Занятие 11

Многомерное шкалирование, кластерный анализ.

Generalized linear models и логистическая регресия.

Многомерные методы анализа

Мы рассмотрели способы получения новых переменных (и их интерпретации) на основе взаимосвязей (матриц ковариаций или корреляций) **между переменными** (размера *p* x *p*) – R-mode analysis.

Существует альтернативный способ получения новых переменных: построить матрицу **дистанций между объектами** (размера *n* x *n)* в пространстве исходных переменных, и получить новые переменные на основе этой матрицы - Q-mode analysis.

Дистанция – это просто сходство между объектами по исходным переменным.

Многомерное шкалирование (*Multidimentional scaling*)

У нас в руках:
p переменных
n объектов

Цель: **уменьшить число исходных переменных** с минимальными потерями информации.

(Мы уже встречали такую цель!)

Новые переменные получаются на основе **матрицы дистанций** между объектами так, чтобы **дистанции** в пространстве новых переменных **изменились как можно меньше**.
То, что было «похожим», должно остаться похожим; большие исходные различия должны остаться большими.

Многомерное шкалирование

Пример:

Исходные переменные – масса тела и рост
Новая переменная – «упитанность»

Различия большие

Масса – большая
Рост - маленький

Различия большие

Масса – маленькая
Рост - большой

Упитанность большая

Упитанность низкая

Можно по-разному померить различия между объектами:

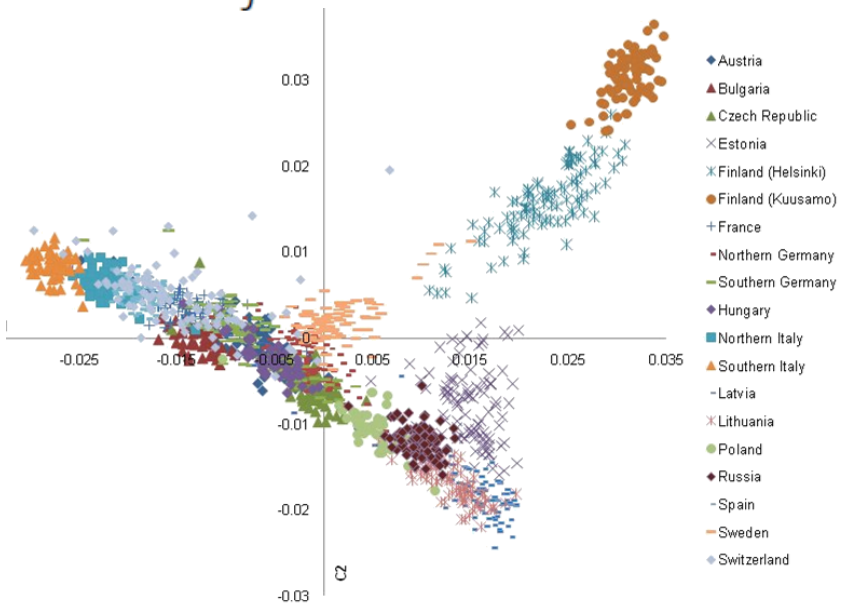
- Евклидовы дистанции – просто расстояния в пространстве переменных, как в геометрии;

$$\sqrt{\sum_{j=1}^{p}(y_{1j}-y_{2j})^2}$$

- Квадрат евклидова расстояния (увеличивает вес больших разностей);
- Манхэттенское расстояние (Manhattan увеличивает вес больших разностей).

$$\sum_{j=1}^{p}\left|(y_{1j}-y_{2j})\right|$$

Многомерное шкалирование м.б. основано на **любых показателях различий** (напр., генетических расстояниях)!



Многомерное шкалирование

Пример:

Мы наблюдаем поведение молодых сурков. У нас есть 15 переменных, описывающих социальное поведение - частоты контактов разных типов (погони, обнюхивания, игра…).

Мы хотим из 15 переменных получить 2-3, которые бы хорошо объясняли изменчивость в выборке (т.е., чтобы сурки, которые ведут себя по-разному, в пространстве новых переменных оказались далеко друг от друга).

1 этап. Подготовительный.

а) трансформируем данные (если между переменными нелинейные связи) и стандартизируем их (если переменные измерены в принципиально разных шкалах);

б) получаем *n* x *n* матрицу дистанций между объектами (сурками);

в) определяем, сколько новых переменных (измерений) мы будем получать (*k*);

Если исходная матрица различий содержит Евклидовы дистанции, и данные по всем переменным центрированы так что среднее = 0, взаиморасположение объектов будет таким же, как и в PCA в пространстве новых компонент.
T.e., classical scaling – обобщённый вариант PCA.

2 этап. Получение новых переменных.

а) из матрицы дистанций получаем новые переменные (так же, как и в R-mode analysis, у этих переменных есть eigenvalues и eigenvectors). Первая переменная описывает максимум различий между объектами.

б) объекты помещаются в пространство *k* новых переменных (первых, самых лучших) и многократно поворачиваются там, приближая дистанции между объектами в новом пространстве (*d-hat*) к реальным различиям между объектами (*d*).

Т.е., это итеративная процедура, новые переменные получаются не одним действием из матриц, а они постепенно подгоняются к данным.

Примечание про проблему сходимости (convergence problem)

Этап 3. оценка качества модели.

а) считается простой показатель того, насколько хорошо модель описывает реальные различия между объектами – stress. Чем он меньше, тем лучше: >0.3 – недопустимо, <0.2 – приемлемо, <0.1 – идеально;

б) Картинка, показывающая качество модели – диаграмма Шепарда

в) если стресс большой, повторяем процедуру, увеличив число измерений – новых переменных.

Стресс оценивается на основе residuals в регрессионной модели «дистанции в модели - реальные различия»

$$\sqrt{\frac{\sum (\tilde{d}_{hi} - \hat{\tilde{d}}_{hi})^2}{\sum \tilde{d}_{hi}^2}}$$

Этап 4. Интерпретация результатов.

а) существенный недостаток метода: **готового способа** выяснить, какие исходные переменные вносят вклад в новые и коррелируют с новыми, **НЕТ**!

б) это потому, что новые переменные НЕ являются линейной комбинацией исходных, они получены вообще не из них а из дистанций.

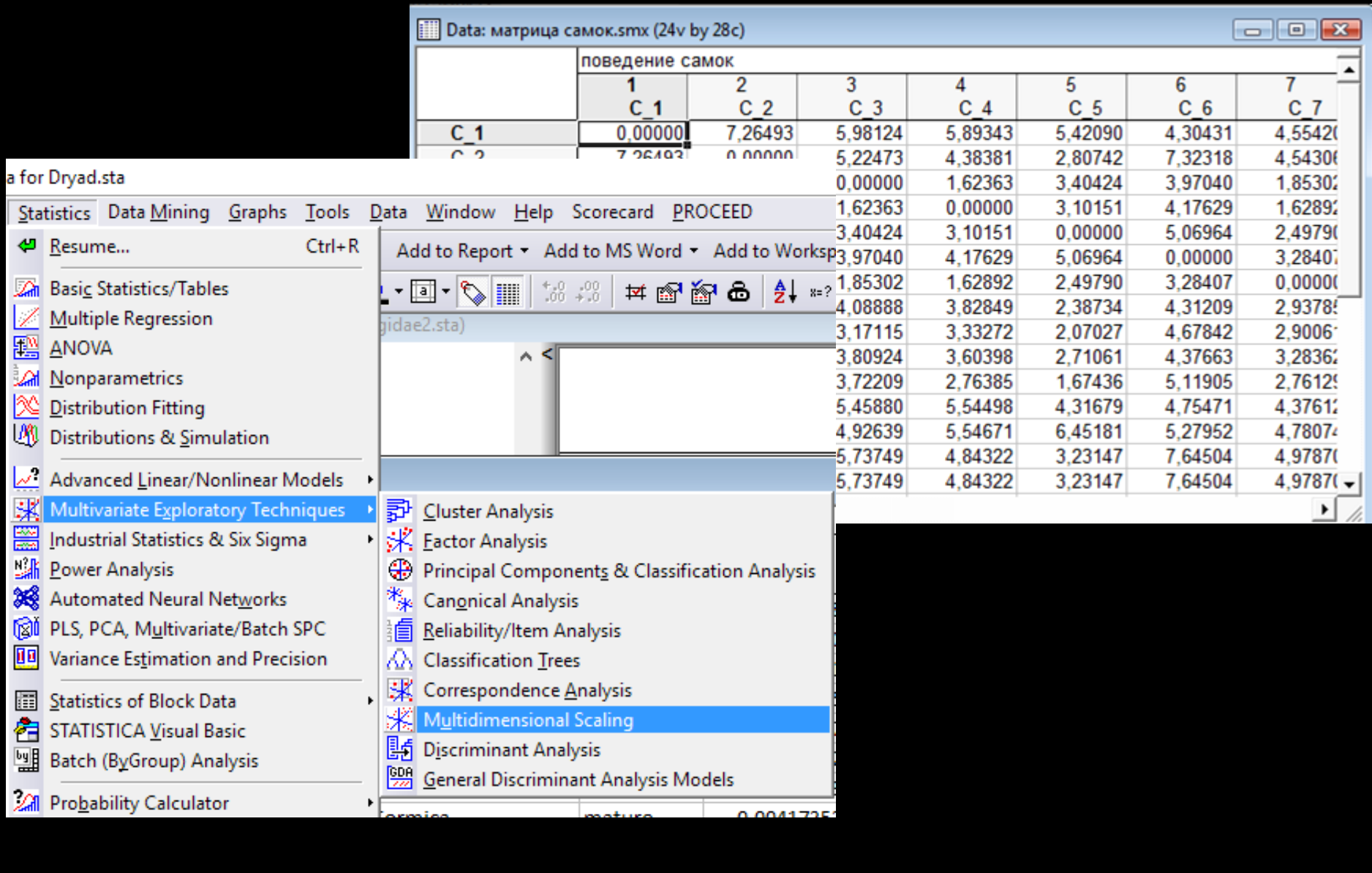
в) Допустимо проанализировать обычные корреляции исходных переменных с новыми измерениями.

