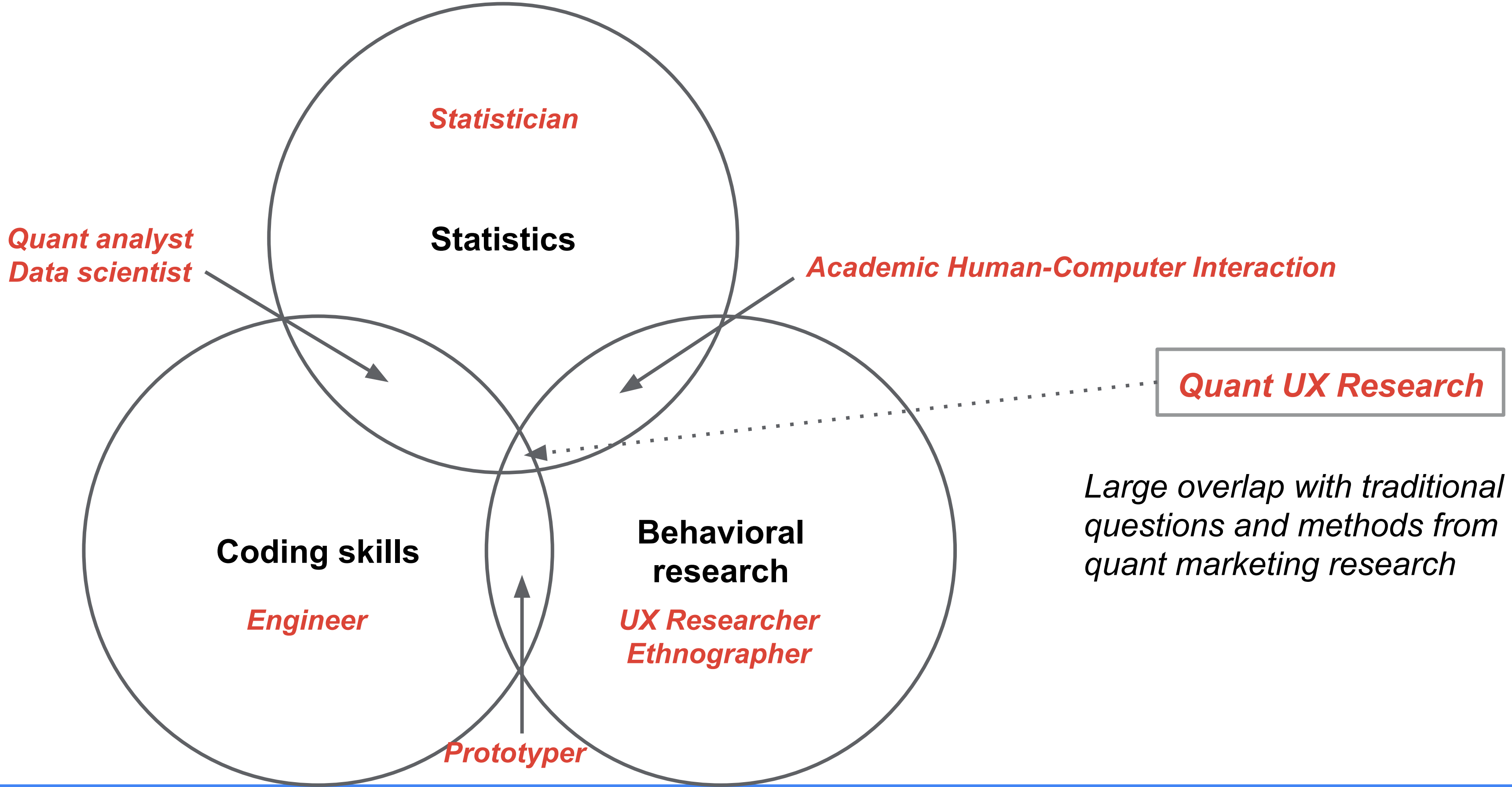
R for Marketing Research and Analytics: Motivation & Brief Tour

Chris Chapman, Google
Elea McDonnell Feit, Drexel University

What Chris does: "Quantitative User Experience Research"

