Methods for identifying causality

Seán Roberts

Overview

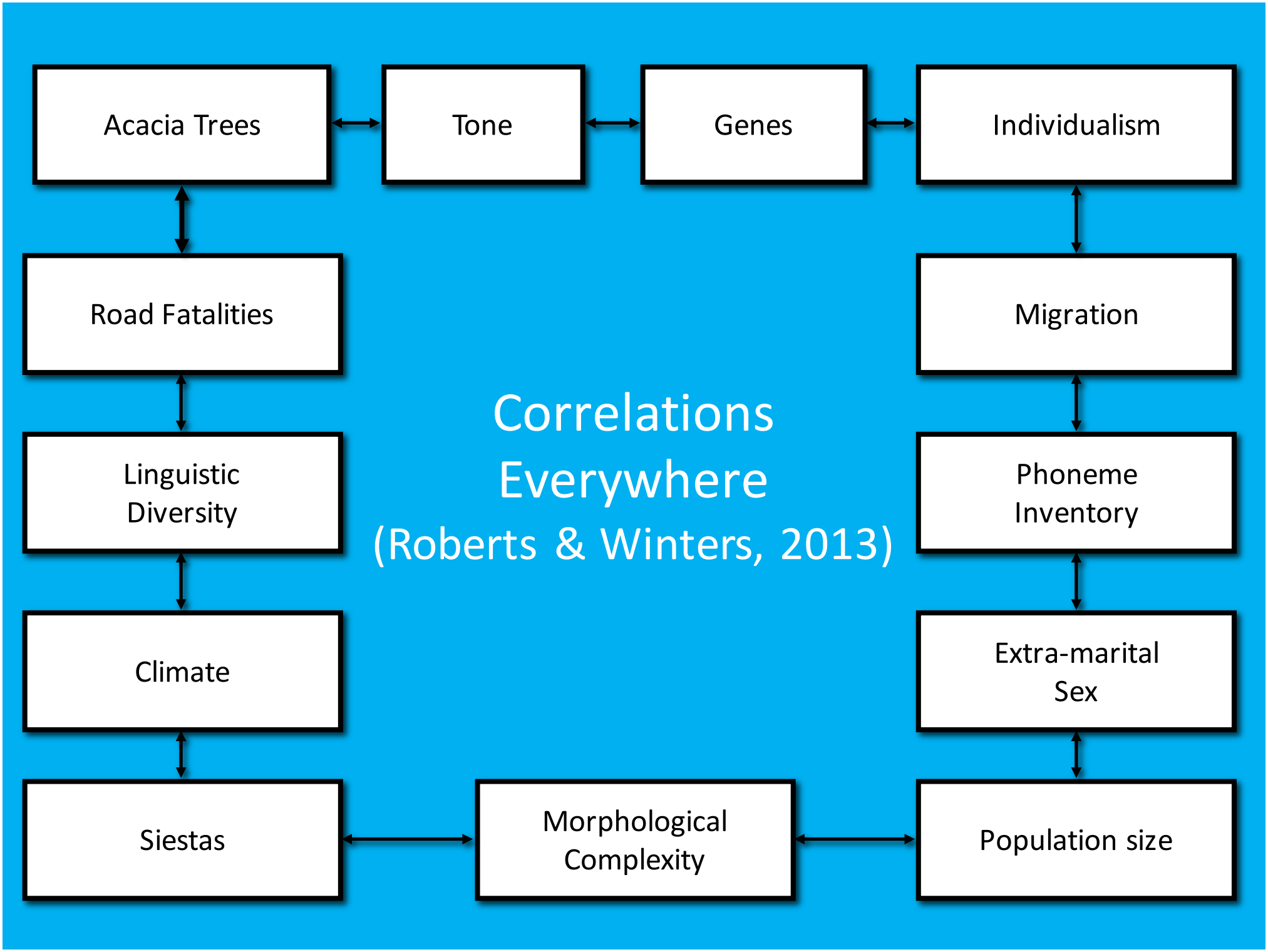
Correlation is not causation?

Correlation, in the absence of reasonable alternative explanations implies causation

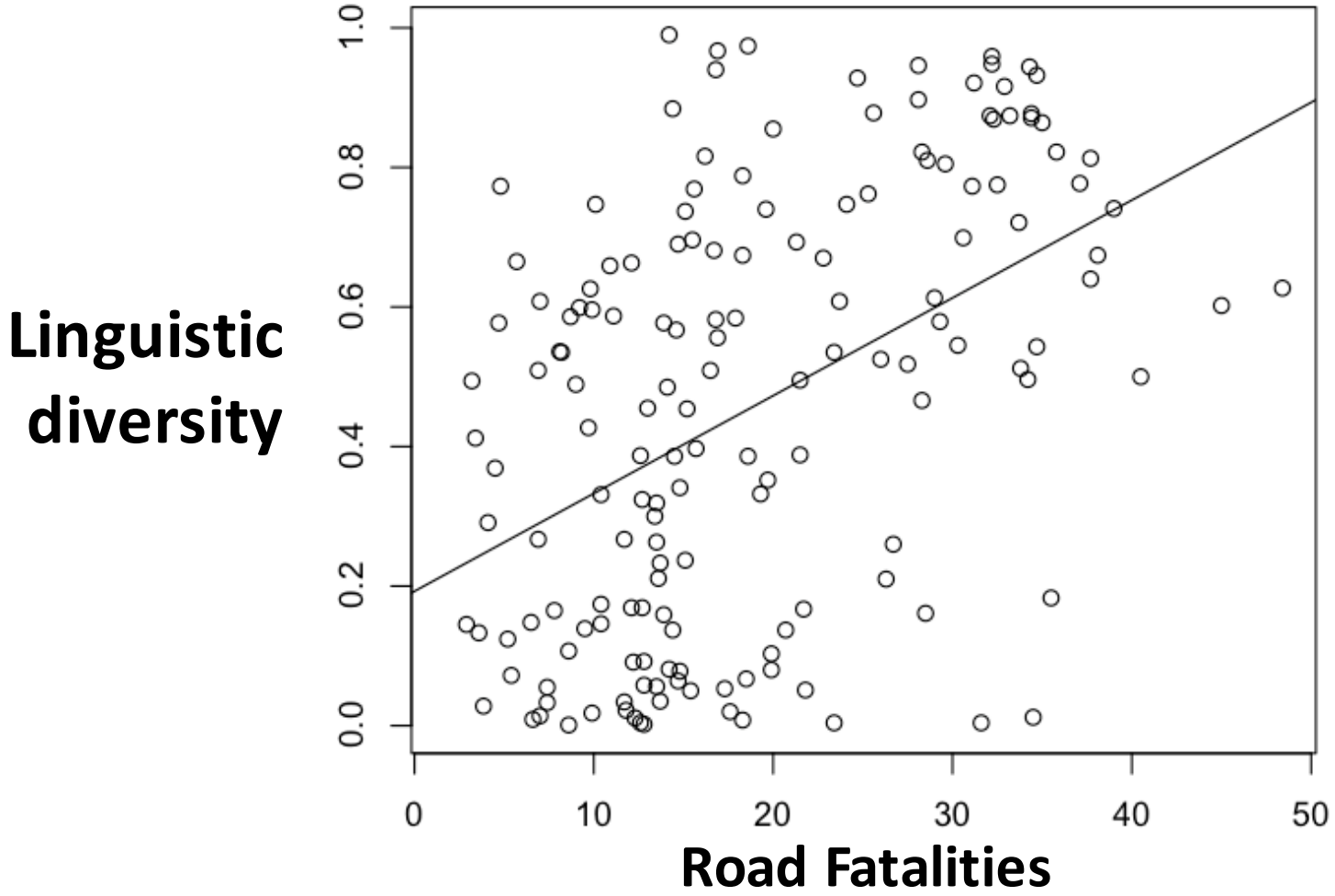
Today:

Deep Understanding ✗ **What are my options? ✔**

- Galton's problem
- Permutation
- Mixed effects modelling
- Decision trees
- Causal graphs