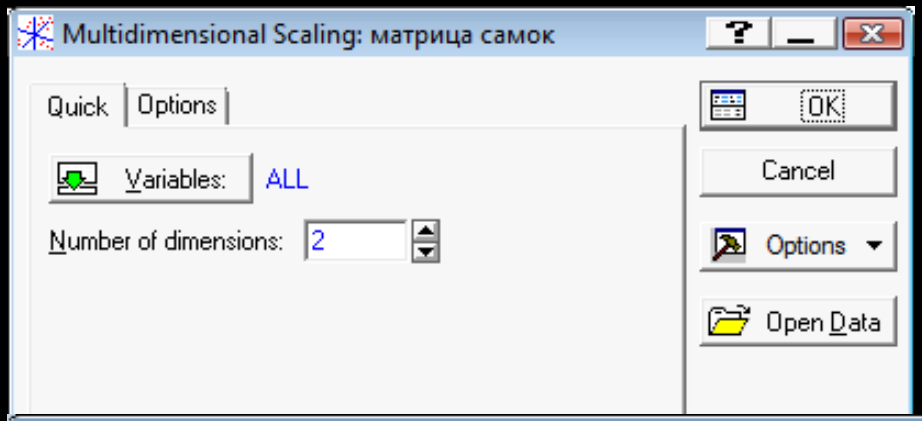
д) теперь можно получить для каждого объекта значения новых переменных и использовать их в дальнейшем анализе (например, для сравнения групп).

Стресс оценивается на основе residuals в регрессионной модели «дистанции в модели - реальные различия»

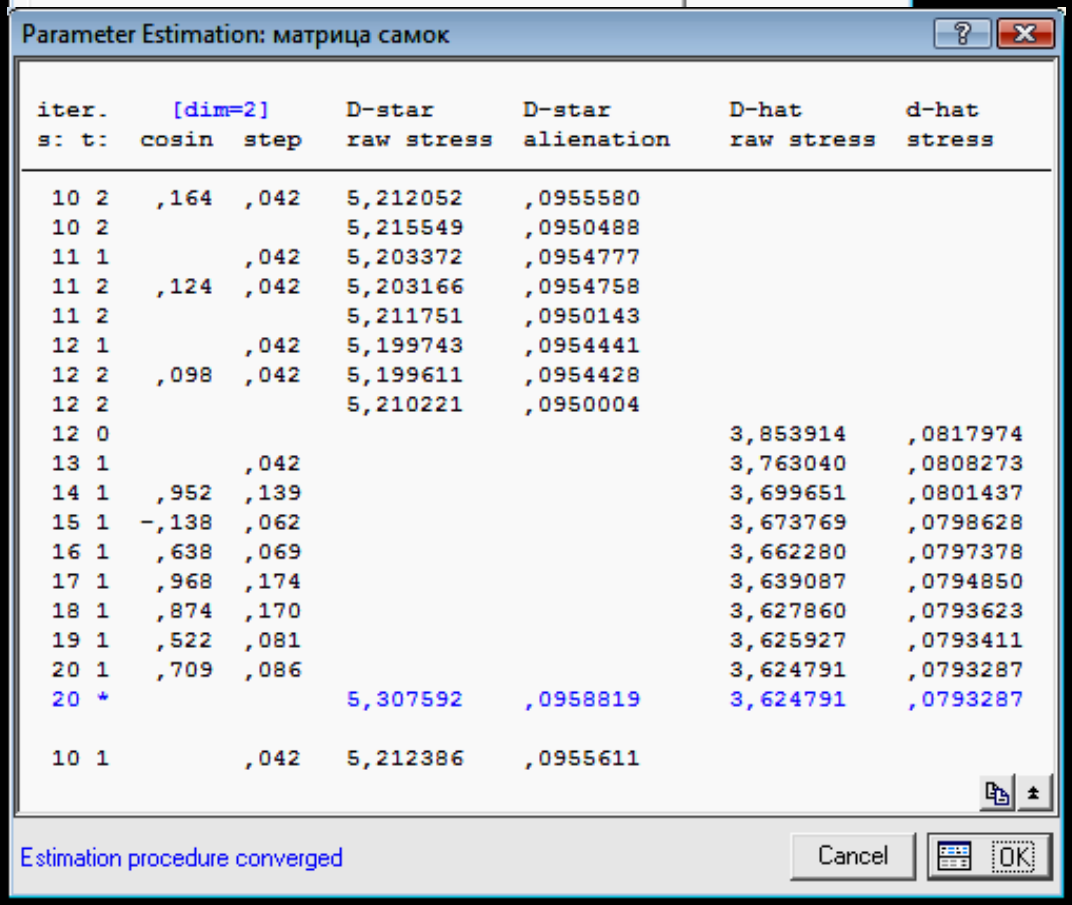
$$\sqrt{\frac{\sum (d_{\tilde{hi}} - \hat{d}_{\tilde{hi}})^2}{\sum d_{\tilde{hi}}^2}}$$

Данные для анализа должны быть представлены **матрицей дистанций** (получается в модуле кластерного анализа)



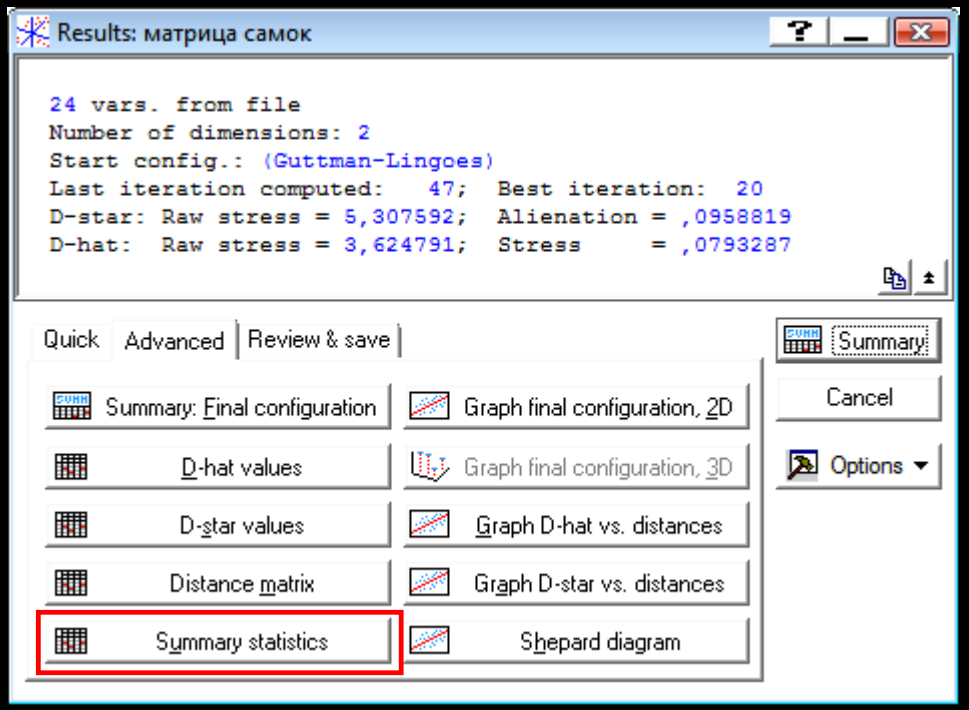


Выбрали число новых переменных



Программа вращает объекты в пространстве так, чтобы расстояния между ними в модели лучше всего соответствовали исходным расстояниям между объектами

(чем больше измерений в модели, тем лучше модель будет отражать реальность, но тем она будет сложнее)



Мы получили итоговую конфигурацию. Посмотрим, насколько она хороша.

Configuration (матрица самок)

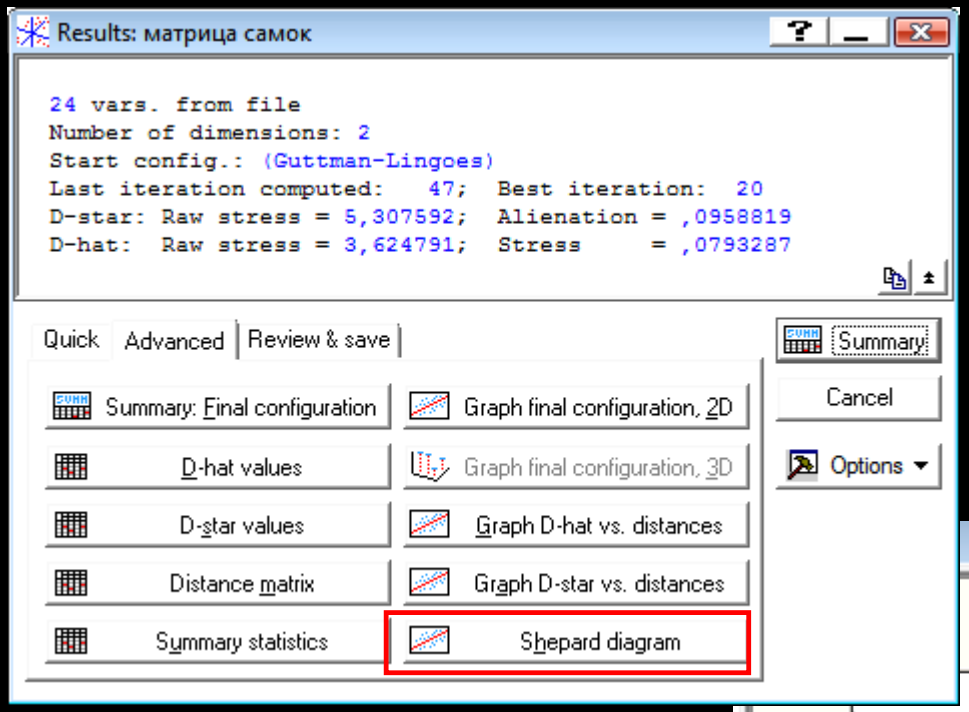
Final Configuration (матрица самок)
D-star: Raw stress = 5,307592; Alienation = ,0958819
D-hat: Raw stress = 3,624791; Stress = ,0793287