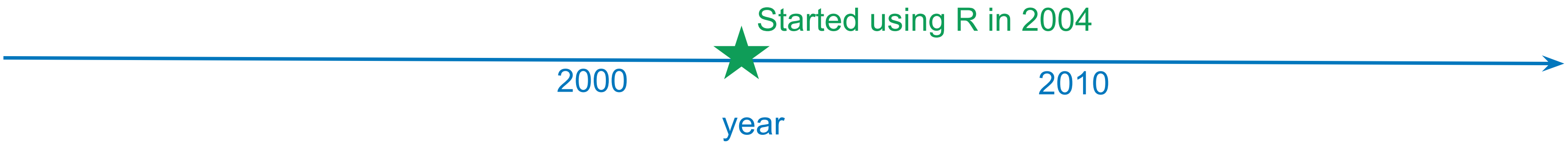


Large overlap with traditional questions and methods from quant marketing research

with similarities to Drew Conway's data science Venn diagram, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What Elea does

I develop statistical methods	choice modeling	ad stock models and experiments
to understand customers	potential buyers	customers
so that companies	GM	retailers
can make better decisions	vehicle designs	designing catalogs
and I teach those methods to others.	colleagues	company partners & students



You may also like to know that Elea is a Bayesian and is working on a second book titled "Business Experiments."

Observations on the state of quant methods in marketing

Stats depth
Essential for analytics, predictive modeling, experimentation

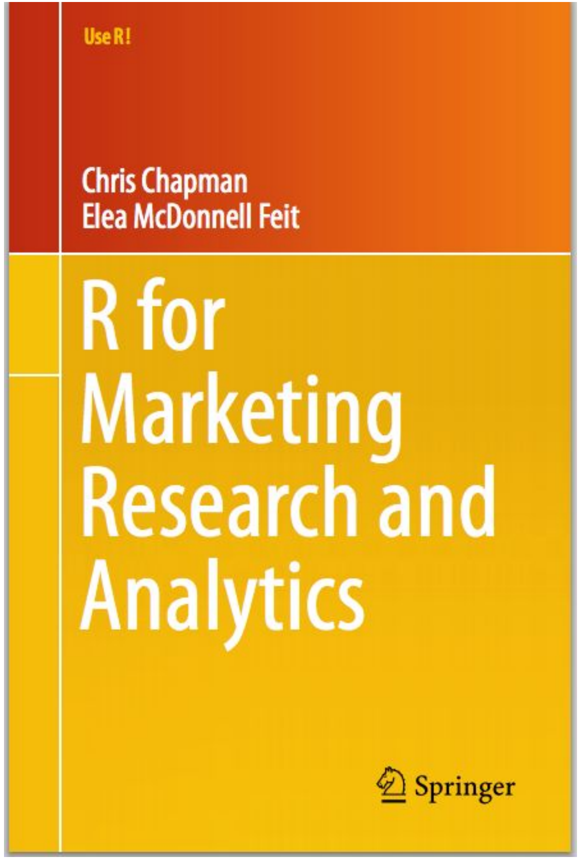
Stats breadth
Needed for customer insight, rapid feedback, strategy impact

Implications

1. Too many models and applications to expect expertise in any one analyst
2. Analysts often recreate the wheel because of siloed knowledge

To date, there have been few references describing a breadth of marketing methods for general researchers and statisticians

The obligatory book photo



Chapter	Key topics
	General
1-3	Basic R
4-6	Descriptives and ANOVA
7	Linear models
	Focused on marketing
8	EFA, PCA, and perceptual mapping
9	Hierarchical linear models
10	CFA and structural equation models
11	Segmentation (clustering and classification)
12	Association rules (market basket analysis)
13	Choice models (conjoint analysis)

Why those methods?

Chapter	Key topics	
	General	
1-3	Basic R	
4-6	Descriptives and ANOVA	
7	Linear models	
	Method	***Common marketing application***
8	EFA, PCA, MDS	Assess brand/product positioning for strategy
9	HLM	Individual- or subgroup- level assessment
10	CFA, SEM	Survey validation; Estimates given many IVs & DVs
11	Cluster/classify	Market & customer insight, profiling, prediction
12	Association rules	Retail optimization, consumer targeting
13	Choice models	Feature prioritization, pricing, product portfolio design

Topics we'll describe in a bit more depth

Chapter	Key topics	
	General	
1-3	Basic R	
4-6	Descriptives and ANOVA	
7	Linear models	
	Method	***Common marketing application***
8	EFA, PCA, MDS	Assess brand/product positioning for strategy
9	HLM	Individual- or subgroup- level assessment
10	**CFA, SEM**	**Survey validation; Estimates given many IVs & DVs**
11	Cluster/classify	Market & customer insight, profiling, prediction
12	Association rules	Retail optimization, consumer targeting
13	**Choice models**	**Feature prioritization, pricing, product portfolio design**

Quick SEM in R

SEM: Why?