Acacia Trees
Tone
Genes
Individualism
Road Fatalities
Migration
Correlations Everywhere
(Roberts & Winters, 2013)
Linguistic Diversity
Phoneme Inventory
Climate
Extra-marital Sex
Siestas
Morphological Complexity
Population size

Galton's problem

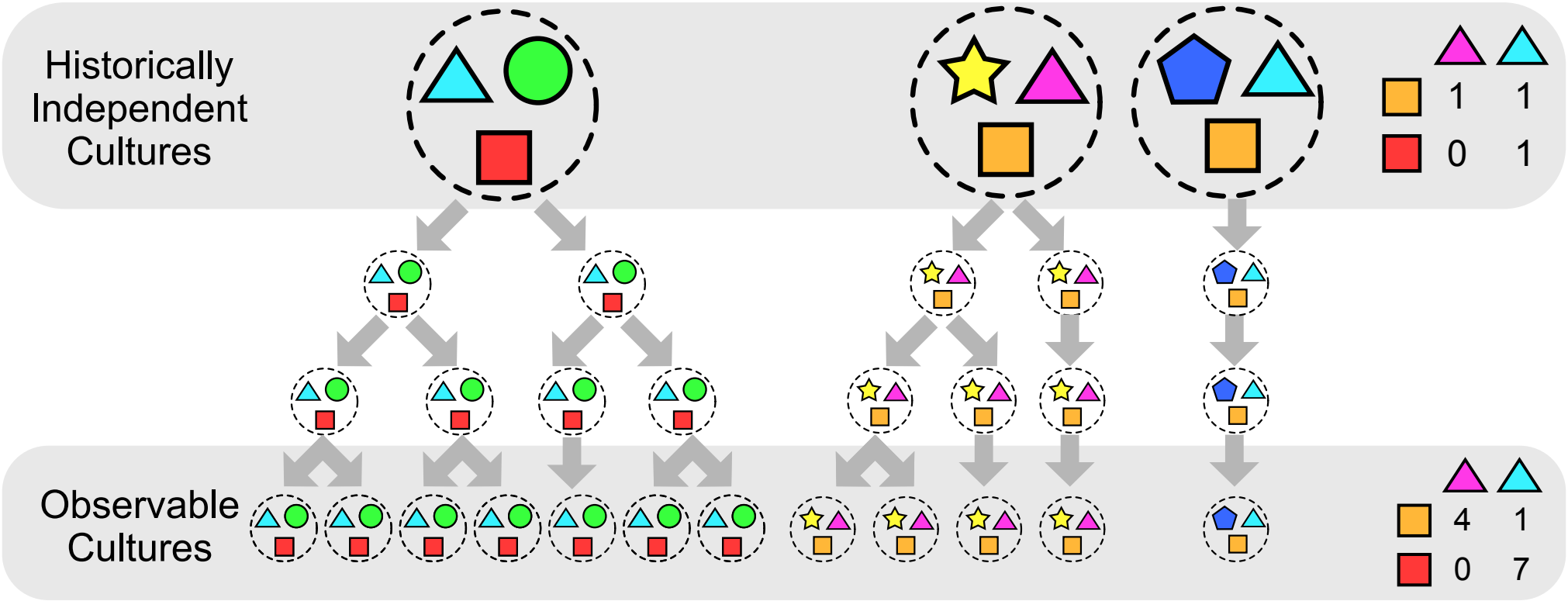


F (97,10) = 4.18, p<0.0001

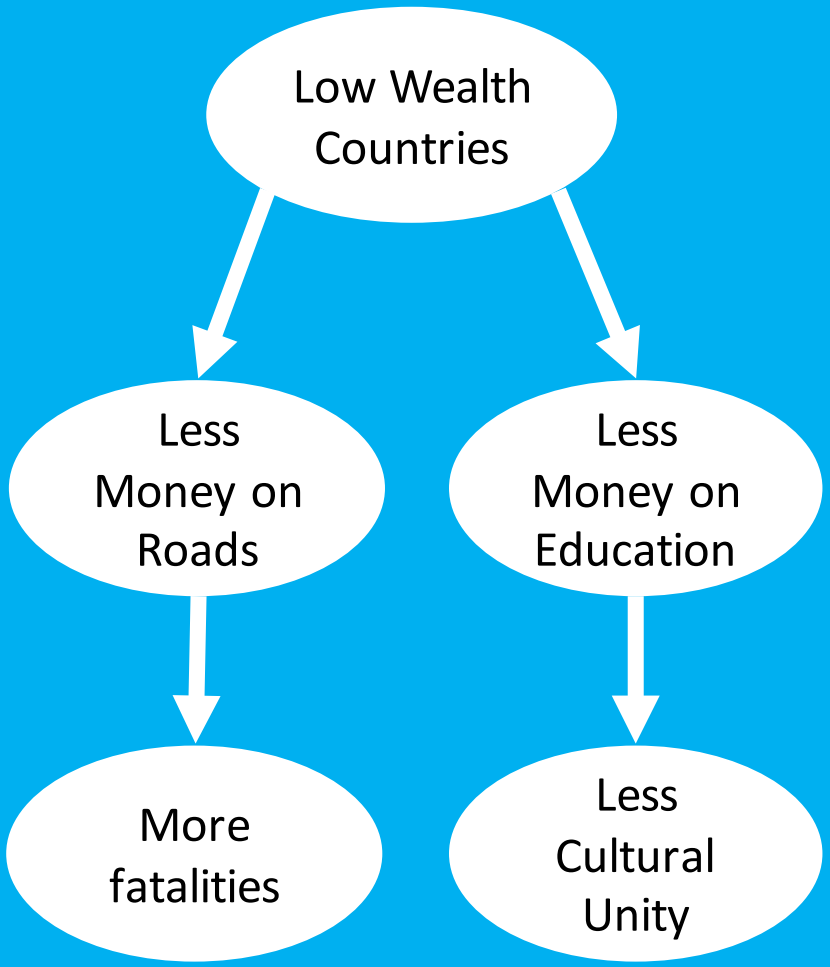
Controlling for:

- Per-capita GDP
- Country nominal GDP
- Population density
- Migration
- Inside / outside Africa
- Distance from the equator

Historical relationships inflate correlations



Correlations from common cause



Standard tests don’t always work

Statistical testing requires clear thinking:

- What is my prediction?
- What are the alternative explanations?
- What is the baseline for chance?

Permutation

Permutation

Are the patterns in the data stronger than chance?

Assumptions: Very few

Advantages

Flexible, adaptable to particular questions

Any measures can be used

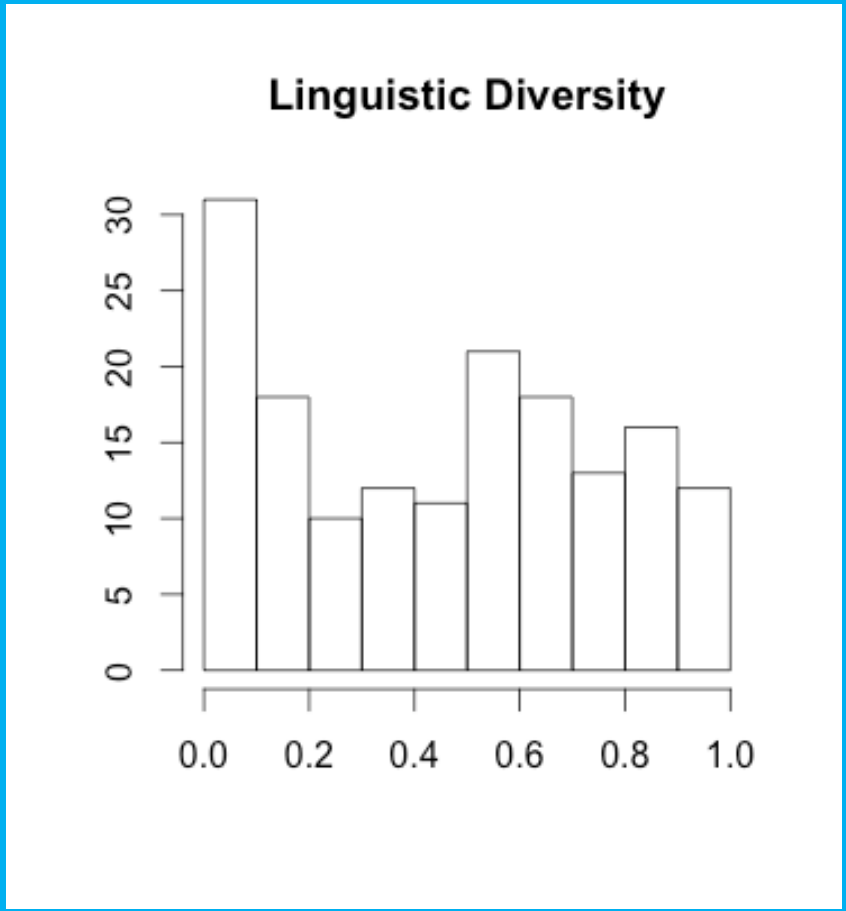
More details:

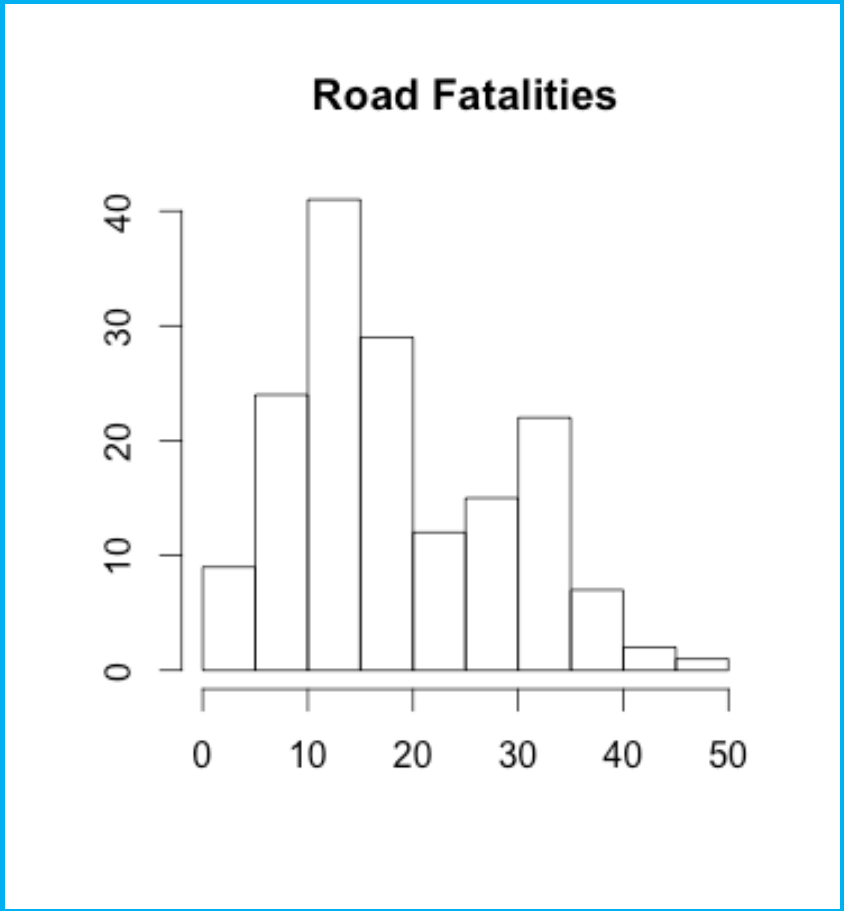
'Permutation Tests' tutorial in the QMSS Intro to R

Permutation

Traffic accidents and linguistic diversity:

Distributions are not normal!

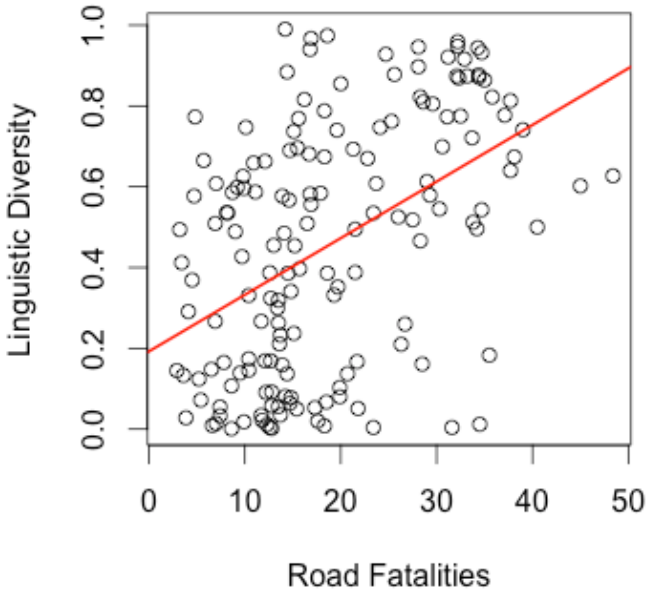




Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.663	Asia
Bangladesh	12.6	0.387	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.535	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.386	North America
Barbados	12.2	0.091	North America
Belize	15.6	0.769	North America
Australia	5.2	0.124	Oceania

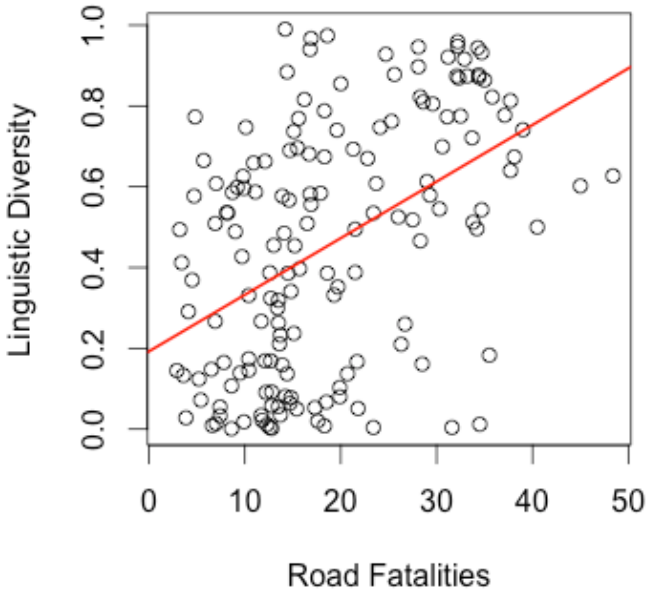
True Data



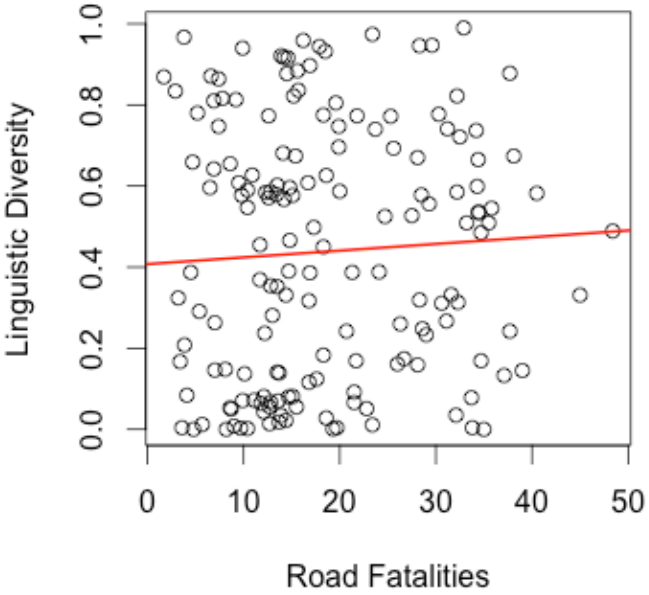
Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.535	Asia
Bangladesh	12.6	0.386	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.663	Europe
Azerbaijan	13	0.769	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.387	North America
Barbados	12.2	0.124	North America
Belize	15.6	0.455	North America
Australia	5.2	0.091	Oceania