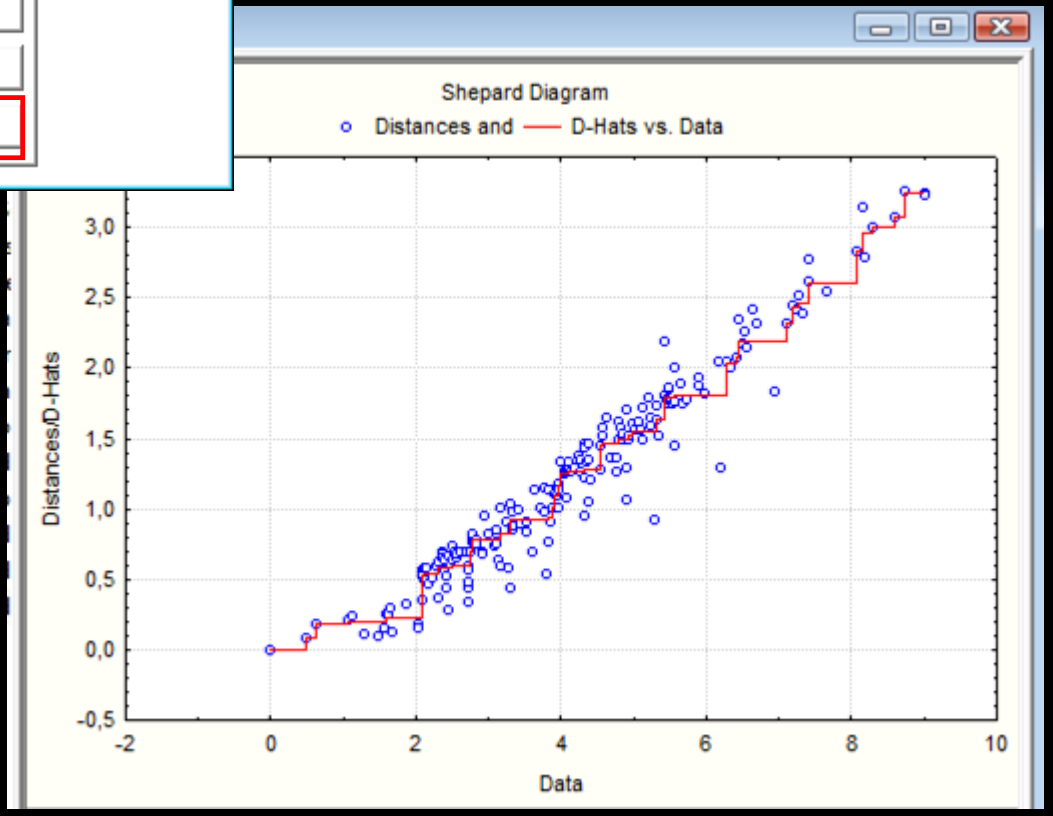
	Distance	D-star	D-hat
D(22,15)	0,000243	0,000074	0,000163
D(15,14)	0,000087	0,000083	0,000163
D(22,16)	0,000328	0,000085	0,000163
D(19,14)	0,000083	0,000087	0,000163
D(22,19)	0,000074	0,000156	0,000163
D(19,16)	0,000254	0,000169	0,000167
D(22,14)	0,000156	0,000172	0,000167
D(16,14)	0,000172	0,000243	0,000167
D(19,15)	0,000169	0,000254	0,000167
D(16,15)	0,000085	0,000328	0,000167
D(21, 2)	0,077300	0,077300	0,077300
D(16, 2)	0,179191	0,099765	0,179191
D(15, 2)	0,179237	0,115304	0,179237
D(22, 2)	0,179369	0,122014	0,179328
D(14, 2)	0,179287	0,160012	0,179328
D(19, 2)	0,179330	0,160621	0,179330
D(24, 5)	0,214672	0,179066	0,198856
D(22,21)	0,239922	0,179191	0,198856
D(21,15)	0,239840	0,179237	0,198856
D(21,19)	0,239898	0,179287	0,198856
D(21,14)	0,239872	0,179330	0,198856
D(21,16)	0,239812	0,179369	0,198856
D(17, 7)	0,115304	0,214672	0,198856
D(24,11)	0,099765	0,239812	0,198856
D(20,11)	0,160621	0,239840	0,198856

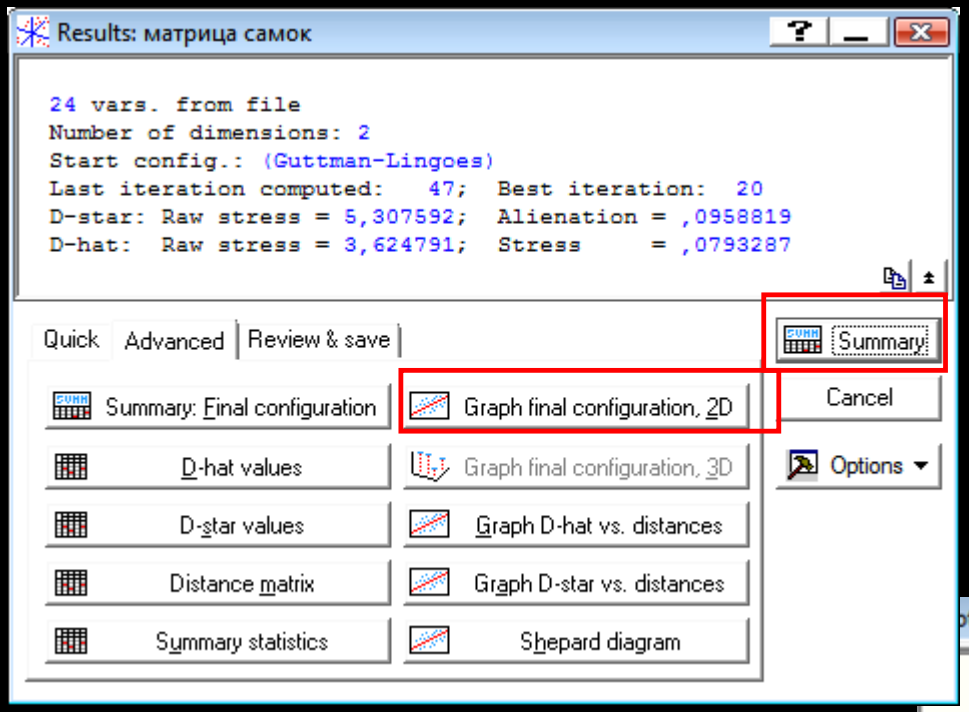
D-star и D-hat –смоделированные дистанции между измерениями (по-разному рассчитанные); расстояния упорядочены по ним.

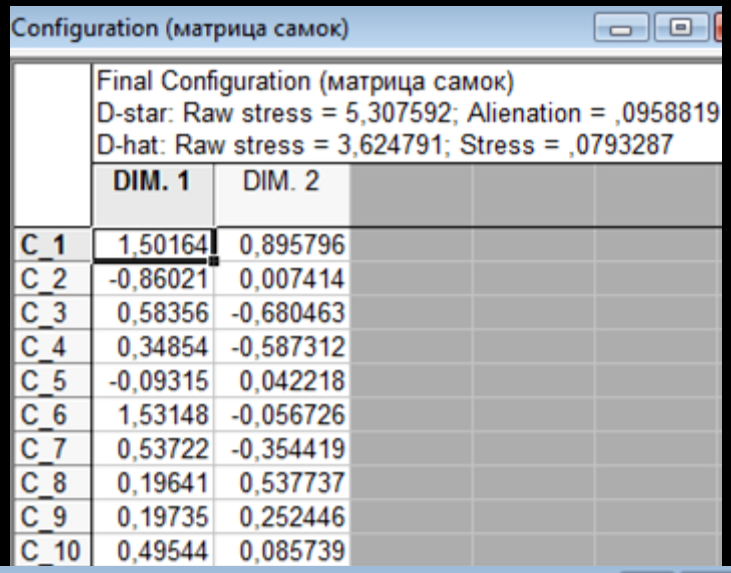
Distance – реальные дистанции; чем лучше они выстроены по возрастанию, тем лучше модель.

Диаграмма Шепарда тоже показывает, хорошо ли модель согласуется с исходными данными: чем ближе точки к красной линии, тем лучше.

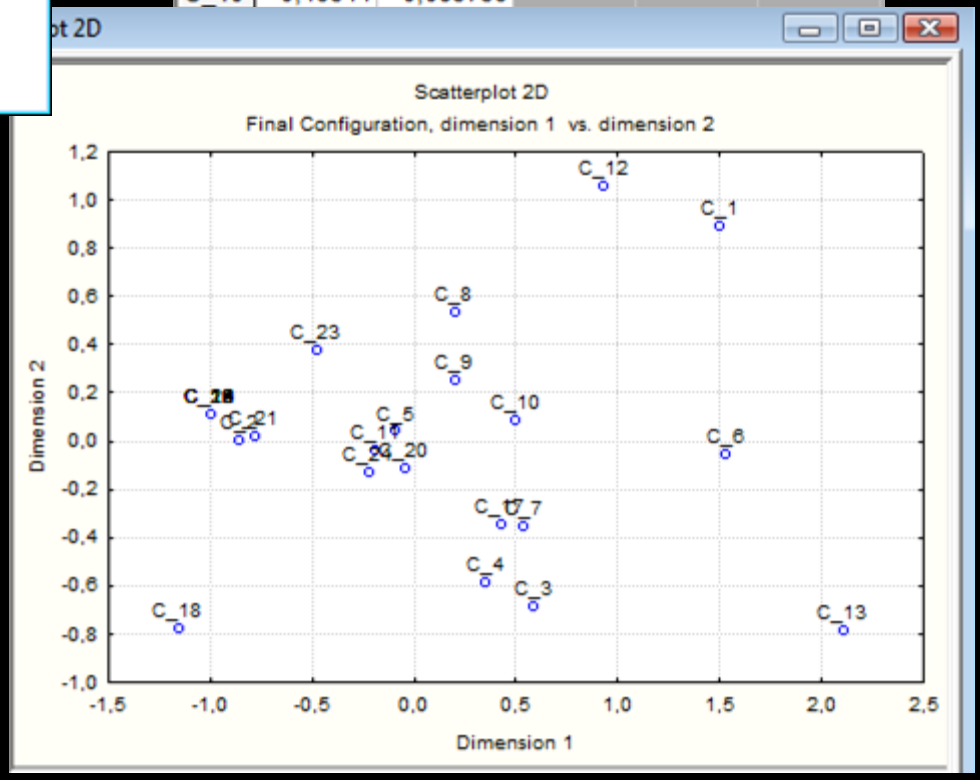








	DIM. 1	DIM. 2
C_1	1,50164	0,895796
C_2	-0,86021	0,007414
C_3	0,58356	-0,680463
C_4	0,34854	-0,587312
C_5	-0,09315	0,042218
C_6	1,53148	-0,056726
C_7	0,53722	-0,354419
C_8	0,19641	0,537737
C_9	0,19735	0,252446
C_10	0,49544	0,085739



Наконец, получаем значения новых переменных для объектов; строим картинку, где они расположены в пространстве этих переменных

Дальнейший анализ с меньшим числом переменных

Для публикаций

В методы: для упрощения структуры данных и графической визуализации различий между объектами использовали метод многомерного шкалирования.