Consider survey asking about satisfaction

Customers are asked scaled items for:

- Sat with the product
- Sat with the salesperson
- Likelihood to recommend product
- Likelihood to recommend salesperson

… and the business wants to know:

How is Sat related to Recommend?

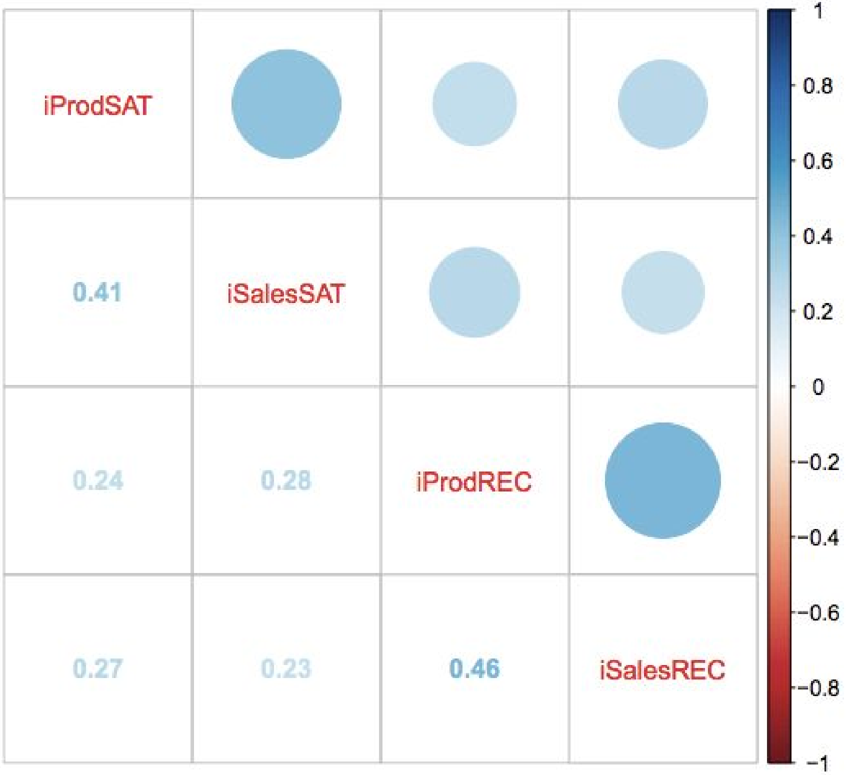
Problem: **the variables are all highly correlated**

Consider survey asking about satisfaction

Customers are asked scaled items for:
- Sat with the product
- Sat with the salesperson
- Likelihood to recommend product
- Likelihood to recommend salesperson

… and the business wants to know:

How is Sat related to Recommend?