True Data



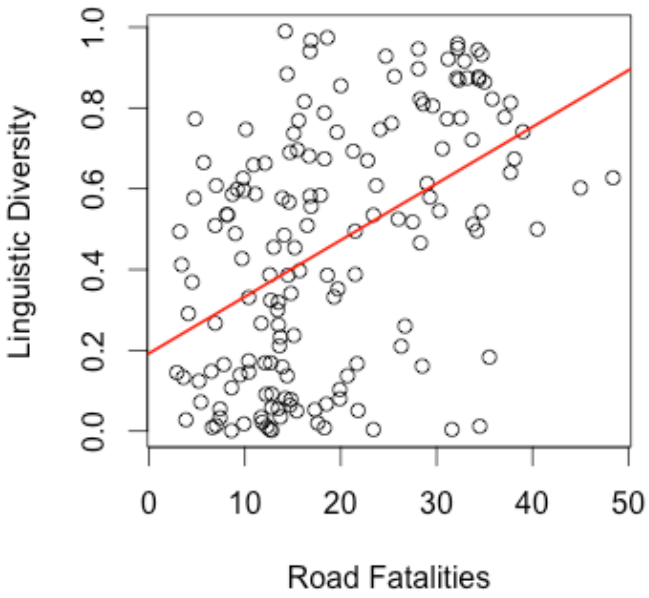
Permuted Data



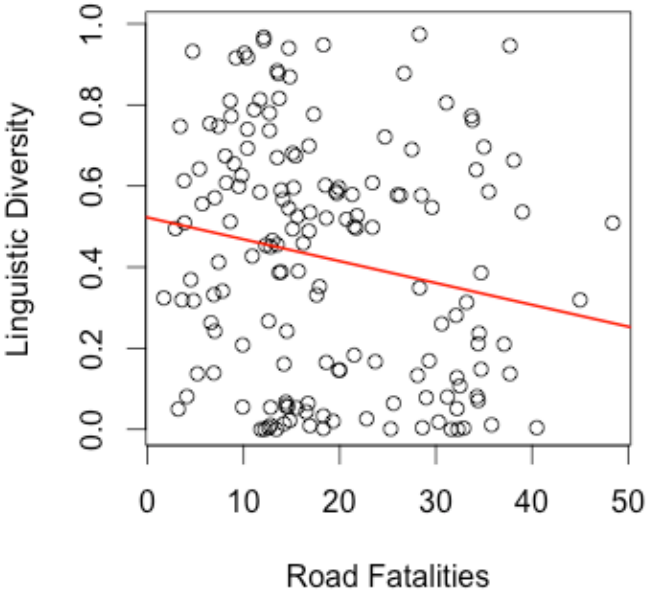
Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.387	Asia
Bangladesh	12.6	0.663	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.091	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.386	Europe
Bahamas	14.5	0.747	North America
Barbados	12.2	0.535	North America
Belize	15.6	0.124	North America
Australia	5.2	0.769	Oceania

True Data



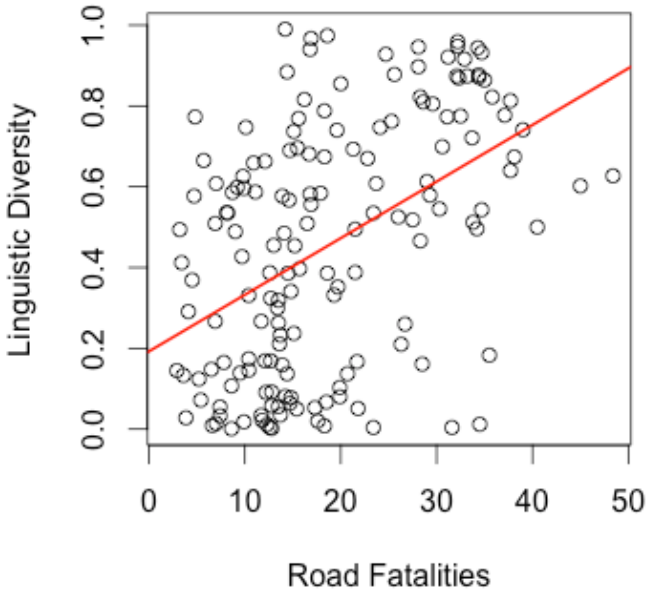
Permuted Data



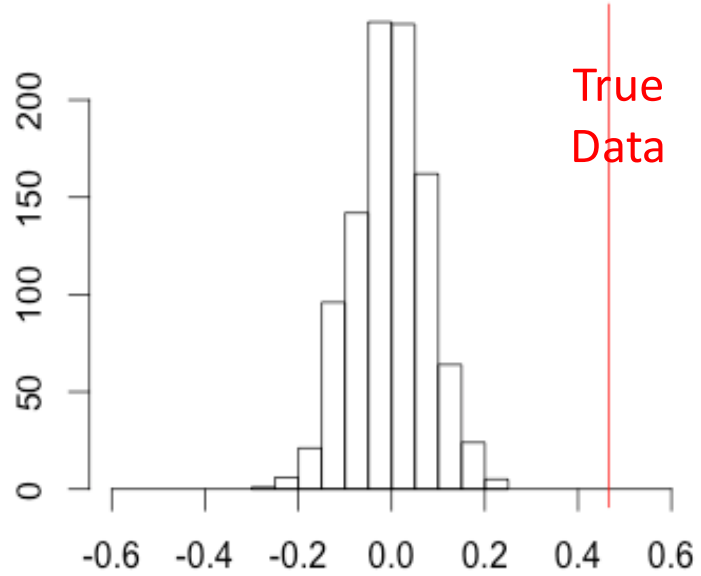
Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.663	Asia
Bangladesh	12.6	0.387	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.535	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.386	North America
Barbados	12.2	0.091	North America
Belize	15.6	0.769	North America
Australia	5.2	0.124	Oceania

