fun*	This brand is *exciting*
1	Apple	6	1	1
1	Microsoft	9	0	1
1	IBM	7	0	0
2	Apple	6	1	1
2	Microsoft	9	0	1
2	IBM	7	0	0
3	Apple	6	1	1
3	Microsoft	9	0	1
3	IBM	7	0	0
4	Apple	6	1	1
4	Microsoft	9	0	1
4	IBM	7	0	0

Tips for stacking

Q

- Get an SPSS .SAV data file. If you do not have an SPSS file:
  - Import your data the usual way
  - **Tools > Save Data as SPSS/CSV** and **Save as type: SPSS**
  - Re-import
- **Tools > Stack SPSS .sav Data File**
- Set the labels for the stacking variable (in Q: `observation`) in **Value Attributes**
- Delete any *None of these* data (e.g., brand associations where respondents were able to select *None of these*

R / Displayr

The R function `reshape`

Standard "best practice" recommendation for driver analysis:

The average improvement in R^2 that a predictor makes across all possible models (aka "Shapley")

LMG

Lindeman, Merenda, Gold (1980)

=

Kruskal

Kruskal (1987)

=

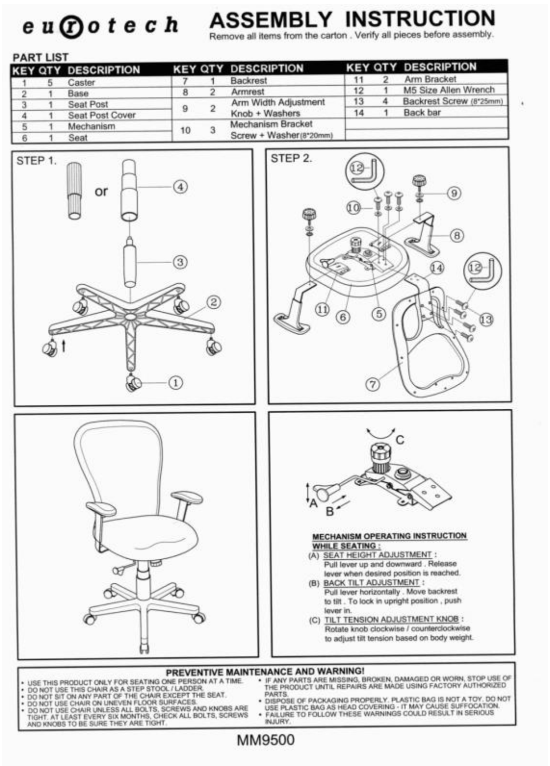
Dominance Analysis

Budescu (1993)

=

Shapley / Shapley Value

Lipovetsky and Conklin(2001)


eurotech ASSEMBLY INSTRUCTION
Remove all items from the carton . Verify all pieces before assembly.

PART LIST

KEY	QTY	DESCRIPTION	KEY	QTY	DESCRIPTION	KEY	QTY	DESCRIPTION
1	5	Caster	7	1	Backrest	11	2	Arm Bracket
2	1	Base	8	2	Armrest	12	1	M5 Size Allen Wrench
3	1	Seat Post	9	2	Arm Width Adjustment Knob + Washers	13	4	Backrest Screw (8*25mm)
4	1	Seat Post Cover				14	1	Back bar
5	1	Mechanism	10	3	Mechanism Bracket Screw + Washer(8*20mm)			
6	1	Seat						

STEP 1.

or

STEP 2.

MECHANISM OPERATING INSTRUCTION WHILE SEATING :
(A) SEAT HEIGHT ADJUSTMENT :
Pull lever up and downward . Release lever when desired position is reached.
(B) BACK TILT ADJUSTMENT :
Pull lever horizontally . Move backrest to tilt . To lock in upright position , push lever in.
(C) TILT TENSION ADJUSTMENT KNOB :
Rotate knob clockwise / counterclockwise to adjust tilt tension based on body weight.

PREVENTIVE MAINTENANCE AND WARNING!
- USE THIS PRODUCT ONLY FOR SEATING ONE PERSON AT A TIME.
- DO NOT USE THIS CHAIR AS A STEP STOOL / LADDER.
- DO NOT SIT ON ANY PART OF THE CHAIR EXCEPT THE SEAT.
- DO NOT USE CHAIR ON UNEVEN FLOOR SURFACES.
- DO NOT USE CHAIR UNLESS ALL BOLTS, SCREWS AND KNOBS ARE TIGHT. AT LEAST EVERY SIX MONTHS, CHECK ALL BOLTS, SCREWS AND KNOBS TO BE SURE THEY ARE TIGHT.
- IF ANY PARTS ARE MISSING, BROKEN, DAMAGED OR WORN, STOP USE OF THE PRODUCT UNTIL REPAIRS ARE MADE USING FACTORY AUTHORIZED PARTS.
- DISPOSE OF PACKAGING PROPERLY. PLASTIC BAG IS NOT A TOY. DO NOT USE PLASTIC BAG AS HEAD COVERING - IT MAY CAUSE SUFFOCATION.
- FAILURE TO FOLLOW THESE WARNINGS COULD RESULT IN SERIOUS INJURY.

MM9500

Much too hard

Best practice:
Bespoke models (e.g., Bayesian multilevel model)

Too hard

GLMs (e.g., linear regression)

Too Soft

Bivariate metrics
E.g., Correlations, Jaccard Coefficients

Just Right

Shapley, Relative Importance Analysis

What makes bespoke models and GLMs too hard?