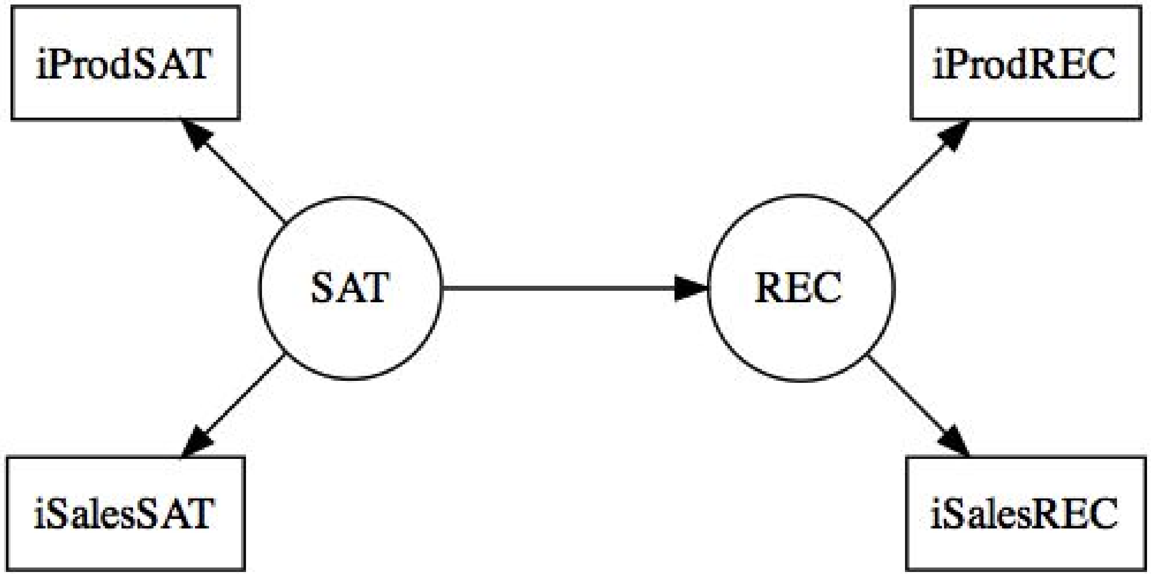


One latent model we might wish to estimate

Sat and Recommend are *latent constructs* with multiple observed variables

There are various ways to deal with collinearity and latent variables

SEM addresses the business question, estimating how *SAT* affects *REC* directly

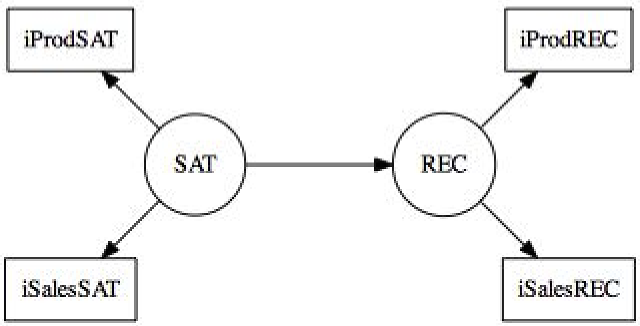


R code: Load the data and set up model (1)

```
# load data
> satData <- read.csv("http://goo.gl/UDv12g")
> head(satData)

  iProdSAT iSalesSAT Segment iProdREC iSalesREC
1        6         2       1        4         3
2        4         5       3        4         4
3        5         3       4        5         4
```

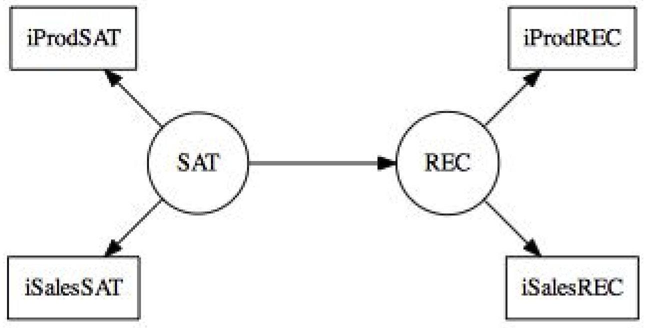
R code: Load the data and set up model (2)

```
# load data
> satData <- read.csv("http://goo.gl/UDv12g")
> head(satData)

  iProdSAT iSalesSAT Segment iProdREC iSalesREC
1        6         2       1        4         3
2        4         5       3        4         4
3        5         3       4        5         4


# set up manifest and LATENT variables
> satModel <- "SAT =~ iProdSAT + iSalesSAT
+              REC =~ iProdREC + iSalesREC
+              REC ~  SAT "
```