В результаты:

✓количество новых переменных;
✓Stress;
✓корреляции (Пирсона или Спирмана) новых переменных с исходными;
✓картинка расположения объектов в пространстве новых переменных.

Требования к выборкам для проведения MDS

1. В качестве исходных данных можно взять любую матрицу дистанций – **никаких ограничений и требований** к выборке!
2. Для интерпретации новых переменных желательно, чтобы связь переменных была ***линейной*** (диагностика - осмотр скаттерплотов; при необходимости - трансформация);
3. Желательно исключить **аутлаеры,** если они меняют результат;
4. Если переменные измерены в **разных шкалах**, их следует **стандартизировать**.
5. Следует исключить переменные, сильно коррелирующие с другими.

MDS – аналог метода главных компонент, не имеющий ограничений PCA, но его возможности крайне ограничены (в сравнении с PCA); лучше пользоваться им лишь в крайнем случае.

Кластерный анализ

У нас в руках:
p переменных
n объектов

Цель – объединить эти объекты в группы, внутри которых они будут более сходными, чем объекты из разных групп.

Результат – получение дендрограммы (иерархического дерева) – только картинка, никакой проверки гипотез и пр..

Можно выявить внутри вида группы, различающиеся по морфологии, экологии, поведению, физиологическим характеристикам. Сгруппировать пробы по химическому составу. Разделить местообитания по видовому составу.

Поведенческие тактики самцов сусликов:
активные и пассивные самцы

Кластерный анализ

Основа кластерного анализа – **матрица дистанций** между объектами (в пространстве *p* переменных). Дистанции могут быть любыми, как и в MDS.

Результат - **объединение сходных объектов** в группы (кластеры).

Иерархический

Объекты по одному присоединяются к кластеру и он становится всё больше и больше

Не иерархический

Объекты присоединяются по одному, но кластеры перегруппировываются в процессе анализа

Переменные, измеренные в разных шкалах, должны быть стандартизированы

Кластерный анализ

Иерархический метод:

Накопительное группирование (agglomerative hierarchical clustering).

Начинает с того, что **каждый объект** – отдельная **группа**, заканчивает одной группой со всеми объектами.

a) Получается матрица дистанций между объектами;

b) Пара самых «похожих» объектов объединяется в первый кластер;

c) Пересчитывается матрица дистанций, как будто новый кластер – это один объект;

d) Получается второй кластер

e) …

Много разных показателей дистанций:

- Евклидовы дистанции;
- Квадрат евклидова расстояния (увеличивает вес больших разностей);
- Манхэттенское расстояние;
- … (Quinn, Keough, 2002)

Как мерить расстояние между кластерами?

- ✓Метод **ближайшего соседа** (Single linkage = nearest neighbor) - расстояние между кластерами = расстоянию между ближайшими объектами в них;
- ✓Полная связь (**complete linkage**); расстояние между кластерами = расстоянию между самыми удалёнными объектами в них (не годится, если кластеры формируют цепочки);
- ✓ **Average linkage** – меряют среднее расстояние между объектами в кластерах (unweighted pair-group average).

Есть ещё Divisive hierarchical clustering, когда, наоборот, от одного большого кластера по одному отделяются объекты.

Кластерный анализ

Пример:

Учёные анализируют численность разных животных вокруг деревень в Африке; у них 24 трансекта (это **объекты**) и 6 категорий животных (человекообразные обезьяны, птицы, слоны, копытные, мартышки, грызуны – это **переменные**).

Попробуем выявить, не образуют ли объекты (трансекты) группы в пространстве этих 6-и переменных.

(трансекты проложены в нацпарках, на вырубках и на прочих территориях)

Koerner SE, Poulsen JR, Blanchard E, Okouyi J, Clark CJ (2016) Data from: Vertebrate community composition and diversity declines along a defaunation gradient radiating from rural villages in Gabon. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.vs97g

Кластерный анализ

: species diversity.sta (26v by 24c)

D:\Nina\statistica\datasets\Копия IvindoData_DryadVersion.xlsx : Data

	RA_Apes	RA_Birds	**RA_Elephant**	RA_Monkeys	RA_Rodent	RA_Ungulate
Neither	0	85,02835	0,287088	9,086342	3,739324	1,858896
Logging	0	72,986541	0,363769	23,121135	3,266642	0,261913
Neither	0	57,820337	0	37,745098	3,191974	1,037392
				35,130475	2,092195	2,563513
				27,707992	3,640998	1,037684
				39,404145	1,252159	1,373057
				21,456161	5,407493	0
				38,430185	2,969605	3,051124
				54,118587	1,287058	8,121782
				23,937286	3,974826	8,7667
				34,954971	1,924775	4,167402
				25,580767	4,051864	1,877364

ta

Statistics | Data Mining | Graphs | Tools | Data | Window | Help | Scorecard | PROCEED

Add to Report | Add to MS Word | Add to Workspa

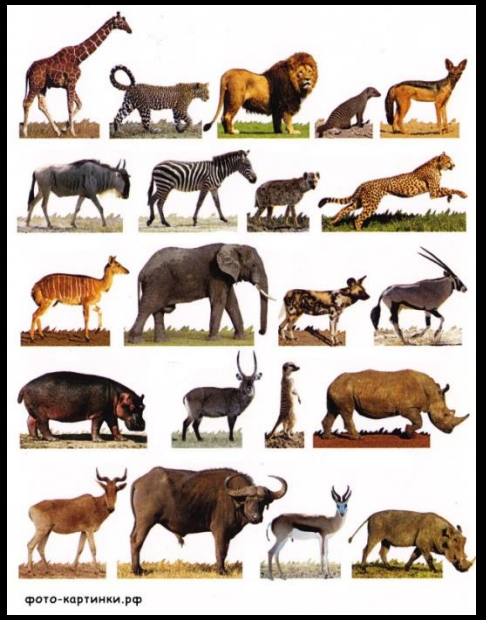
(26v by 24c)

Nina\statistica\datasets\Копия IvindoData_D

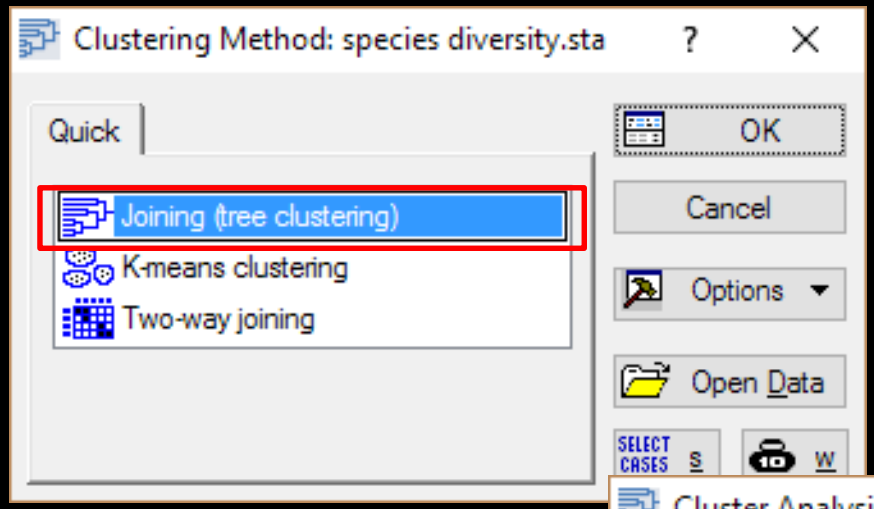
Apes | RA_Birds | **RA_Elephant** | RA

- Resume... Ctrl+R
- Basic Statistics/Tables
- Multiple Regression
- ANOVA
- Nonparametrics
- Distribution Fitting
- Distributions & Simulation
- Advanced Linear/Nonlinear Models
- Multivariate Exploratory Techniques
- Industrial Statistics & Six Sigma
- Power Analysis
- Automated Neural Networks
- PLS, PCA, Multivariate/Batch SPC
- Variance Estimation and Precision
- Statistics of Block Data
- STATISTICA Visual Basic
- Batch (ByGroup) Analysis
- Probability Calculator

- Cluster Analysis
- Factor Analysis
- Principal Components & Classification Analysis
- Canonical Analysis
- Reliability/Item Analysis
- Classification Trees
- Correspondence Analysis
- Multidimensional Scaling
- Discriminant Analysis
- General Discriminant Analysis Models