To estimate an OK bespoke model, you need to have a few week, and know lots of things, including:

- Joint interpretation of parameter estimates, the predictor covariance matrix, and the parameter covariance matrix
- Conditional effects
- Multicollinearity
- Confounding (e.g., suppressor effects)
- Estimation (ML, Bayesian)
- Specification of informative priors
- Specification of random effects

To understand importance in a GLM (e.g., linear regression), you need to know quite a lot about:

- Joint interpretation of parameter estimates, the predictor covariance matrix, and the parameter covariance matrix
- Conditional effects
- Multicollinearity
- Confounding (e.g., suppressor effects)

Shapley and similar methods allow us to be less careful when interpreting results

Bespoke models & GLMs

Relative Importance Analysis

AKA Relative Weight: Johnson (2000)

Random Forest

(for importance analysis)

Shapley

Shapley

With coefficient adjustment
Lipovetsky and Conklin(2001)

Kruskal's Squared partial correlation

Called **Kruskal** in Q

Proportional Marginal Variance Decomposition

Creating Shapley analysis in Q

- Open `Initial.Q.` This already contains the cola data.
- **File > Data Sets > Add to Project > From File >** `Stacked Technology`
- **Create > Regression > Driver (Importance) Analysis > Shapley**
- Dependent variable: **Q3. Likelihood to recommend [Stacked Technology]**
- Dependent variable: **Q4** variables from `Stacked Technology`
- **No** when asked about confidence intervals (clicking Yes is **OK** as well)
- *Note that* High Quality *is the most important, with a score of 18.2*
- Right-click: **Reference name**: `shapley`

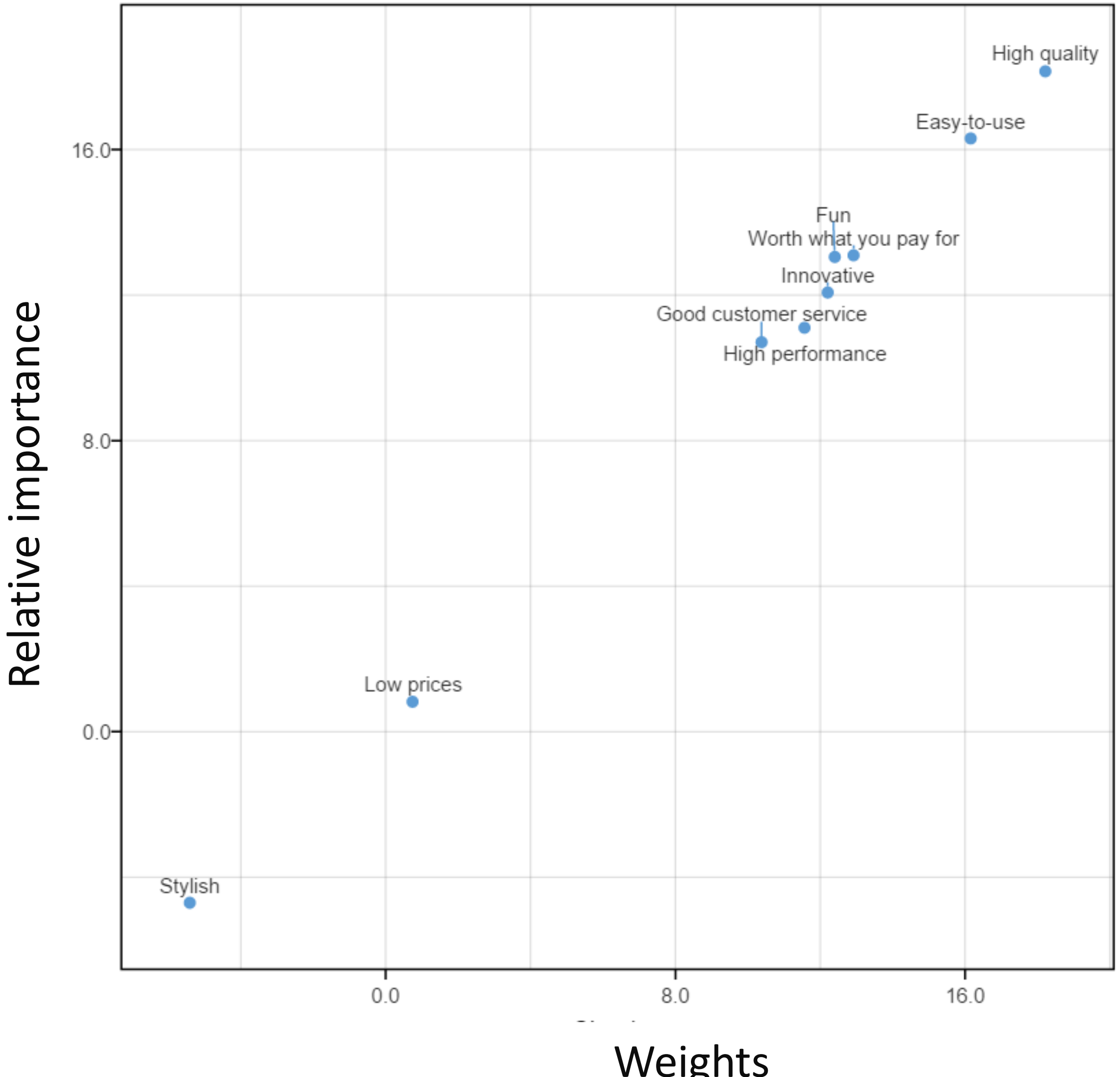
Everything I demonstrate in this webinar is described on a slide like this. The rest of them are hidden in this deck, but you can get them if you download the slides. So, there is no need to take detailed notes.

Shapley and Relative Importance Analysis give very similar results (Case Study 2)

The plot on the right shows that we get very similar results from performing driver analysis using Shapley and Relative Importance Analysis.

Please see the following blog posts for more on this:

- *4 reasons to compute importance using Relative Weights rather than Shapley Regression*
- *The difference between Shapley Regression and Relative Weights*



Basic process for driver analysis

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

1: There is no multicollinearity/correlations between predictors (if using GLMs, e.g., linear regression)

Issue

The bigger the correlations between predictors, the more difficult it is to accurately interpret estimates from traditional GLMs (e.g., linear regression)

Test

1. Inspect the *Variance Inflation Factors (VIF)* or *Generalized Variance Inflation Factors (GVIF).* Q automatically computes these and warns you if they are high.
2. Inspect the coefficients. Do they make sense?
3. Look at the correlations.