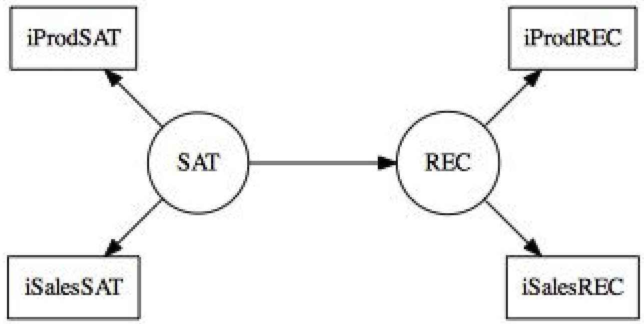


Estimate the SEM (1)

```
> satModel <- "SAT =~ iProdSAT + iSalesSAT
+              REC =~ iProdREC + iSalesREC
+              REC ~  SAT "

# estimate the model
> library(lavaan)
> sat.fit <- cfa(satModel, data=satData)
```



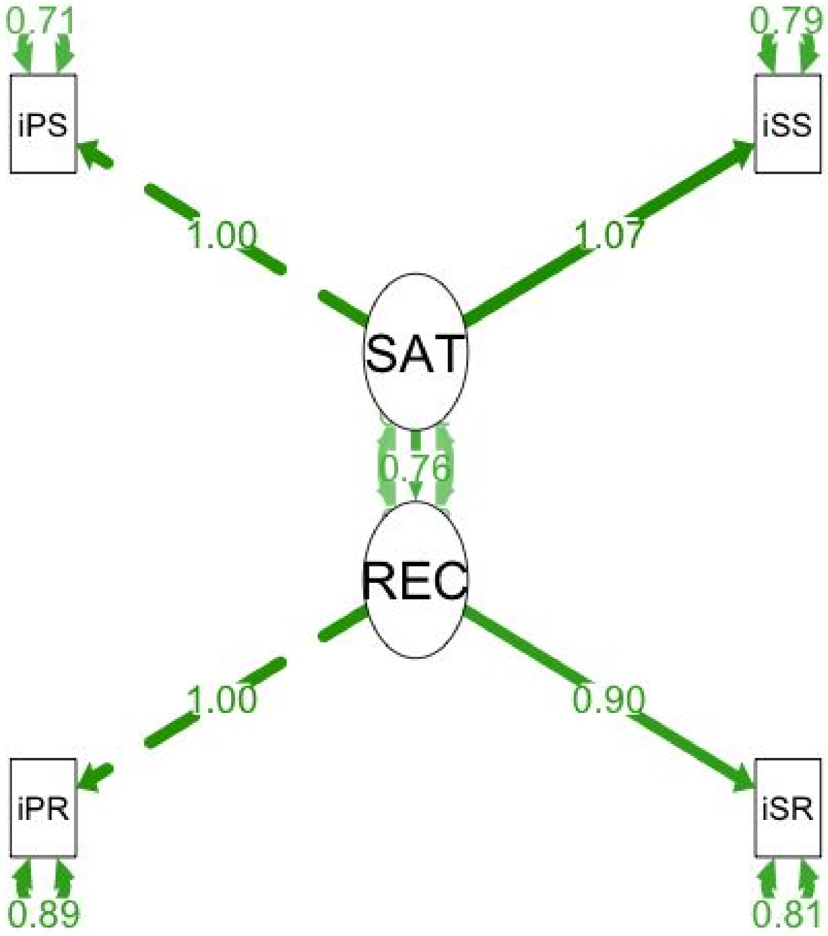
Estimate the SEM (2)




```
> satModel <- "SAT =~ iProdSAT + iSalesSAT
+              REC =~ iProdREC + iSalesREC
+              REC ~  SAT "

# estimate the model
> library(lavaan)
> sat.fit <- cfa(satModel, data=satData)

# inspect it
> summary(sat.fit, fit.m=TRUE)
User model versus baseline model:
  Comparative Fit Index (CFI)                    0.995
…
Regressions:
  REC ~ SAT           0.758    0.131    5.804    0.000
```