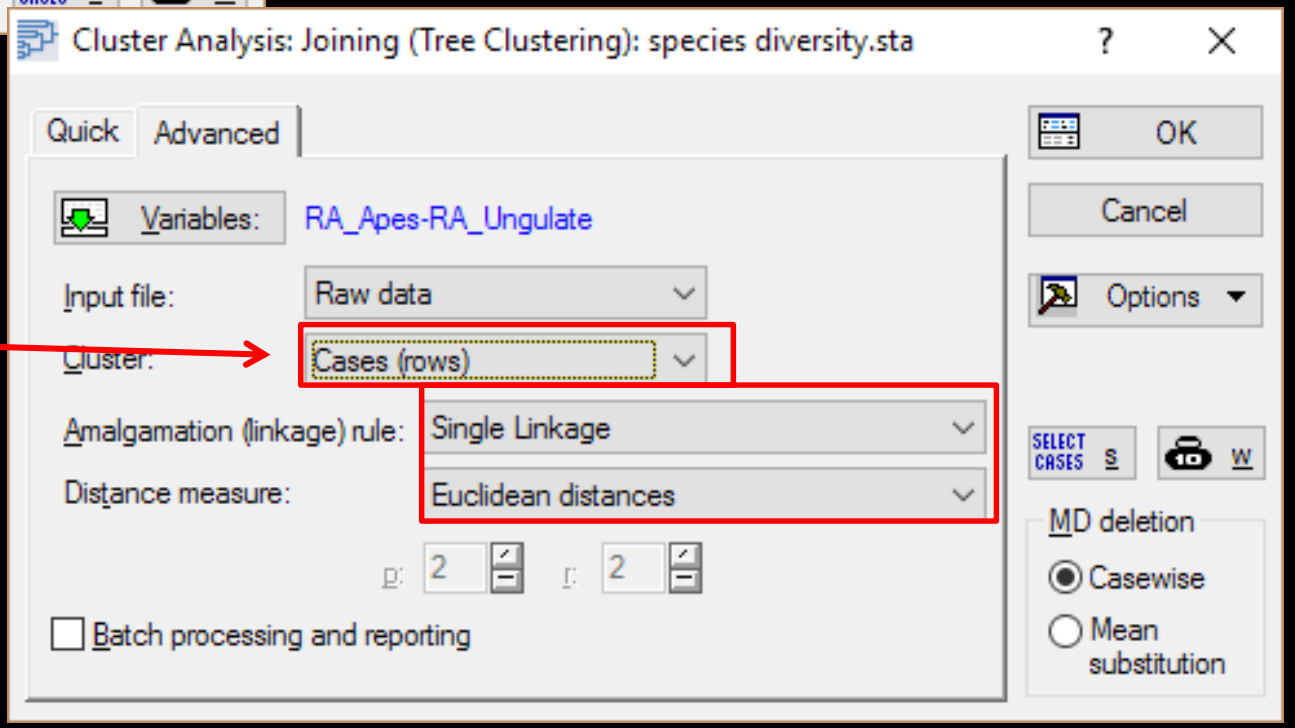


Кластерный анализ: иерархический метод



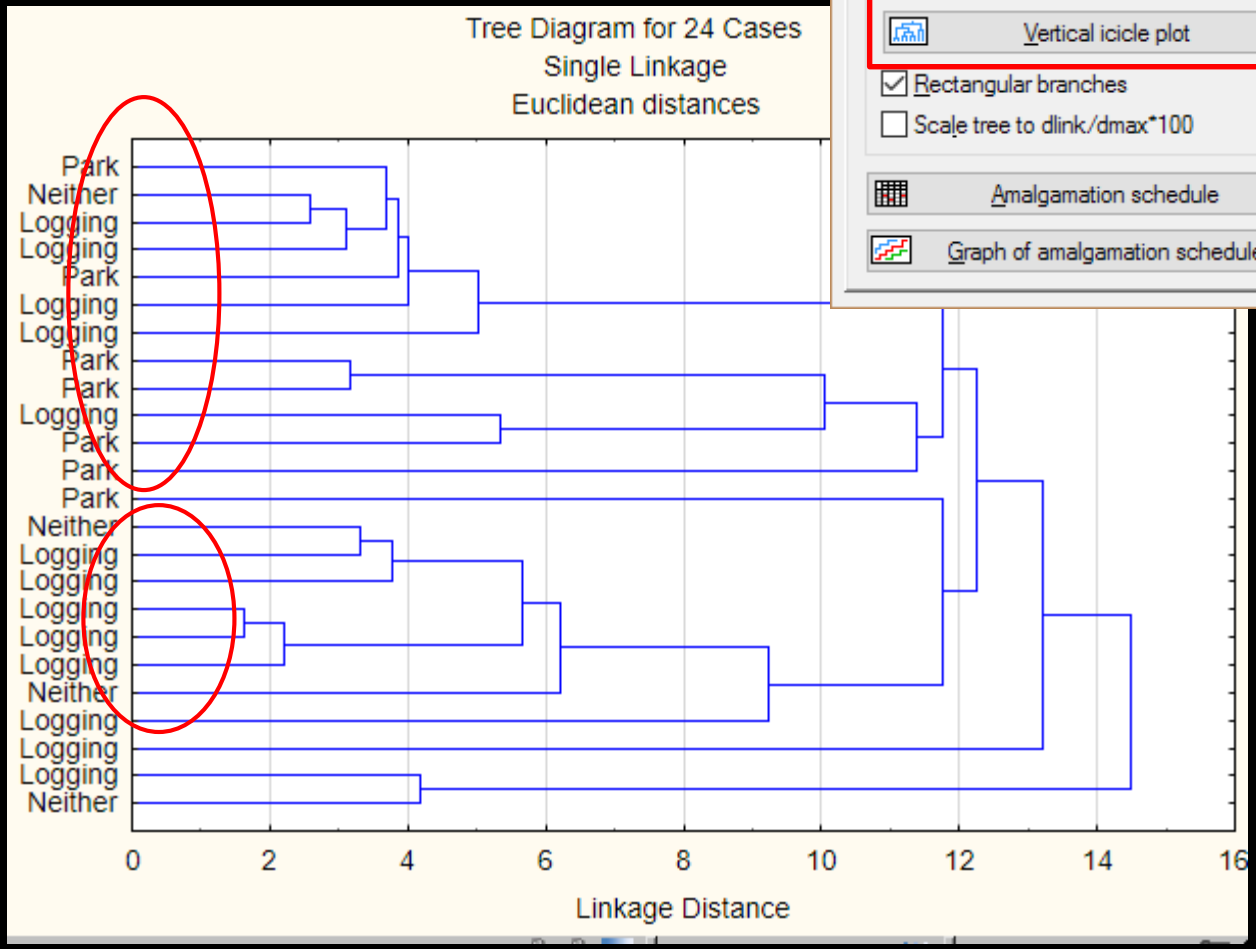
Задаем переменные, способ вычисления дистанций и метод объединения объектов в кластеры

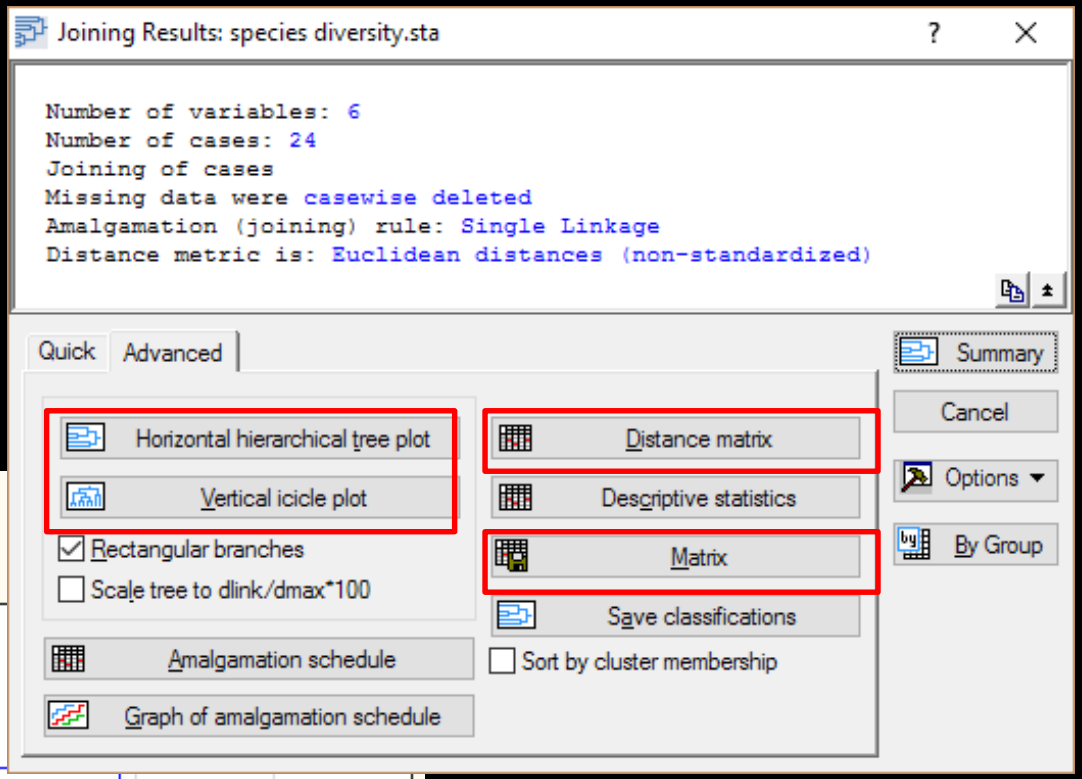
Здесь важно выставить строки, а не столбцы



Кластерный анализ

Получаем картинки с деревьями и смотрим, на каких уровнях выделяются кластеры (чтобы у объектов появились обозначения, надо копировать столбец с ними в названия строк)





Можно получить матрицу дистанций между наблюдениями (например, для многомерного шкалирования)

Кластерный анализ

Можно посмотреть, на каких дистанциях какие трансекты объединяются в кластеры

Joining Results: species diversity.sta

```
Number of variables: 6
Number of cases: 24
Joining of cases
Missing data were casewise deleted
Amalgamation (joining) rule: Single Linkage
Distance metric is: Euclidean distances (non-standardized)
```