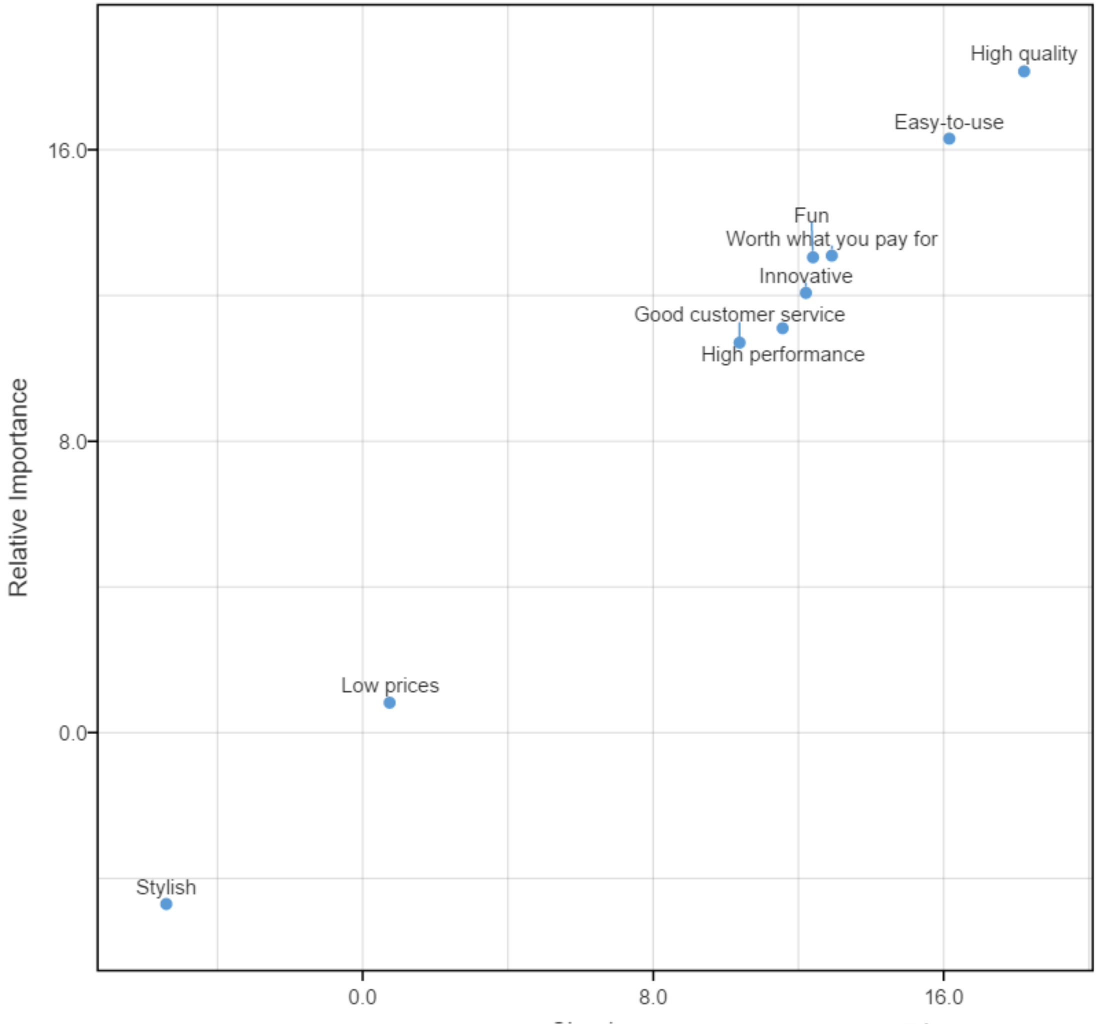
Options (ranked from best to worst)	**Comments**
Take all the relevant theory into account when interpreting the results.	This requires a strong technical and intuitive understanding of the underlying maths. Even if you possess that understanding, it is really difficult to explain to clients (particularly if it is a tracking study and they are seeing results fluctuate from period-to-period)
Use *Shapley* or *Relative Importance Analysis*.	These techniques are designed to address this problem. They are not perfect, but they are easier to interpret than linear regression and other GLMs when predictor variables are correlated.

2: There are 15 or fewer predictors (if using Shapley)

- *With the cola study, we have 34 variables, and that will take an infinite amount of time to compute, so using Shapley is not an option and we have to use Relative Importance Analysis.*
- *We can use the technology data set, which only has 9 predictors, to explore how similar the techniques are.*
- **Create > Regression > Linear Regression**
  - **Reference name:** `relative.importance`
  - **Select variables**
  - **Output: Relative importance analysis**
  - Check **Automatic *Note that High Quality is again most important***
- Right-click: **Add R Output:**

```
comparison = cbind(shapley = shapley[-10],
       "Relative Importance" =
relative.importance$relative.importance$importance)
```

- **Calculate**
- Change `shapley` to `shapley[-10]`
- **Calculate**
- Right-click: **Add R Output:** `correlation = cor(comparison)`
- Increase number of decimal places. Note the correlation is 0.999
- Rename output: **Correlation**
- **Insert > Charts > Visualization > Labeled Scatterplot,**
  - **Table: comparison**
  - **Automatic**



3: The outcome variable is monotonically increasing

Issue

All the standard *driver analysis* algorithms assume that the *outcome variable* contains categories ordered from lowest to highest, and which are believed to be associated with greater levels of preference.

Test

This is usually best checked by creating a *summary table.*

Options (not mutually exclusive)	**Comments**
Set Don't Knows to missing	
Merge categories	• Do this when there are categories that have ambiguous orderings (e.g., *OK* and *Good*). • The more categories you merge, the less significant the results will be.
Recode the data in some meaningful way (e.g., reverse the scale, Likelihood to recommend, recoded as NPS)	The specific values tend to make little difference, so using a recoding that is easy to explain to stakeholders, such as NPS, is often desirable.