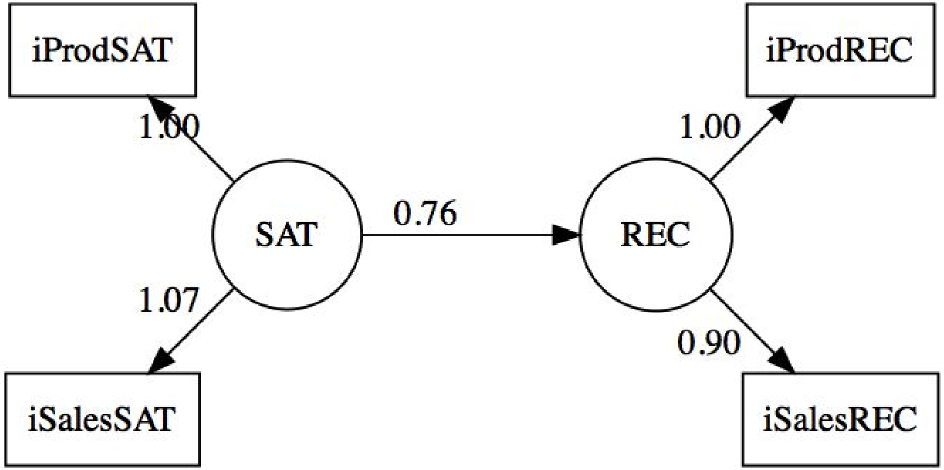
Plot it

```
# plot it
> library(semPlot)
> semPaths(sat.fit, what="est",
+            edge.label.cex=1)
```



Or a cleaner plot with DiagrammeR

```
library(DiagrammeR)

grViz("
      digraph SEM {
        graph [layout = neato, overlap = true,
               outputorder = edgesfirst]
        node [shape = rectangle]

        a [pos='-2, 1!', label='iProdSAT']
        b [pos='-2,-1!', label='iSalesSAT']
        c [pos='-1, 0!', label='SAT', shape=circle]
        d [pos=' 1, 0!', label='REC', shape=circle]
        e [pos=' 2, 1!', label='iProdREC']
        f [pos=' 2,-1!', label='iSalesREC']

        c->a [label='1.00']
        c->b [label='1.07']
        c->d [label='0.76']
        d->e [label='1.00']
        d->f [label='0.90']
      } ")
```



based on example by Rich Iannone, April 2015, http://stackoverflow.com/questions/29208525/using-diagrammer-for-path-diagrams-sem

R code: complete SEM

```
# set up manifest and LATENT variables
satModel <- "SAT =~ iProdSAT + iSalesSAT
             REC =~ iProdREC + iSalesREC
             REC ~  SAT "

# estimate the model
library(lavaan)
sat.fit <- cfa(satModel, data=satData)

# inspect it
summary(sat.fit, fit.m=TRUE)

# plot it
library(semPlot)
semPaths(sat.fit, what="est")
```

SEM: Did we answer the question?

Customers are asked scaled items for:

Sat with the product
Sat with the salesperson
Likelihood to recommend product
Likelihood to recommend salesperson

… and the business wants to know:

How is Satisfaction related to Recommending?
⇒ *Recommend* goes up 0.76 units for each unit of latent *Satisfaction*
⇒ This is stronger than any single effect in the raw, bivariate correlations

Quick Choice Models in R

Choice Modeling: Why?

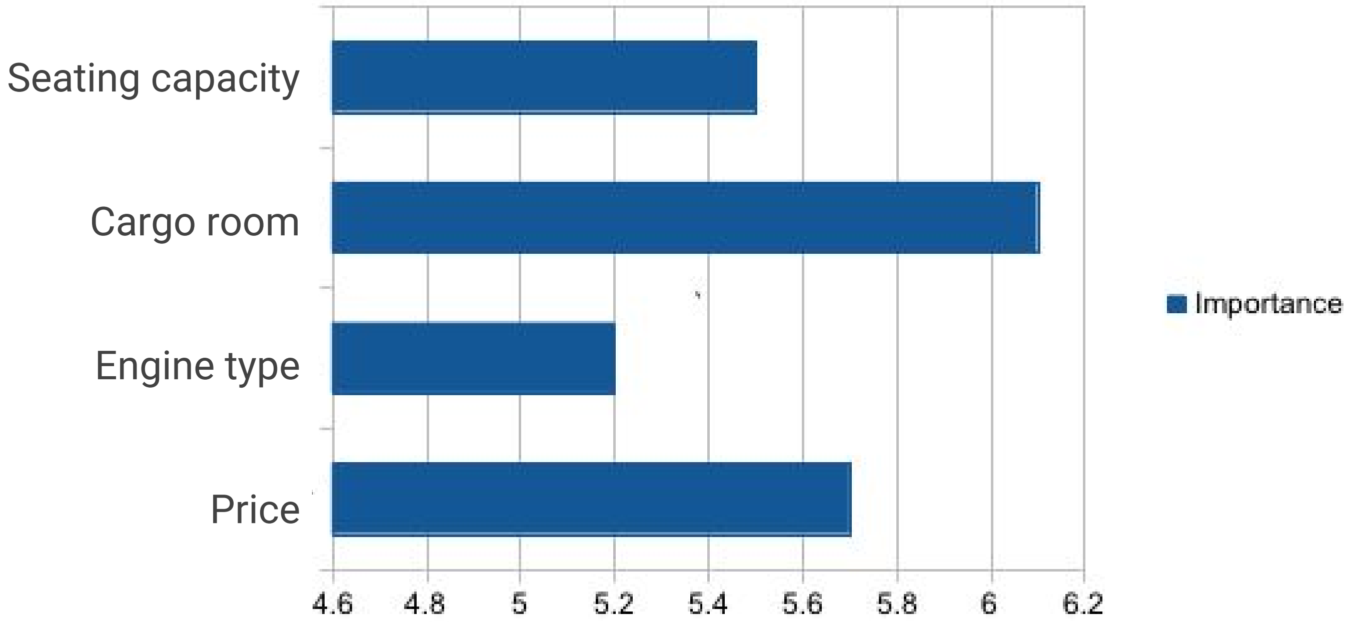
Traditional scaled responses rarely give good answers