True Data



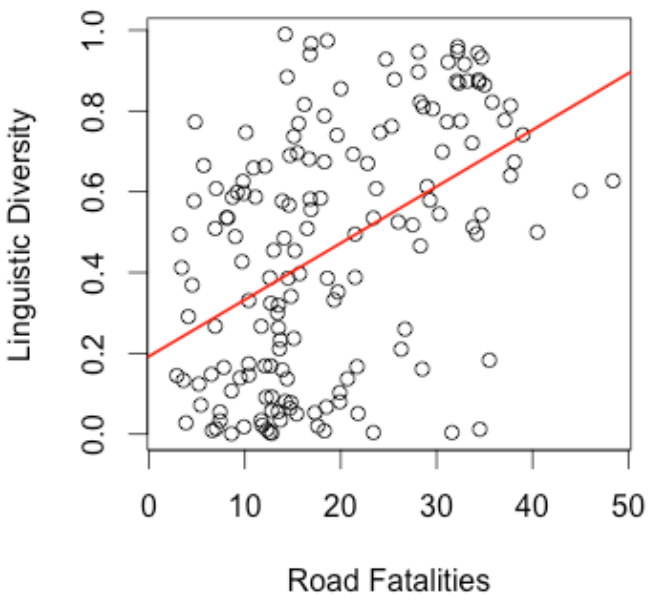
Permuted Data



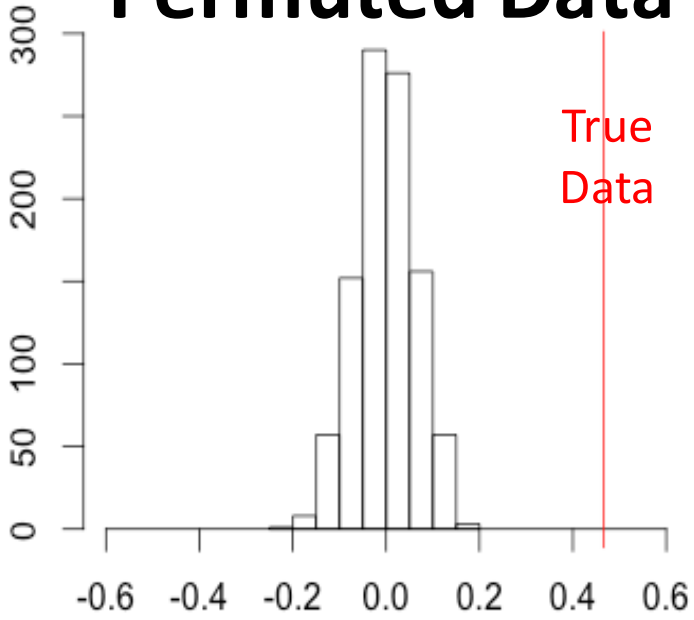
Stratified Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.663	Asia
Bangladesh	12.6	0.387	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.535	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.386	North America
Barbados	12.2	0.091	North America
Belize	15.6	0.769	North America
Australia	5.2	0.124	Oceania

True Data



Permuted Data



Mixed effects modelling

Mixed effects modelling

Are variables correlated, controlling for relationships?

Data Type: Continuous, Categorical

Assumptions: Linear correlations, normal distributions

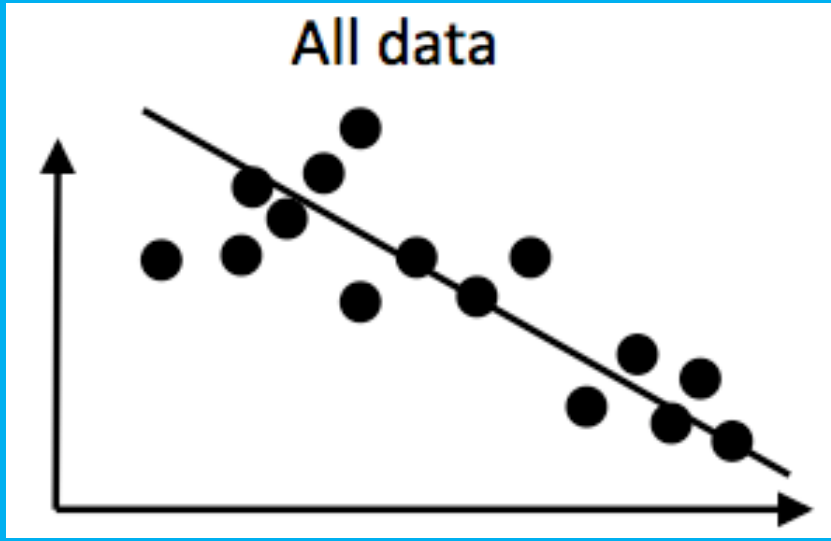
Advantages

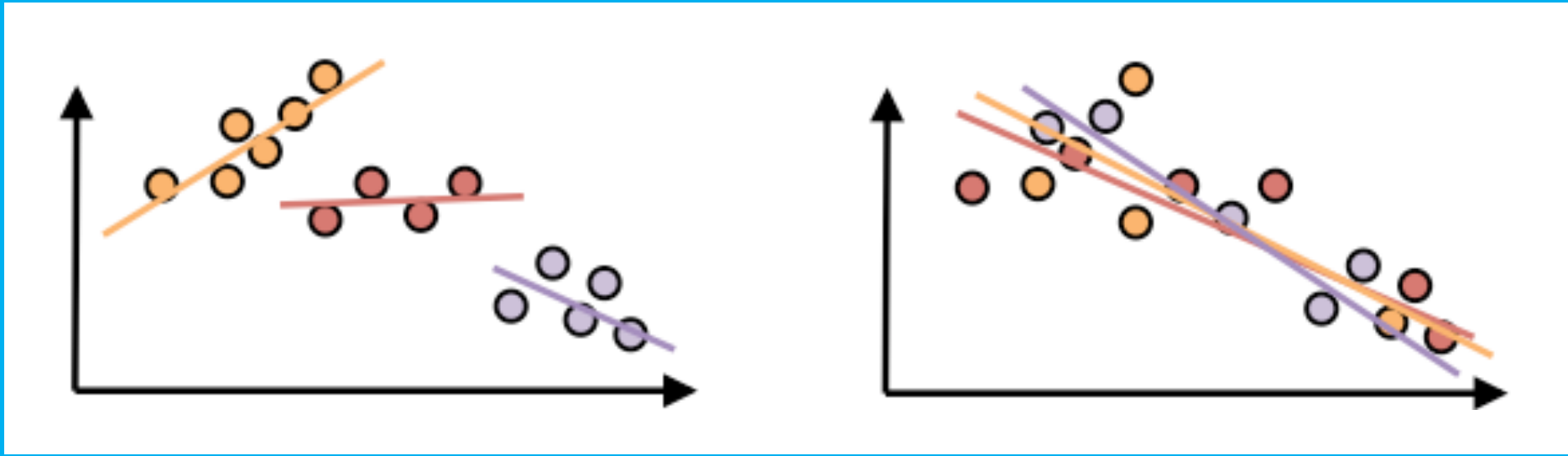
No need to average data – can use all datapoints (greater statistical power)

Controls for random effects, but doesn't require full tree

Can ask more detailed questions

Mixed effects modelling





Mixed effects modelling

Fixed effects
(individual measures)

Basic word order
Subsistance type
Frequency of word
Tenure norm

Random effects
(group membership)

Participant identity
Language family
Linguistic area
Clan / Group / Pack

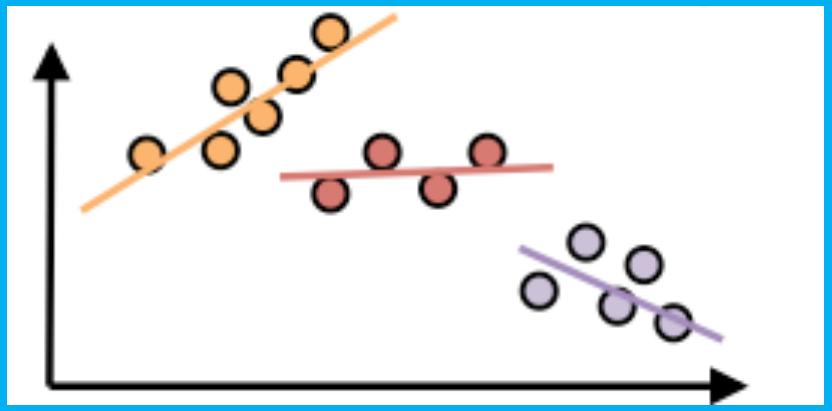
Controlling for continent

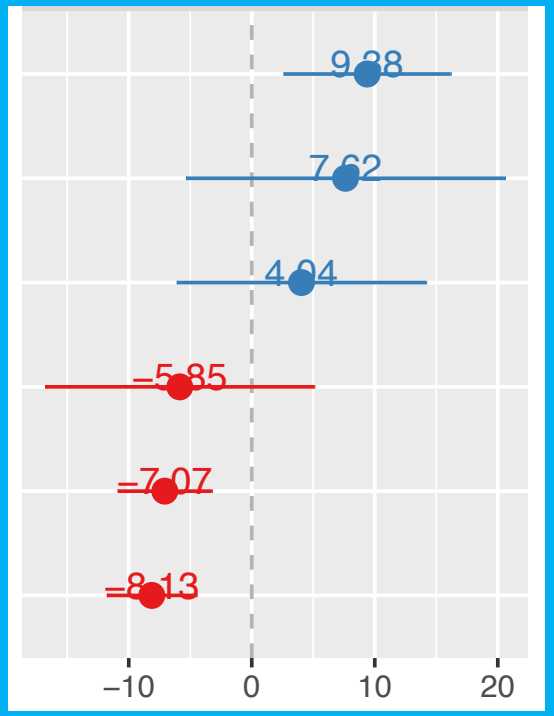
```
Road fatalities ~   Pop.Size + Ling.diversity
                  + (1 + Ling.diversity | continent)
```

Linguistic diversity not significant ($\chi^2 = 2.66$, $p = 0.11$)

Groups with the strongest effect?