Quick | Advanced

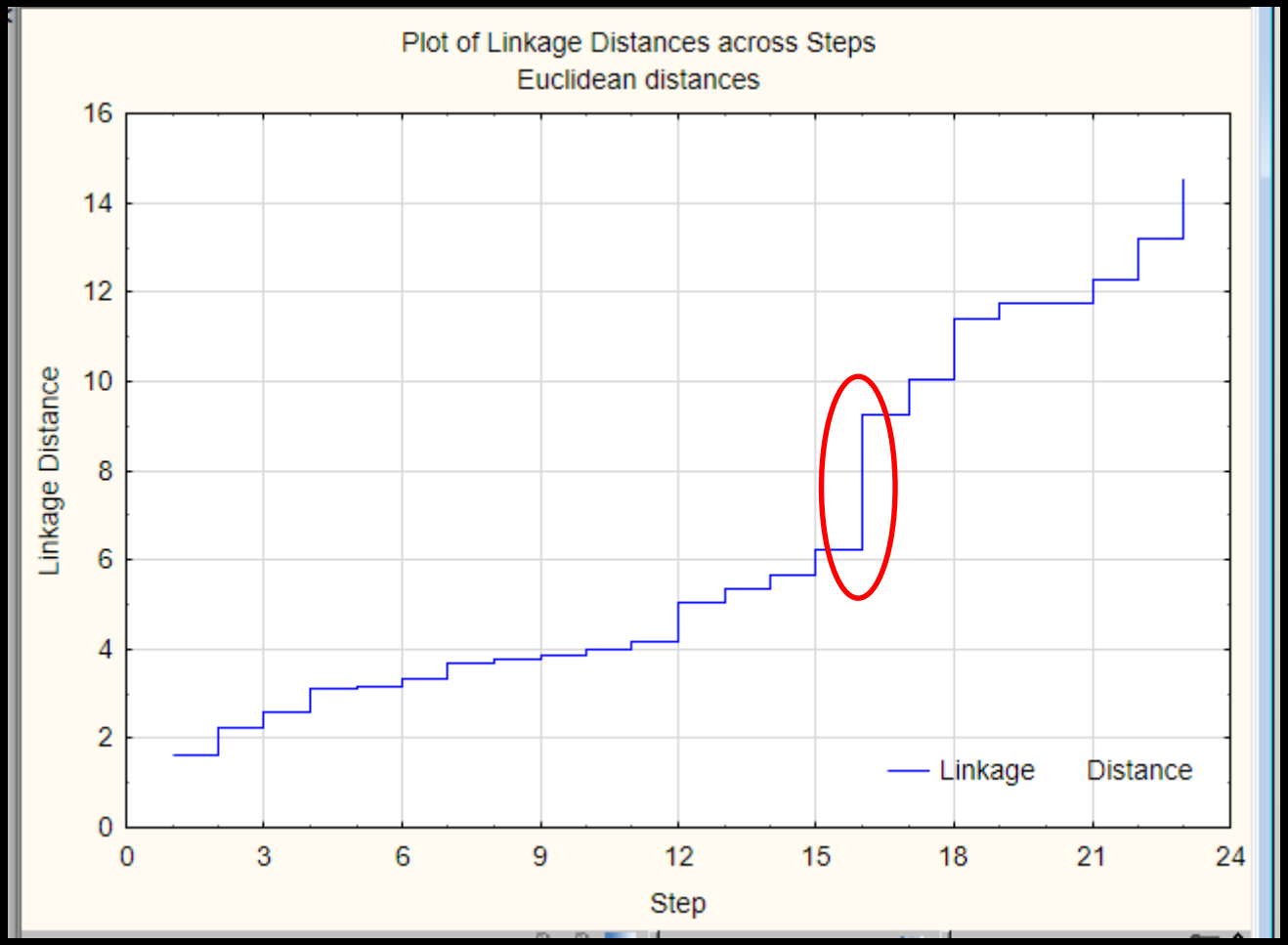
Horizontal hierarchical tree plot
Vertical icicle plot
Rectangular branches
Scale tree to dlink/dmax*100
Amalgamation schedule
Graph of amalgamation schedule
Distance matrix
Descriptive statistics
Matrix
Save classifications
Sort by cluster membership

Summary
Cancel
Options
By Group

Amalgamation Schedule (species diversity.sta)
Single Linkage
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj			
1,634135	Logging	Logging						
2,219885	Logging	Logging	Logging					
2,578872	Neither	Logging						
3,119694	Neither	Logging	Logging					
3,162199	Park	Park						
3,316061	Neither	Logging						
3,688090	Park	Neither	Logging	Logging				
3,769812	Neither	Logging	Logging					
3,870564	Park	Neither	Logging	Logging	Park			
4,010718	Park	Neither	Logging	Logging	Park	Logging		
4,178964	Logging	Neither						
5,022119	Park	Neither	Logging	Logging	Park	Logging	Logging	
5,339507	Logging	Park						
5,650817	Neither	Logging	Logging	Logging	Logging	Logging		
6,222314	Neither	Logging	Logging	Logging	Logging	Logging	Neither	
9,235194	Neither	Logging	Logging	Logging	Logging	Logging	Neither	Log
10,04474	Park	Park	Logging	Park				
11,39306	Park	Park	Logging	Park	Park			
11,76157	Park	Neither	Logging	Logging	Park	Logging	Logging	
11,76466	Park	Neither	Logging	Logging	Logging	Logging	Logging	Ne
12,26400	Park	Neither	Logging	Logging	Park	Logging	Logging	
13,19922	Park	Neither	Logging	Logging	Park	Logging	Logging	
14,49395	Park	Neither	Logging	Logging	Park	Logging	Logging	

Кластерный анализ



По этому графику можно посмотреть, на каком расстоянии происходят скачки в дистанциях присоединения. Если такие скачки есть, значит, есть и кластеры.

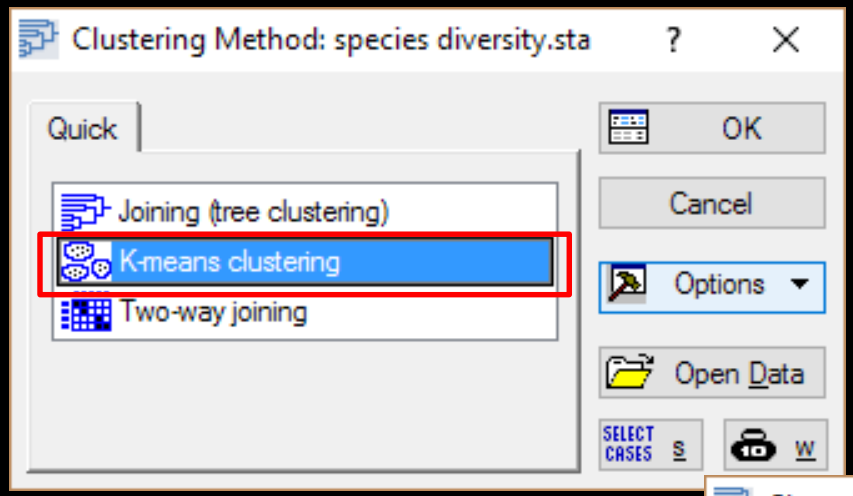
Неиерархический метод: метод К-средних (k-means clustering) – начинает с одиночных объектов, но в процессе объединения объекты могут быть **перегруппированы**. Они организуются в **заданное число групп** (k).

Объекты перераспределяются в кластеры в ходе нескольких итераций, так, чтобы кластеры различались как можно сильнее.

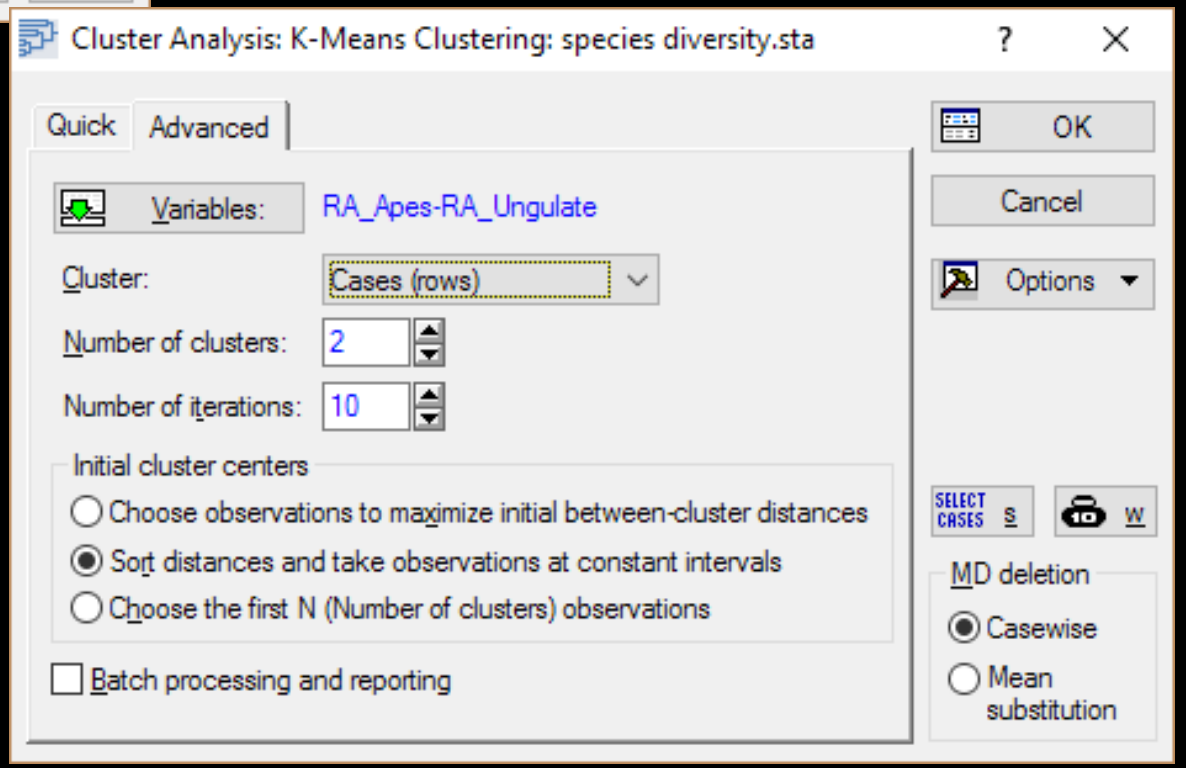
Можно выбирать метод, который даёт лучший результат.