Typical survey approach:

How important is each auto feature for you?
(check an answer for each feature)

	Not important						*Very important*
Seating capacity	1	2	3	4	5	6	7
Cargo room	1	2	3	4	5	6	7
Engine type	1	2	3	4	5	6	7
Price	1	2	3	4	5	6	7

Mean consumer ratings of auto attributes (fictional)



Unclear interpretation … "How many people would buy our product if we do X or Y?"

Better is to give respondents a more natural task

Which of the following minivans would you buy?
Assume all three minivans are identical other than the features listed below.

	Option 1	Option 2	Option 3
	6 passengers 2 ft. cargo area gas engine $35,000	8 passengers 3 ft. cargo area hybrid engine $30,000	6 passengers 3 ft. cargo area gas engine $30,000
I prefer (check one):	☐	☐	☑

Consumers give meaningful answers, and we can model choice likelihood by feature

The model

Multinomial logit model, aka conditional logit model

Estimates the *part-worth value* (***utility***) for each ***feature***, for each ***respondent***

Utility of respondent ***i*** for product ***j*** η_{ij}

Total utility of **all products** under consideration (set *k*) $\sum \exp\{\eta_{ik}\}$

Likelihood to choose *j* (π_{ij}) is the ratio of exponentiated utility share for product ***j*** vs. **all products**

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum \exp\{\eta_{ik}\}}$$

formulas adapted from G. Rodriguez, http://data.princeton.edu/wws509/notes/c6s3.html

Choice data

```
> cbc.df <- read.csv("http://goo.gl/5xQObB",
+                      colClasses = c(seat = "factor", price = "factor"))
```

Option 1	Option 2	Option 3
6 passengers	8 passengers	6 passengers
2 ft. cargo area	3 ft. cargo area	3 ft. cargo area
gas engine	hybrid engine	gas engine
$35,000	$30,000	$30,000
☐	☐	☑

For Question 1, Respondent 1 saw 3 products, and chose #3

```
> head(cbc.df)
  resp.id ques alt carpool seat cargo  eng price choice
1       1    1   1     yes    6   2ft  gas    35      0
2       1    1   2     yes    8   3ft  hyb    30      0
3       1    1   3     yes    6   3ft  gas    30      1
-----------------------------------------------------------
4       1    2   1     yes    6   2ft  gas    30      0
```

Estimation using `mlogit`

```
> library(mlogit)
> cbc.mlogit <- mlogit.data(data=cbc.df, choice="choice", shape="long",
+                           varying=3:6, alt.levels=paste("pos",1:3),
+                           id.var="resp.id")

> m1 <- mlogit(choice ~ 0 + seat + cargo + eng + price, data = cbc.mlogit)
> summary(m1)

          Estimate Std. Error  t-value  Pr(>|t|)
seat7    -0.535280   0.062360  -8.5837 < 2.2e-16 ***
seat8    -0.305840   0.061129  -5.0032 5.638e-07 ***
cargo3ft  0.477449   0.050888   9.3824 < 2.2e-16 ***
enggas    1.530762   0.067456  22.6926 < 2.2e-16 ***
enghyb    0.719479   0.065529  10.9796 < 2.2e-16 ***
price35  -0.913656   0.060601 -15.0765 < 2.2e-16 ***
price40  -1.725851   0.069631 -24.7856 < 2.2e-16 ***
```