Groups with opposite effect?





Effect of linguistic diversity on road fatalities

Mixed effects tutorial

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf]

http://www.bodo-winter.net/tutorials.html

Decision Trees

Decision trees

Find the most efficient set of questions for identifying clusters in the data

Data Type: Continuous, Categorical

Assumptions: No missing data

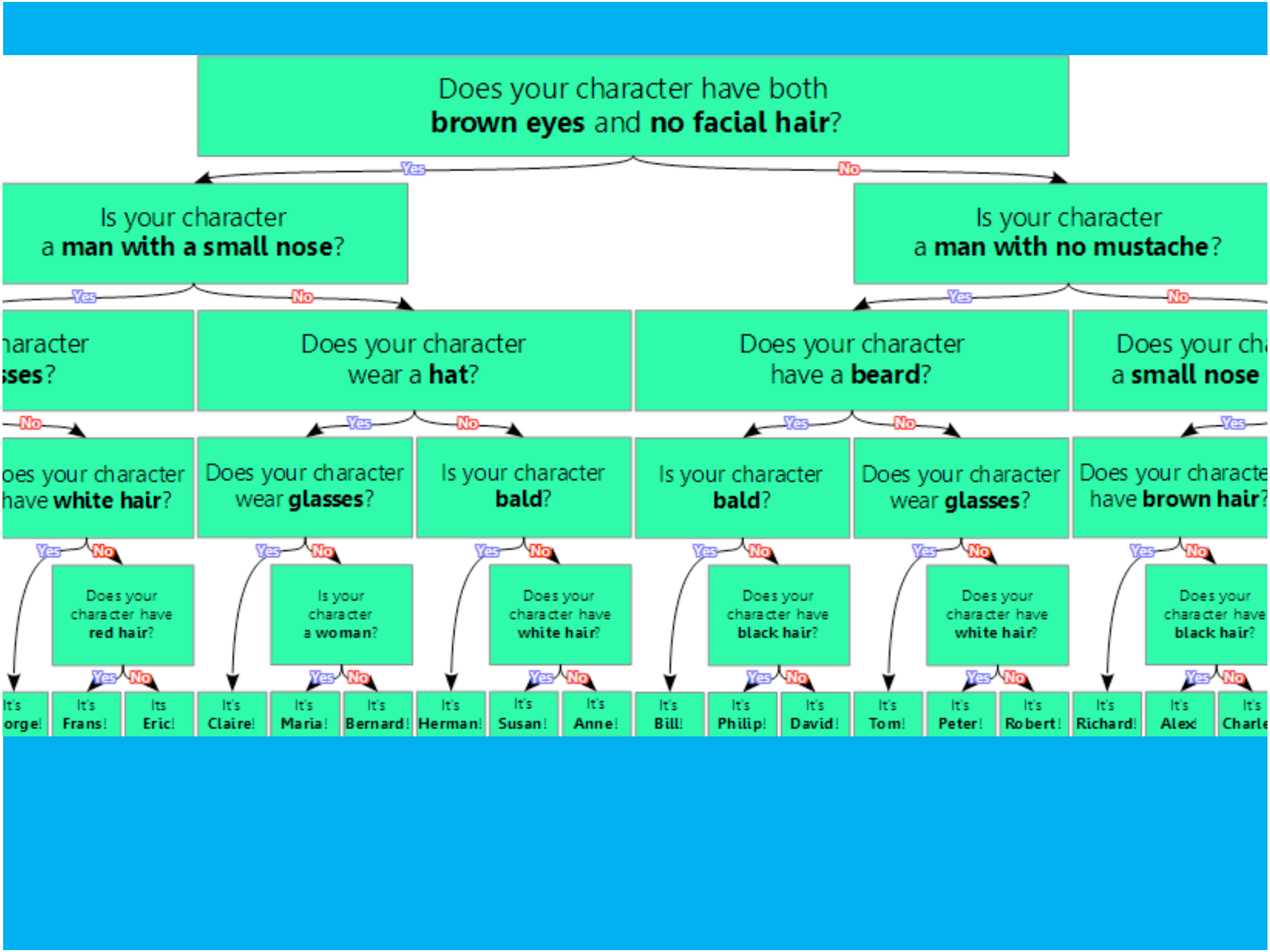
Advantages

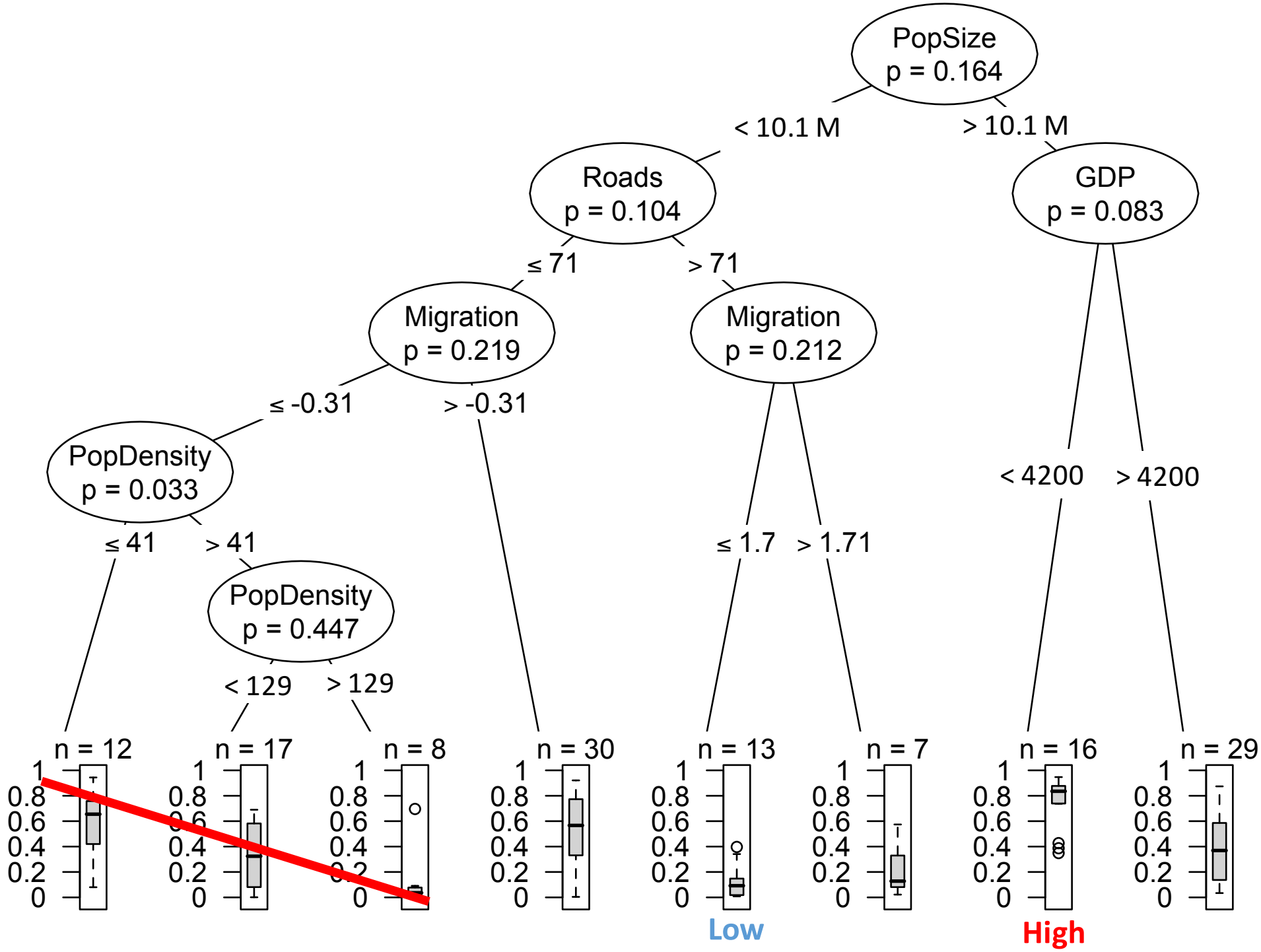
Flexible, easy to implement

Can handle multiple variables, multicolinearity

Easy to interpret


David
Eric
Frans
George
Herman
Joe
Maria
Paul
Peter
Philip
Richard
Robert
Sam
Susan
Tom
Robert
MB
MB GAMES
The original deduction game!
Can you guess the mystery face before your opponent?
GUESS WHO?
THE FLIP 'N' FIND FACE GAME
6+