4: The outcome variable is numeric (if using Shapley)

Issue

Shapley assumes that the outcome variable is numeric (theoretically, it can deal with non-numeric outcome variables, but for more than about 10 or so variables, it is impractical).

Options (ranked from best to worst)	**Comments**
Use limited dependent variable versions of *Relative Importance Analysis* (e.g., *Ordered Logit*)	• The less numeric the variable, the better this option is. • This approach is also preferable because it can take non-linear relationships into account automatically.
Ignore the problem and use *Shapley*.	Where the variable is close to being numeric, there is probably little lost by this approach.

5: The predictor variables are numeric or binary

Issue

Both *Shapley* and *Relative Importance* Analysis assume that the predictor variables are numeric or binary.

Options (not mutually exclusive)	**Comments**
Set Don't Knows to missing	This can be problematic as the variables as the missing values may not be missing at random. This is discussed later.
Merge categories	• Do this when there are categories that have ambiguous orderings (e.g., *OK* and *Good*). • The more categories you merge, the less significant the results will be.
Recode the data in some meaningful way (midpoint recoding)	
Use a bespoke or *Generalized Linear Model (GLM)*, with *dummy variables* and/or *splines*, computing importance as the difference between the lowest and largest effect sizes for each variable.	In theory this is the best approach to dealing with non-numeric data, but it requires quite a lot to get right and, when interpreting the data, the sampling error of the categorical and spline effects will make them hard to compare.

6: People do not differ in their needs/wants (segmentation)

Issue

Traditional driver analysis techniques assume that people have the same needs/wants, and apply these consistently from situation to situation.

How to test

- Compare by brand
- Compare by other data
- Latent class analysis

Options (not mutually exclusive)	**Comments**
Estimate an appropriate bespoke model (e.g., latent class analysis) and then estimate the driver analysis models within each segment	In Q: In a non-stacked data file, set up the data as an **Experiment**, and use **Create > Segment > Latent Class Analysis**
Form segments by judgment, and estimate separate relative importance analyses for each segment.	
Ignore the problem, interpreting results as "average" effects	Rightly-or-wrongly, this is how 99.9%* of all modelling is done. * Made-up number

7: The causal model is plausible

Issue

All driver analysis techniques assume that the analysis is a plausible explanation of the causal relationship between the predictor variables and the outcome variable.

This assumption is never true.

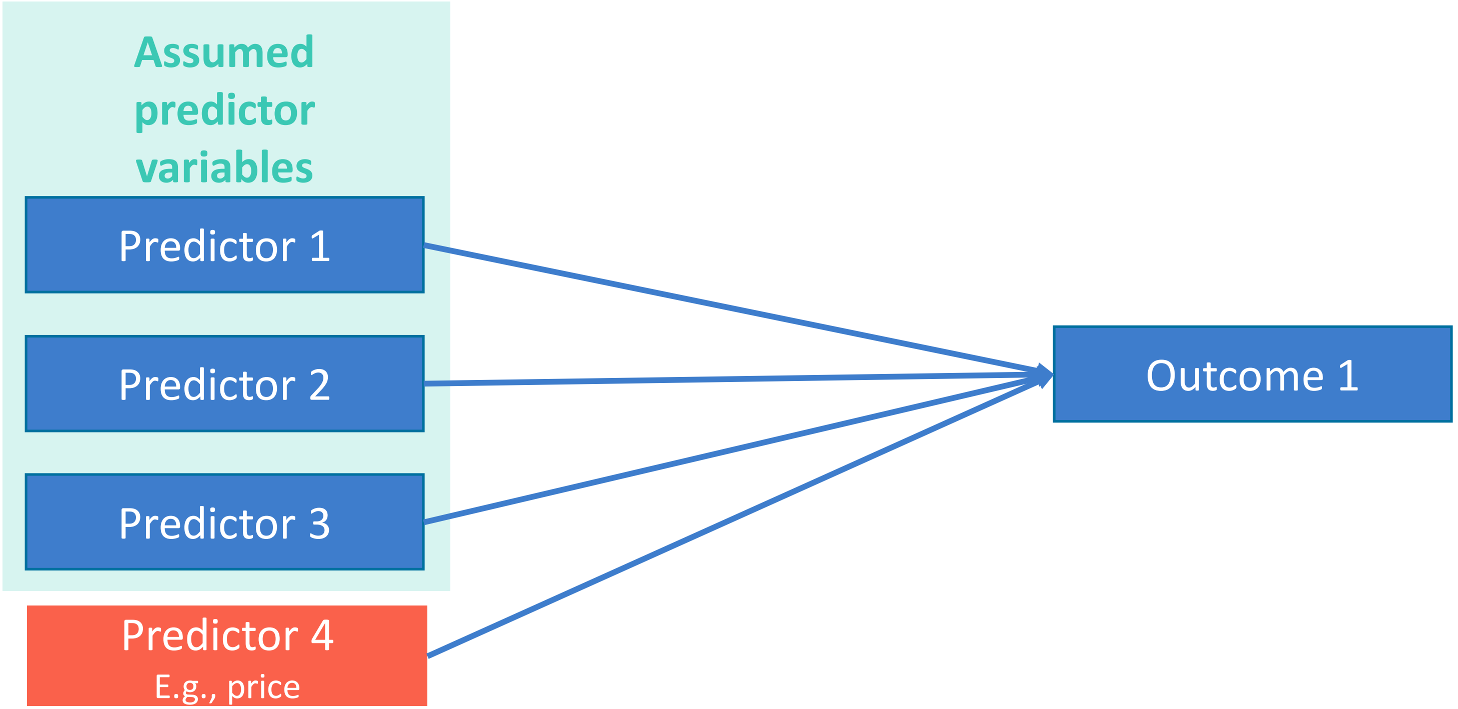
How to test

Common sense. Four common examples are shown on the next slides.