Разобьём трансекты из Африки на 2 кластера

Кластерный анализ

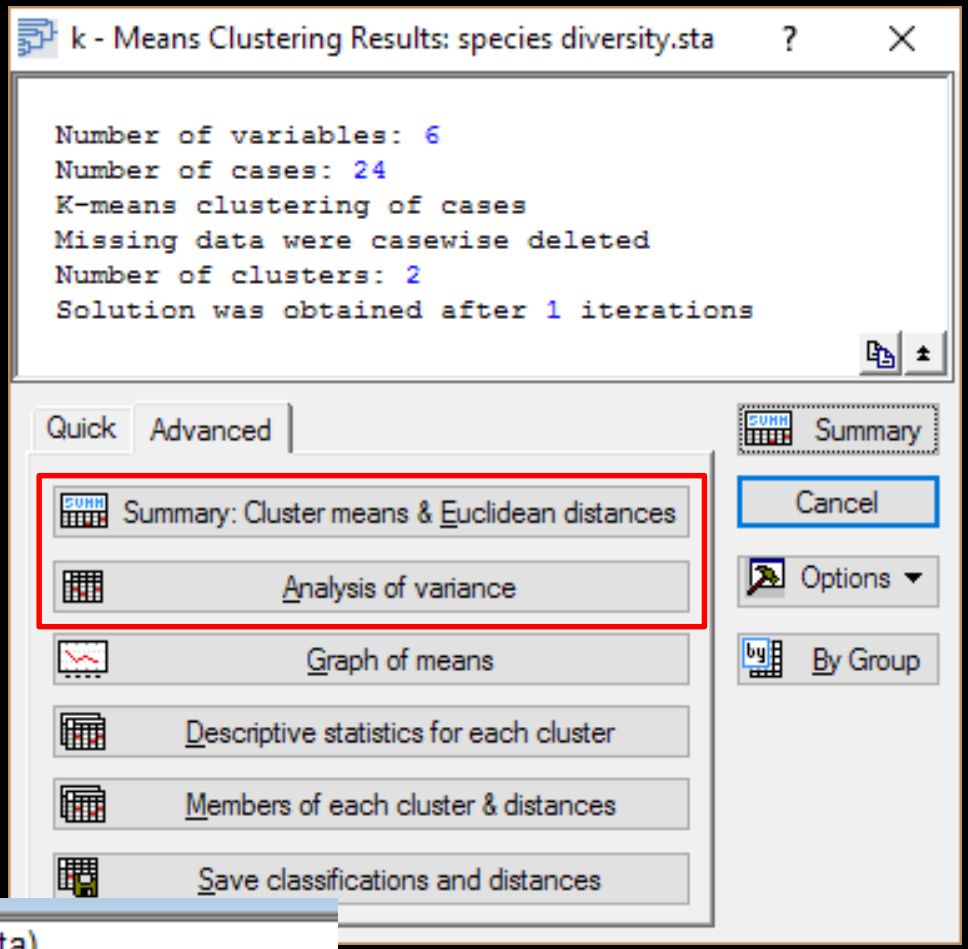


Выбираем переменные и задаем число групп



Кластерный анализ

Variable	Cluster Means (species Cluster No. 1	Cluster No. 2
RA_Apes	2,16302	1,90510
RA_Birds	48,10677	71,09298
RA_Elephant	0,54106	0,55071
RA_Monkeys	40,70191	20,17682
RA_Rodent	2,51893	4,17513
RA_Ungulate	5,95253	2,05563



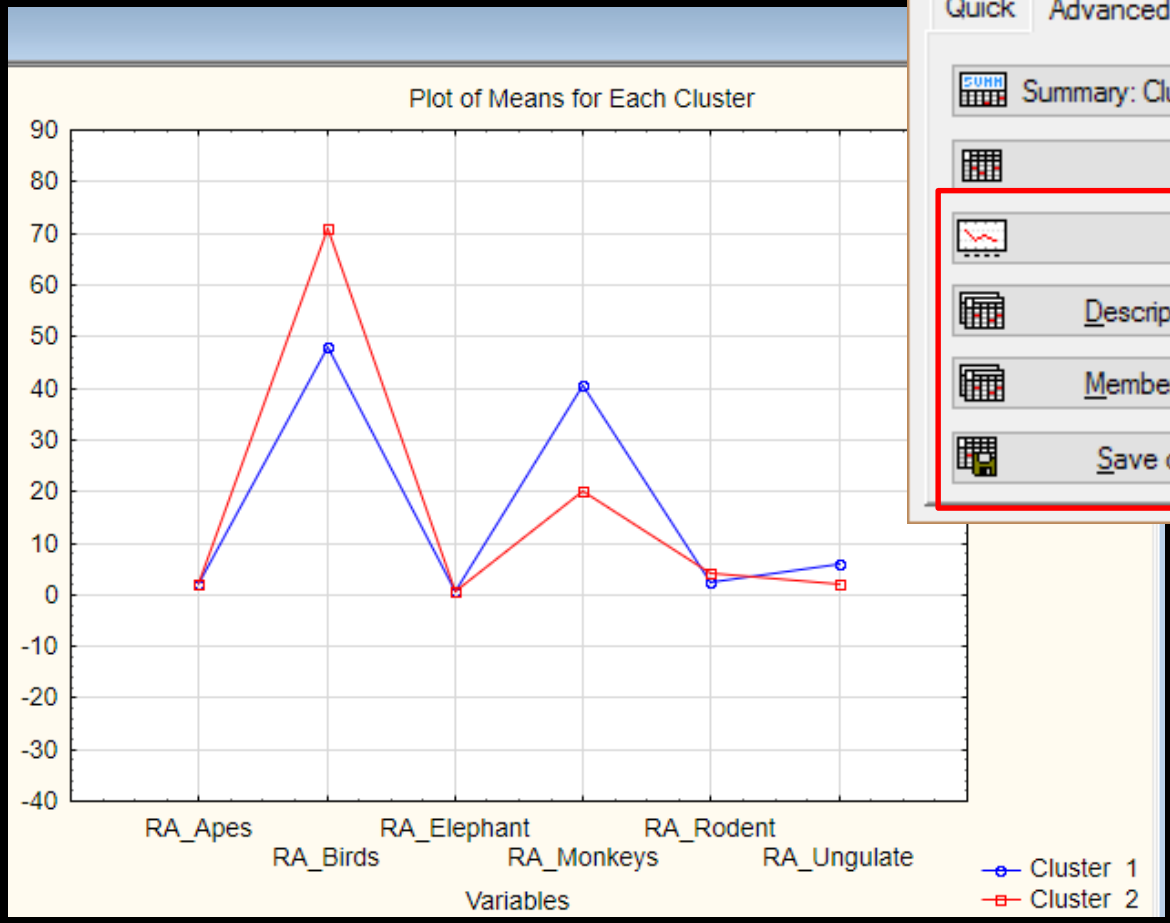
Программа получила 2 кластера; посмотрим, какие переменные принимают в них какие значения

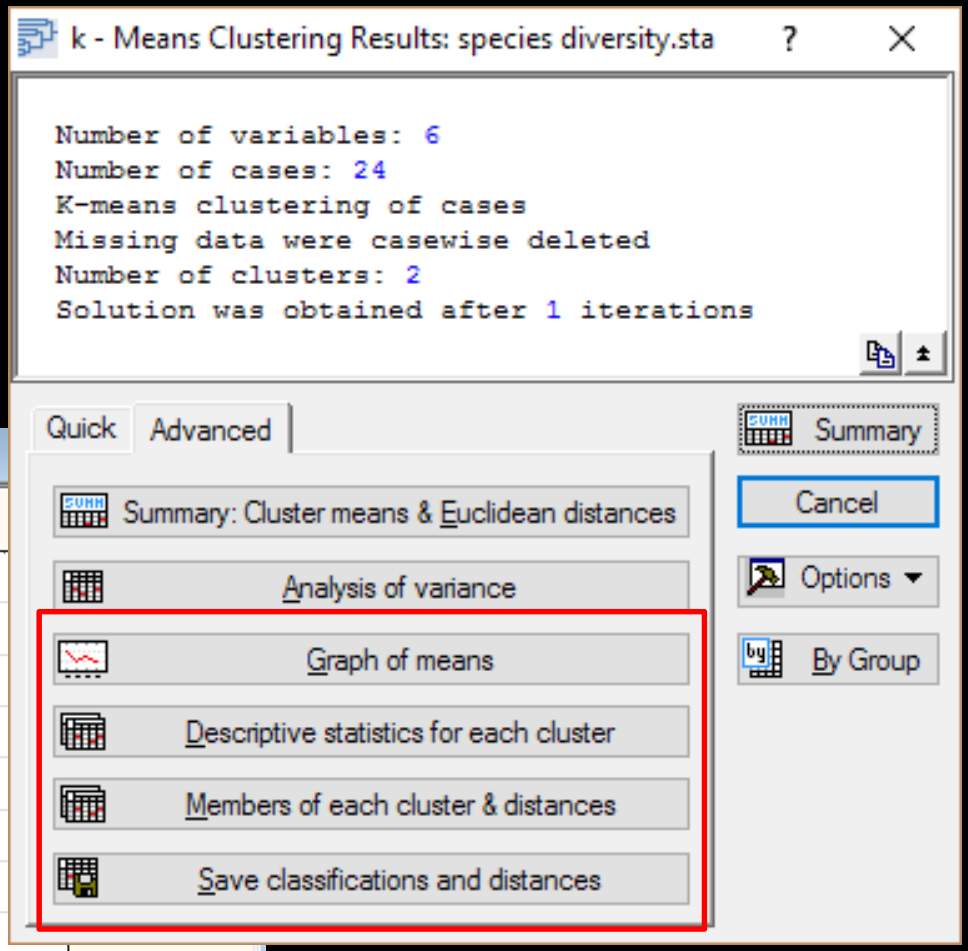
Variable	Analysis of Variance (species diversity.sta) Between SS	df	Within SS	df	F	signif. p
RA_Apes	0,396	1	211,348	22	0,04126	0,840905
RA_Birds	3148,179	1	1828,676	22	37,87437	0,000003
RA_Elephant	0,001	1	10,426	22	0,00117	0,973006
RA_Monkeys	2510,123	1	1012,611	22	54,53497	0,000000
RA_Rodent	16,344	1	33,430	22	10,75572	0,003424
RA_Ungulate	90,482	1	337,412	22	5,89964	0,023768

