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

<table>
<thead>
<tr>
<th>Likelihood to recommend</th>
<th>This brand is fun</th>
<th>This brand is exciting</th>
<th>This brand is youthful</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Case study 1: Cola brand attitude

<table>
<thead>
<tr>
<th>Outcome variable(s)</th>
<th>34 Predictor variable(s)</th>
<th>If the brand was a person, what would its personality be?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hate/Dislike/Neither/Like/Love/Don’t know: Coke Zero Coke Diet Coke Diet Pepsi Pepsi Max Pepsi</td>
<td>Brand associations: Beautiful Carefree Charming Confident Down-to-earth Feminine Fun Health-conscious Hip Honest Humorous</td>
<td>Imaginative Individualistic Innocent Intelligent Masculine Older Open to new experiences Outdoorsy Rebellious Reckless Reliable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sexy Sleepy Tough Traditional Trying to be cool Unconventional Up-to-date Upper-class Urban Weight-conscious Wholesome Youthful</td>
</tr>
</tbody>
</table>
## Case study 2 (time permitting): Technology

### Outcome variable(s)
- Likelihood to recommend:
  - Apple
  - Microsoft
  - IBM
  - Google
  - Intel
  - Hewlett-Packard
  - Sony
  - Dell
  - Yahoo
  - Nokia
  - Samsung
  - LG
  - Panasonic

### Predictor variable(s)
- 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

<table>
<thead>
<tr>
<th>ID</th>
<th>Brand</th>
<th>Likelihood to recommend</th>
<th>This brand is fun</th>
<th>This brand is exciting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apple</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Microsoft</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>IBM</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Apple</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Microsoft</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>IBM</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Apple</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Microsoft</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>IBM</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Apple</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Microsoft</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>IBM</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
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)
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 weeks, 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
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)

<table>
<thead>
<tr>
<th>Options (ranked from best to worst)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take all the relevant theory into account when interpreting the results.</td>
<td>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).</td>
</tr>
<tr>
<td>Use Shapley or Relative Importance Analysis.</td>
<td>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.</td>
</tr>
</tbody>
</table>

**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.
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:
```r
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

<table>
<thead>
<tr>
<th>Options (not mutually exclusive)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set Don’t Knows to missing</td>
<td></td>
</tr>
</tbody>
</table>
| 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. |

### 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.
## The outcome variable is numeric (if using Shapley)

<table>
<thead>
<tr>
<th>Options (ranked from best to worst)</th>
<th>Comments</th>
</tr>
</thead>
</table>
| 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. |

### 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).
5: The predictor variables are numeric or binary

<table>
<thead>
<tr>
<th>Options (not mutually exclusive)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set Don’t Knows to missing</td>
<td>This can be problematic as the variables as the missing values may not be missing at random. This is discussed later.</td>
</tr>
</tbody>
</table>
| 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) | 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. |
| 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. | |
6: People do not differ in their needs/wants (segmentation)

<table>
<thead>
<tr>
<th>Issue</th>
<th>How to test</th>
</tr>
</thead>
</table>
| Traditional driver analysis techniques assume that people have the same needs/wants, and apply these consistently from situation to situation. | • Compare by brand  
• Compare by other data  
• Latent class analysis |

<table>
<thead>
<tr>
<th>Options (not mutually exclusive)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate an appropriate bespoke model (e.g., latent class analysis) and then estimate the driver analysis models within each segment</td>
<td>In Q: In a non-stacked data file, set up the data as an <strong>Experiment</strong>, and use <strong>Create &gt; Segment &gt; Latent Class Analysis</strong></td>
</tr>
<tr>
<td>Form segments by judgment, and estimate separate relative importance analyses for each segment.</td>
<td></td>
</tr>
</tbody>
</table>
### 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.

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

* 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).

![Diagram showing assumed predictor variables and their relationship to outcome 1.](image-url)

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.
Example causality problem: Backdoor path

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

<table>
<thead>
<tr>
<th>Issue</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>When people interpret importance scores, they assume that higher means better. This assumption is not always right.</td>
<td>Correlate each predictor variable with the outcome variable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Options (ranked from best to worst)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigate the data to make sense of the unexpected relationships.</td>
</tr>
<tr>
<td>Remove problematic variables from the analysis.</td>
</tr>
</tbody>
</table>
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

**Options (ranked from best to worst)**

<table>
<thead>
<tr>
<th>Options</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a bespoke model that appropriately models the process(es) that cause the values to be missing.</td>
<td>This is really hard!</td>
</tr>
<tr>
<td>Multiple imputation of missing values</td>
<td>If using <em>Relative Importance Analysis</em>, set <strong>Missing Data</strong> to <strong>Multiple Imputation</strong></td>
</tr>
<tr>
<td>Leave out observations with missing values from the analysis (i.e., <em>complete case analysis</em>)</td>
<td>This implicitly assumes that the data is <strong>Missing Completely At Random</strong> (<em>MCAR</em>; 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 <strong>Automate &gt; Browse Online Library &gt; Missing Data &gt; Little’s MCAR Test</strong></td>
</tr>
</tbody>
</table>

**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)
### 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)

<table>
<thead>
<tr>
<th>Options</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inspect each unusual observation, and understand if it is an error or not</td>
<td>Difficult/time consuming</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Ignore the problem</td>
<td>This is, by far, the most common approach.</td>
</tr>
</tbody>
</table>
12: There is no serial correlation (aka autocorrelation)

<table>
<thead>
<tr>
<th>Options (ranked from best to worst)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>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)</td>
<td>This is a lot of work.</td>
</tr>
<tr>
<td>Don’t report statistical test results (i.e., <em>p</em>-values).</td>
<td>The importance scores will be OK. The significance tests will be misleading to an unknown extent.</td>
</tr>
</tbody>
</table>

**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)
**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: Performance-Importance Chart (aka Quad Chart)
Example output: Correspondence Analysis with Importance