Options (not mutually exclusive)	**Comments**
Build a bespoke model	This is usually too hard
Include all the relevant (non-outcome) variables and cross your fingers (if you have not collected the data, you cannot magic it into existence)	Rightly-or-wrongly, this is how 99.9%* of all modelling is done * Made-up number

Example causality problem: Omitted variable bias

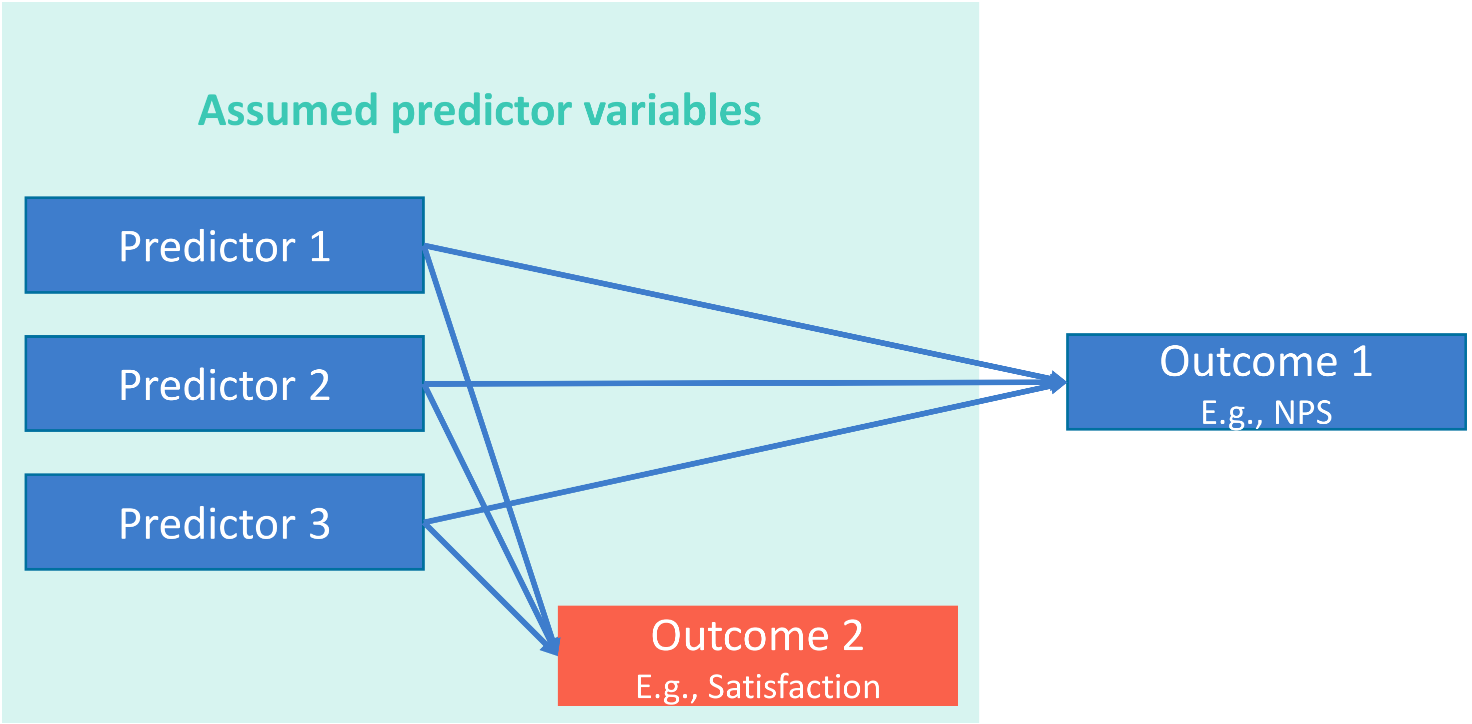
If we fail to include a relevant predictor variable, and that variable is correlated with the predictor variables that we do include, the estimates of importance will be wrong. If your R-square is less than 0.9, you may have this problem (a typical R-square is closer to 0.2 than 0.9).



Arrows denote the true causal relationship

Example causality problem:
Outcome variable included as a predictor

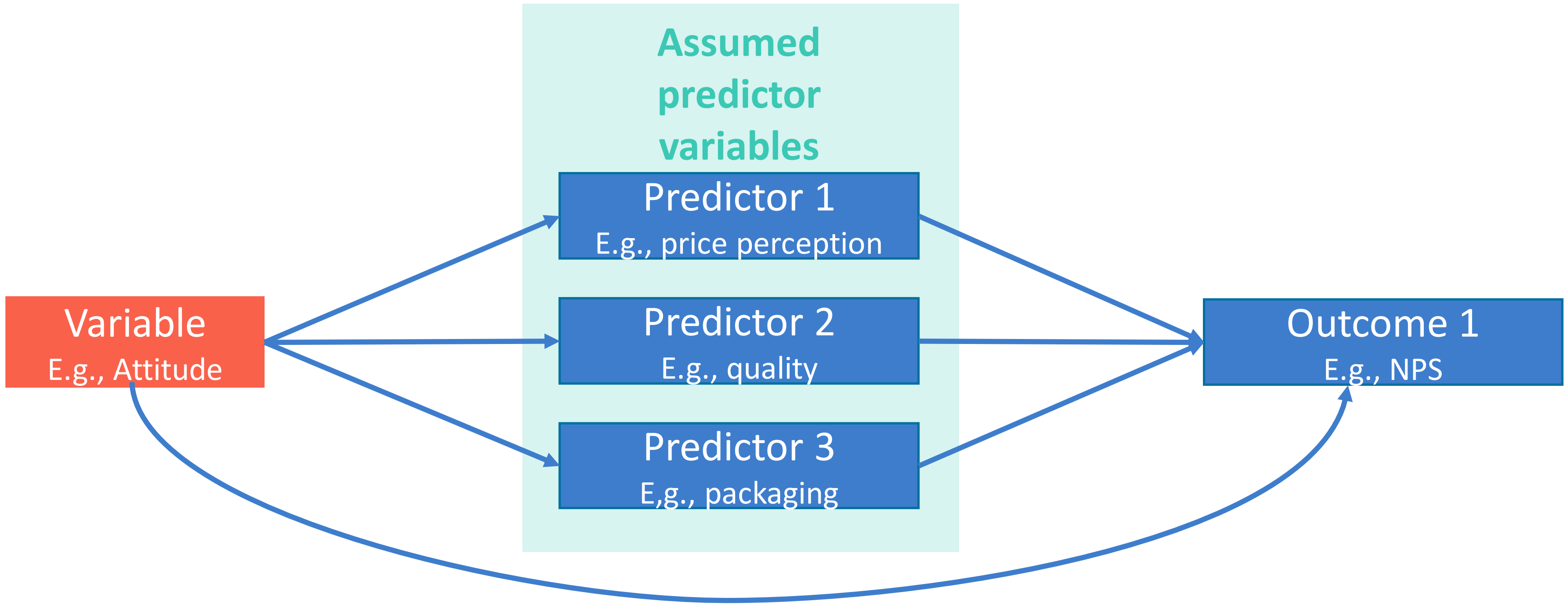
If we include a predictor variable that is really an outcome variable, the estimates of importance will be wrong.