The coefs are the aggregate (upper-level) part worth utilities for MNL

*(`mlogit` is one method. We more typically use a hierarchical Bayes model and estimate with **`bayesm`**)*

Predicting share preference

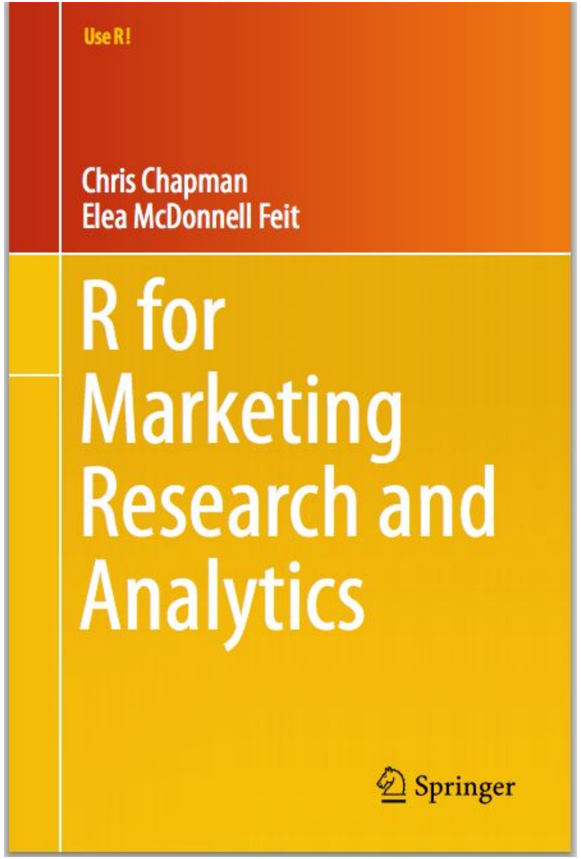
```
> predict.mnl <- function(model, data) {
+     data.model <- model.matrix(update(model$formula, 0 ~ .), data = data)[,
-1]
+     utility <- data.model %*% model$coef
+     share <- exp(utility)/sum(exp(utility))
+     cbind(share, data) }

> attrib <- list(seat = c("6", "7", "8"),  cargo = c("2ft", "3ft"),
+                eng = c("gas", "hyb", "elec"),  price = c("30", "35", "40"))

> new.data <- expand.grid(attrib)[c(8, 1, 3, 41, 49, 26), ]
> predict.mnl(m1, new.data)
        share seat cargo  eng price
8  0.44643895    7   2ft  hyb    30
1  0.16497955    6   2ft  gas    30
3  0.12150814    8   2ft  gas    30
41 0.02771959    7   3ft  gas    40
49 0.06030713    6   2ft elec    40
26 0.17904663    7   2ft  hyb    35
```

Basic MNL preference share estimate

*Many respondents prefer "**auto 8**" … but depending on what is available in market, autos **1**, **3**, or **26** could be good alternatives to produce*

A next step could be a hierarchical (mixed) model to examine individual differences and correlates

Finally



Chapter	**Key topics**
	General
1-3	Basic R
4-6	Descriptives and ANOVA
7	Linear models
	Focused on marketing
8	EFA, PCA, and perceptual mapping
9	Hierarchical linear models
10	CFA and structural equation models
11	Segmentation (clustering and classification)
12	Association rules (market basket analysis)
13	Choice models (conjoint analysis)

Contacts

Book site	Code and data Also classroom slides!	http://r-marketing.r-forge.r-project.org
Twitter	Chris Chapman Elea McDonnell Feit	@cnchapman @eleafeit
Email	Chris Chapman Elea McDonnell Feit	cnchapman+r@gmail.com emf75@drexel.edu ⇐ For Instructors

Thank you!