Does your character have both **brown eyes** and **no facial hair**?
Yes
No
Is your character a **man with a small nose**?
Is your character a **man with no mustache**?
Yes
No
Yes
No
haracter sses?
Does your character wear a **hat**?
Does your character have a **beard**?
Does your cha a **small nose**
No
Yes
No
Yes
No
Yes
oes your character have **white hair**?
Does your character wear **glasses**?
Is your character **bald**?
Is your character **bald**?
Does your character wear **glasses**?
Does your characte have **brown hair**?
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Does your character have **red hair**?
Is your character a **woman**?
Does your character have **white hair**?
Does your character have **black hair**?
Does your character have **white hair**?
Does your character have **black hair**?
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
It's orge!
It's Frans!
Its Eric!
It's Claire!
It's Maria!
It's Bernard!
It's Herman!
It's Susan!
It's Anne!
It's Bill!
It's Philip!
It's David!
It's Tom!
It's Peter!
It's Robert!
It's Richard!
It's Alex
It's Charle


PopSize
p = 0.164
< 10.1 M
> 10.1 M
Roads
p = 0.104
GDP
p = 0.083
≤ 71
> 71
Migration
p = 0.219
Migration
p = 0.212
≤ -0.31
> -0.31
PopDensity
p = 0.033
< 4200
> 4200
≤ 41
> 41
≤ 1.7
> 1.71
PopDensity
p = 0.447
< 129
> 129
n = 12
n = 17
n = 8
n = 30
n = 13
n = 7
n = 16
n = 29
1
0.8
0.6
0.4
0.2
0
Low
High

Importance measures

Creating decision trees

R Code:

```
library(party)
d.tree = ctree(Ling.Diversity ~ popSize + migration + roads)
plot(d.tree)
```

More details:

Roberts, S. G., Torreira, F., and Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. Frontiers in Psychology, 6.

Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and trees of york english: was/were variation as a case study for statistical practice. Lang. Variat. Change 24, 135–178. doi: 10.1017/S0954394512000129

Decision trees with random effects:

REEMtree package in R (new)

Causal Graph Inference

Causal Graphs

Description

Infer the most likely graph of causal relations between variables from observational data

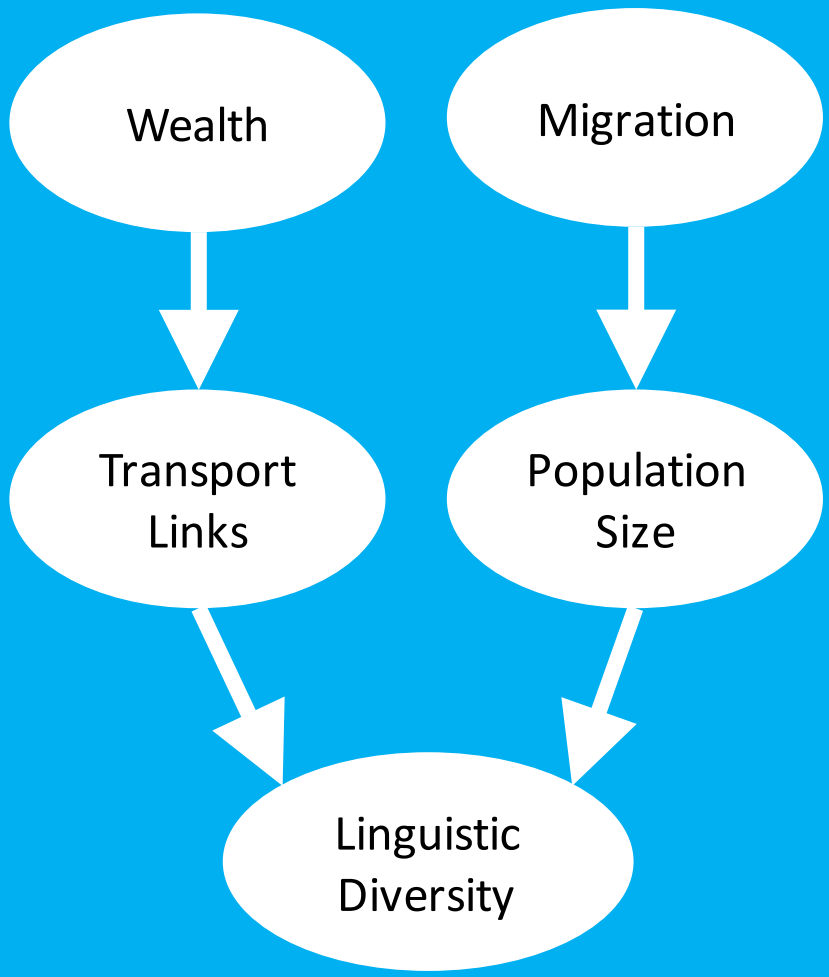
Advantages

Handles large numbers of variables

Can handle complex relationships

Easy to interpret

Causes of Diversity



Pearl, J. (2000). *Causality: models, reasoning, and inference*.

Inferring causal links

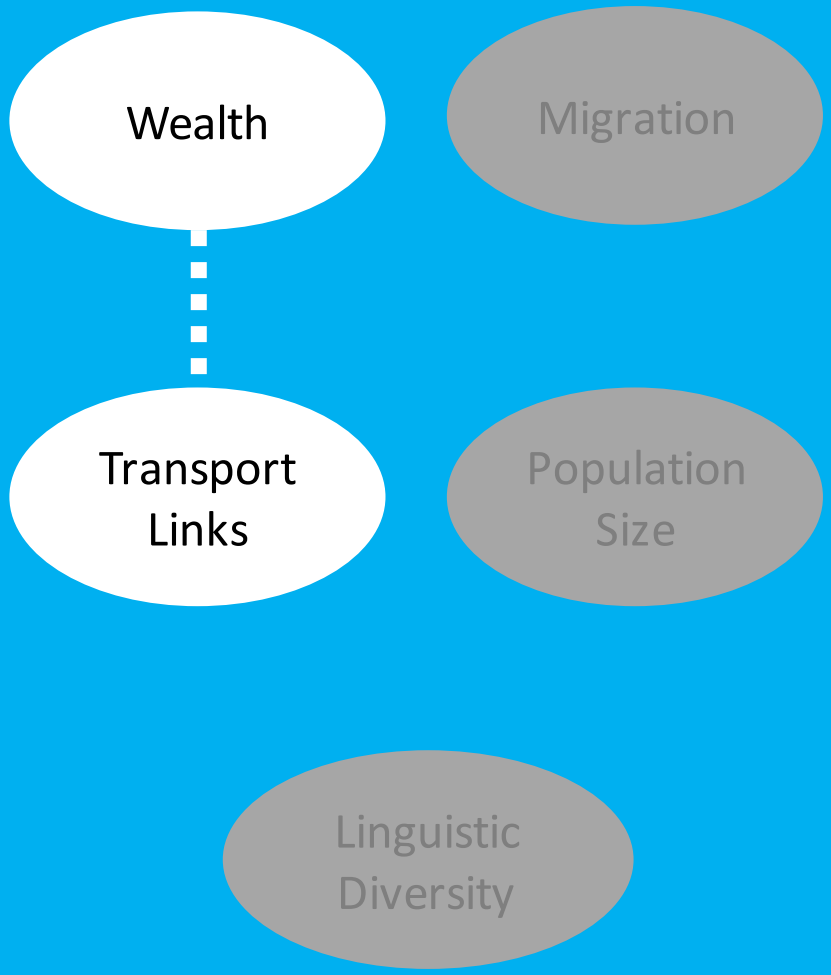
Look at one link at a time:

Are they correlated, when controlling for other variables?