Arrows denote the true causal relationship

Example causality problem: Backdoor path

If *backdoor path* exists from the predictors to the outcome variable, the estimates of importance will be wrong *(spurious).*



Arrows denote the true causal relationship

Example causality problem: Functional form

If we have the wrong functional form (i.e., assumed equation), the estimates of importance will be wrong.

Assumed functional form

Outcome = Predictor 1 + Predictor 2 + Predictor 3

True functional form

Outcome = Predictor 1 × Predictor 2 + Predictor 3

Arrows denote the true causal relationship

8: There are no unexpected correlations between the predictors and the outcome variable

Issue When people interpret importance scores, they assume that higher means better. This is assumption is not always right. **Test** Correlate each predictor variable with the outcome variable	**Options** (ranked from best to worst)
	Investigate the data to make sense of the unexpected relationships.
	Remove problematic variables from the analysis.

9: The signs of the importance scores are correct

Issue

The underlying *Shapley* and *Relative Importance Analysis* algorithms always compute a positive importance scores.

However, the true effect of a predictor can be negative, resulting in people misinterpreting the results.

Test

Compute a GLM (e.g., linear regression). Any negative coefficients warrant investigation. For this reason, Q automatically does this and puts the signs of the multiple regression coefficients onto the driver analysis outputs (both *Shapley* and *Relative Importance Analysis*).

If the correlation is also negative, it means that the effect is negative. If positive, it suggests that the multiple regression is picking up a non-interesting artefact.

Recommendation

If all the effects should be positive, select the **Absolute importance scores** option. Otherwise, manually change the results when reporting.

10: The predictor variables have no missing values

Issue

There are missing values of predictor variables (e.g., some attributes were not collected for some respondents, or there were “don’t know” response)

Options (ranked from best to worst)	**Comments**
Create a bespoke model that appropriately models the process(es) that cause the values to be missing.	This is really hard!
Multiple imputation of missing values	If using *Relative Importance Analysis*, set **Missing Data** to **Multiple Imputation**
Leave out observations with missing values from the analysis (i.e., *complete case analysis*)	This implicitly assumes that the data is **M**issing **C**ompletely At **R**andom (**MCAR**; i.e., other than that some variables have more missing values than others, there is no pattern of any kind in the missing data). Test this assumption using **Automate > Browse Online Library > Missing Data > Little’s MCAR Test**

11: There are no outliers/unusual data points

Issue

A few outliers/unusual observations can skew the results of importance analysis.

Test

- Hat/influence scores
- Standardized residuals
- Cook's distance

Options (ranked from best to worst)	**Comments**
Inspect each unusual observation, and understand if it is an error or not	Difficult/time consuming
Filter out all the unusual observations, and check to see if the model has changed. If it has changed, and the number of unusual observations is small, use the new model.	
Ignore the problem	This is, by far, the most common approach.

12: There is no serial correlation (aka autocorrelation)

Issue

The standard tests for the significance of a predictor assume that there is no serial correlation/autocorrelation (a particular type of pattern in the residuals).

Whenever you stack data you are highly likely to have this problem.

Test

Regression > Diagnostic > Serial Correlation (Durbin-Watson)

Options (ranked from best to worst)	Comments
Create a bespoke model that addresses the serial correlation (e.g., a random effects model if the serial correlation is due to repeated measures, or a time series model if it is measures over time)	This is a lot of work.
Don't report statistical test results (i.e., *p*-values).	The importance scores will be OK. The significance tests will be misleading to an unknown extent.

13: The residuals have constant variance (i.e., no heteroscedasticity in a model with a linear outcome variable)

Issue

The standard tests for the significance of a predictor in a linear model assume that the variance of the residuals is constant.

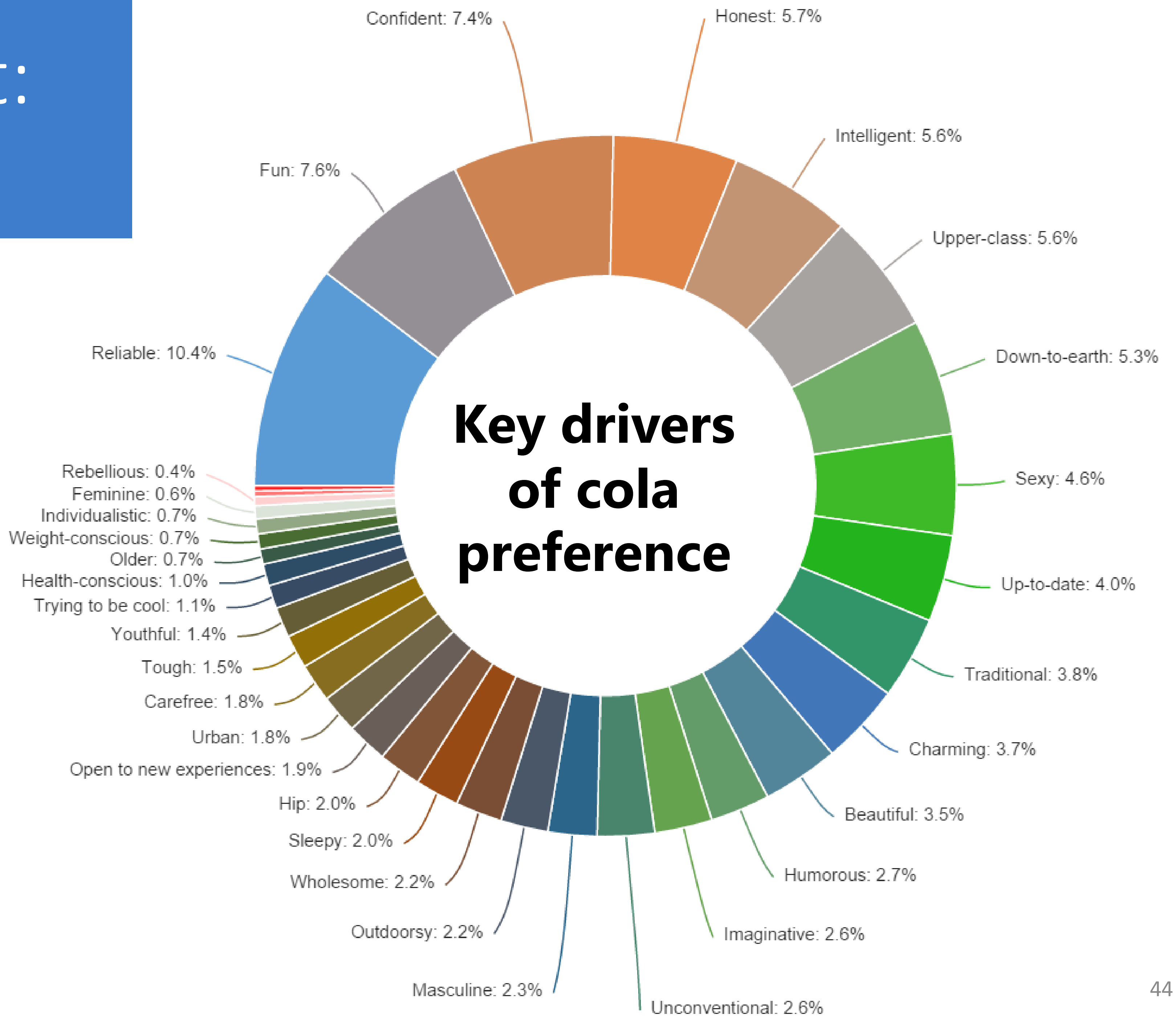
This is rarely the case in driver analysis, as usually the data is from a bounded scale (e.g., if it is a rating out of 10, it is impossible for a value to be observed that is greater than 10).

Test

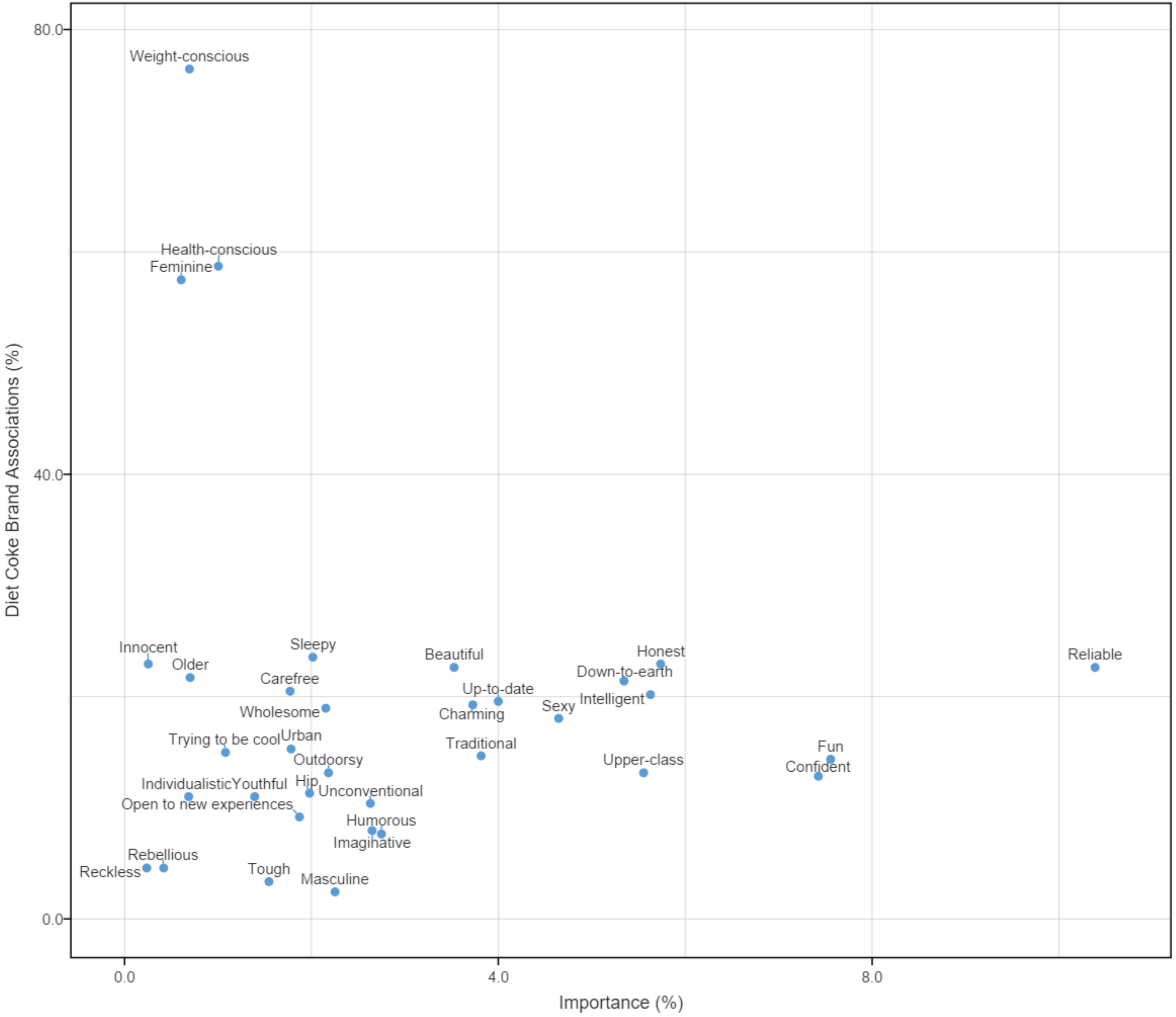
Displayr automatically performs the *Breusch-Pagen Test* **Type** = **Linear**