If Yes, draw a link between them.
If No, remove the link.

After all tests, orient the edges using *colliders.*

See the *pcalg* package in R.



Pearl, J. (2000). *Causality: models, reasoning, and inference*.

Inferring causal links

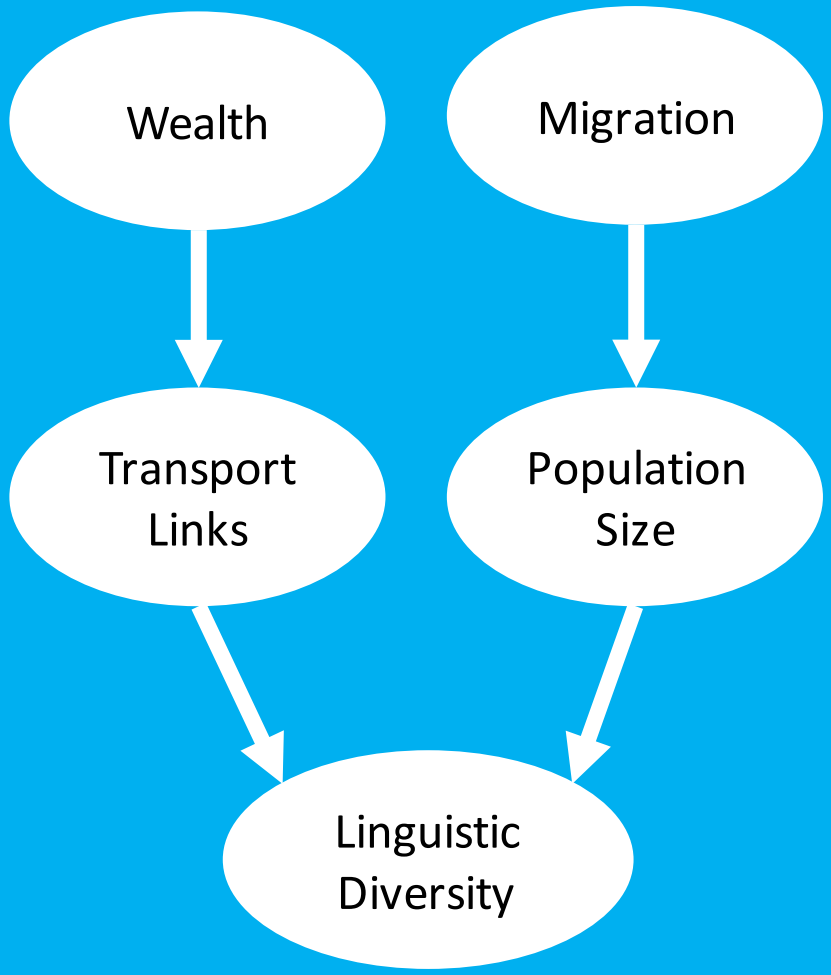
Look at one link at a time:

Are they correlated, when controlling for other variables?

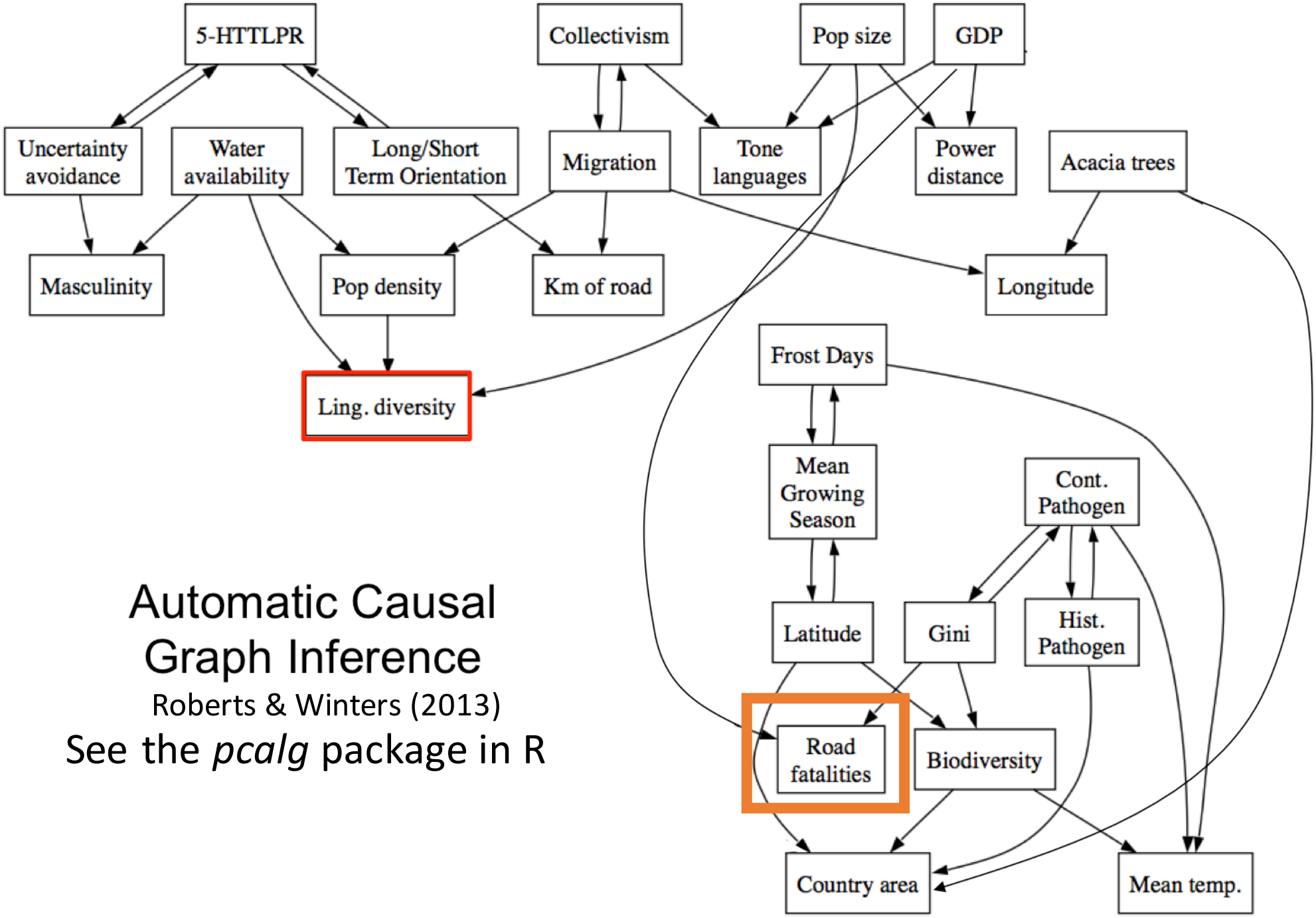
If Yes, draw a link between them.
If No, remove the link.

After all tests, orient the edges using *colliders.*

See the *pcalg* package in R.



Pearl, J. (2000). *Causality: models, reasoning, and inference*.



Automatic Causal Graph Inference

Roberts & Winters (2013)

See the *pcalg* package in R

Extra tips

Missing data? Imputation (mice package)

Regression with a phylogenetic tree? PGLS in ape, caper

Big data? Data tables (data.table, dplyr package)

Code for publication? R markdown in RStudio

Full course on R (Hadley et al.):

http://r4ds.had.co.nz/transform.html