Options (ranked from best to worst)	**Comments**
Use a more appropriate model (e.g., *ordered logit*)	This is not possible with Shapley. This models make other, hopefully less problematic, assumptions (beyond the scope of this webinar)
Use *robust standard errors*	This is not possible with Shapley. In Q: check **Robust standard error**

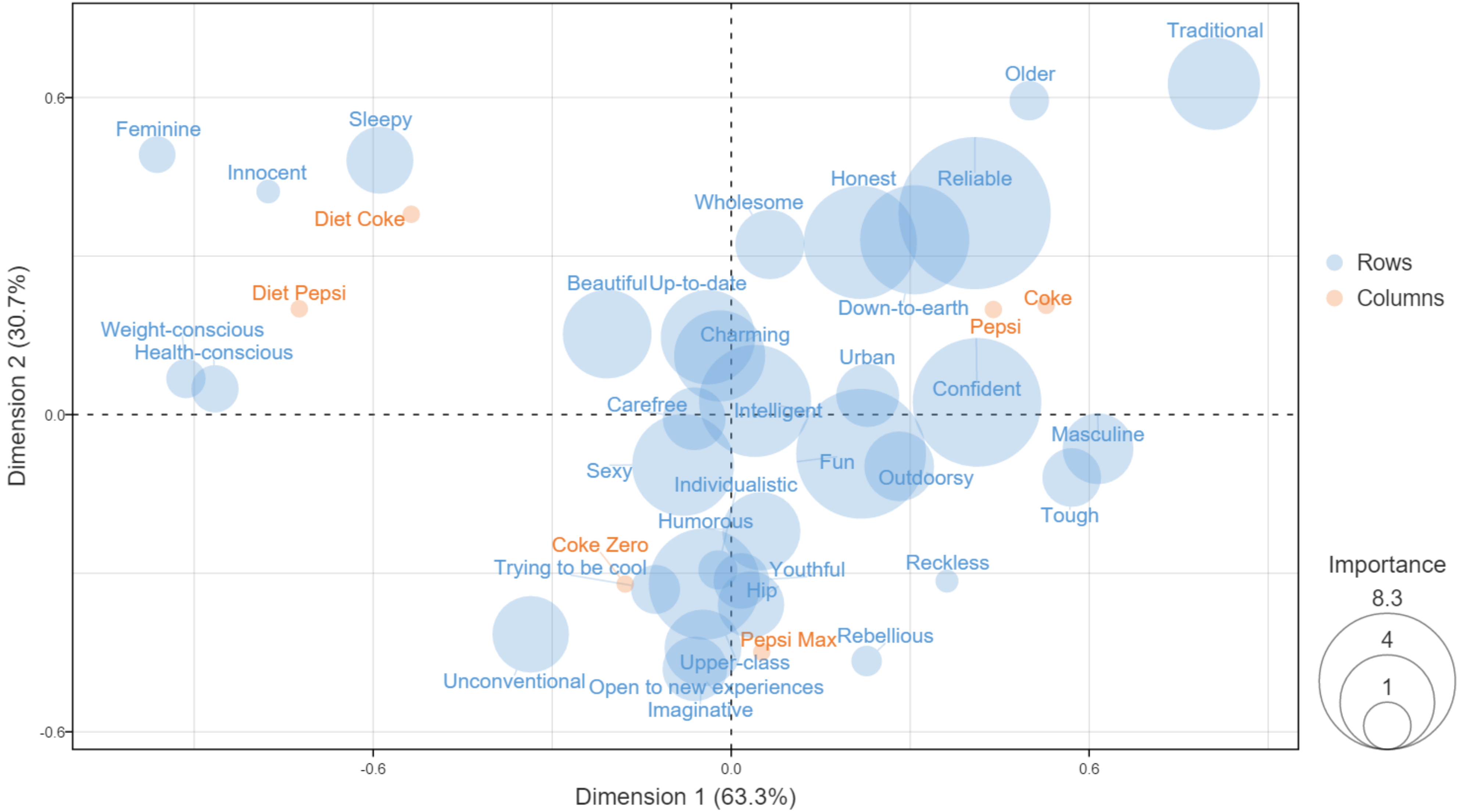
Example output: Importance scores



Example output: Performance-Importance Chart (aka Quad Chart)



Example output: Correspondence Analysis with Importance



RESEARCHSOFTWARE
A DIVISION OF DISPLAYR

TIM BOCK PRESENTS

Q&A Session