Можно посмотреть, по каким переменным различия между кластерами достоверны

Кластерный анализ

Видно, по каким переменным различаются кластеры





Можно сохранить в новый файл классификацию объектов в кластеры

Кластерный анализ

linkage	Members of Cluster Number 1 (species diversity and Distances from Respective Cluster Center Cluster contains 13 cases **Distance**
Park	2,558751
Park	6,103410
Park	5,562892
Park	2,495944
Park	5,131427
Park	3,461835
Logging	7,147769
Neither	4,702465
Logging	4,744735
Logging	4,592163
Logging	3,515239
Park	8,829162
Logging	3,827867

Можно посмотреть, какие объекты попали в какой кластер

linkage	Members of Cluster Number 2 (species diversity. and Distances from Respective Cluster Center Cluster contains 11 cases **Distance**
Logging	6,635179
Neither	1,483423
Logging	8,245093
Logging	1,617373
Logging	3,301316
Neither	7,315745
Logging	1,823993
Logging	3,570219
Neither	1,887791
Logging	4,862472
Logging	2,595359

Здесь в первом кластере оказались все трансекты из нацпарков

Для публикаций

В методы: если переменные были стандартизированы – написать; для метода К-средних – написать, какие переменные использовалсь и сколько кластеров получали.

“To classify males according to their behavioral tactics during the mating season, we performed K-means clustering with three variables (the date of spring emergence, home range size and the number of potential female mates) and two clusters. All variables were standardized before clustering.”

В результаты:

✓Для иерархического метода главный результат – картинка с деревом; в подписи – информация о том, какие дистанции использовали, как измеряли расстояния между кластерами.

✓Для метода К-средних – результаты ANOVA сравнения групп; можно – картинку средних в группах; описать состав групп.

Требования к выборкам для проведения кластерного анализа

1. В качестве исходных данных можно взять любую матрицу дистанций – **никаких ограничений и требований** к выборке!
2. Нет необходимости исключать **аутлаеры**;
3. Если переменные измерены в **разных шкалах**, их следует **стандартизировать**.

Кластеры, полученные методом К-средних, можно использовать в дальнейшем анализе (принадлежность к кластеру – группирующая переменная)

Дискриминантный анализ

Кластерный анализ

У нас есть *исходно существующие группы*. Мы ищем переменные, которые лучше всего их разделяют.

У нас есть несколько *переменных*. Мы на основе них хотим классифицировать выборку – проверить, не объединяются ли наблюдения в группы.

Обобщённые линейные модели (Generalized linear models, GLZ) и логистическая регрессия

В **GLM** мы анализировали действие разных (непрерывных и категориальных) предикторов на зависимую переменную.

Зависимая переменная:

- ✓ одна;
- ✓ количественная;
- ✓ распределена нормально.

Но бывают задачи, когда зависимая переменная качественная или по своей сути не может иметь нормального распределения.

GLZ придуманы для анализа действия разных предикторов на зависимую переменную, имеющую строго ***определённое****, но* ***не обязательно нормальное*** *распределение.*

Generalized linear model

Варианты зависимой переменной (основные):

✓**бинарная** (биномиальное распределение);

✓дискретная количественная (распределение Пуассона);

✓нормальное распределение (тоже можно использовать).

*Примеры с **бинарной** переменной:*

✓Как от массы тела, возраста, социального положения, показателей крови матери зависит пол ребёнка?

✓Как масса тела и количество сибсов влияют на вероятность дожить до года у детёныша (переменная выжил/не выжил).

✓Как местообитание, плотность популяции, пол, возраст влияют на заболеваемость Болезнью X? (есть/нет антитела X)

*Примеры с распределением **Пуассона**:*

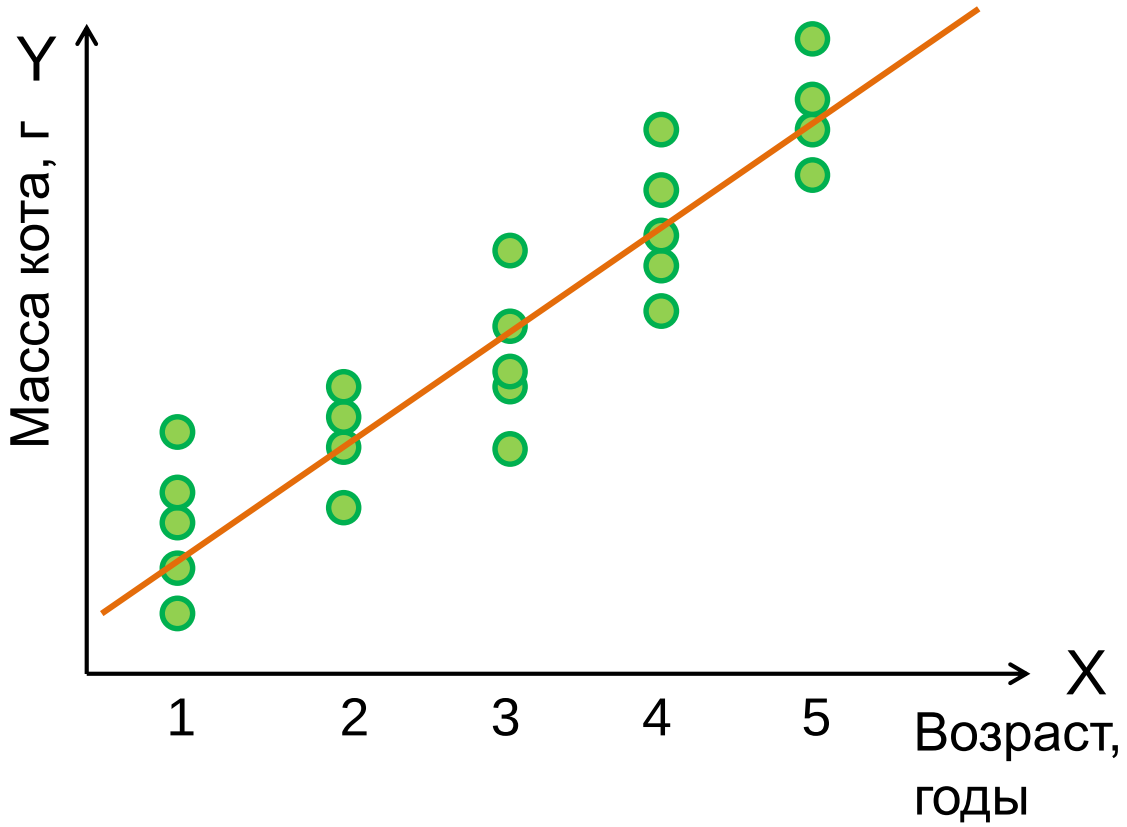
✓Как от pH, освещённости, концентрации разных ионов зависит количество зелёных водорослей в пробе воды?

✓Как тип субстрата влияет на число норок пескожила на 1 м^2?

Generalized linear model

Модель GLM:

$$Y = a + bX + e$$

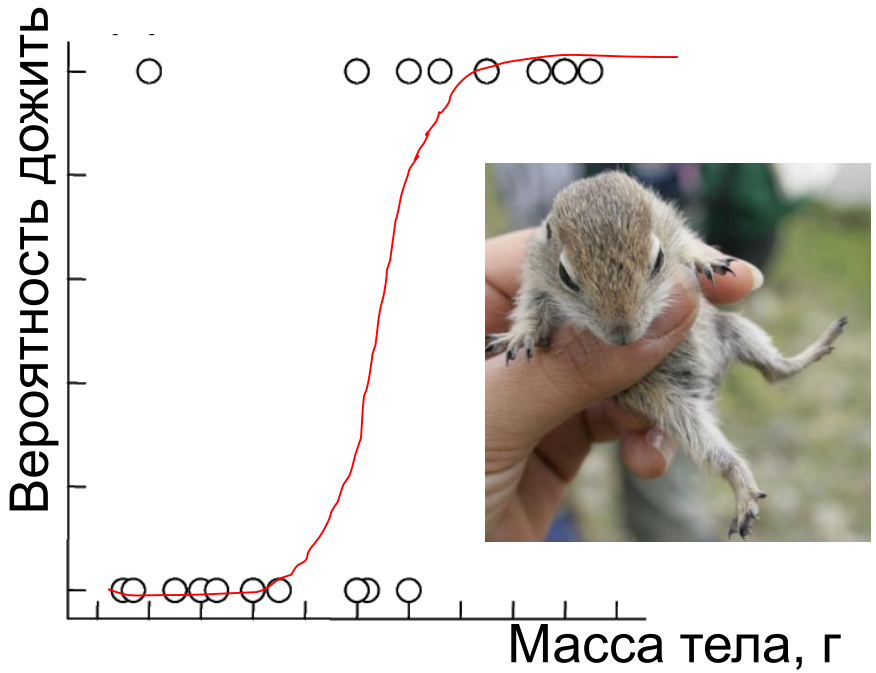


Нормальная линейная связь.
2 составляющие: **зависимая переменная + предикторы** (количественные и категориальные).

Модель GLZ

Пример: проверим, как масса детёныша суслика (в конце лактации) влияет на вероятность дожить до расселения? (умер – 0, дожил - 1)

$$f(Y) = a + bX + e$$



Связь переменных явно **не линейная**.

3 составляющие: **зависимая переменная + предикторы** (количественные и категориальные) **+ ДОПОЛНИТЕЛЬНАЯ ФУНКЦИЯ** (связующая = link function).

Связующая функция нужна чтобы выправить эту кривую зависимость, сделать её линейной (можно было бы просто построить нелинейную зависимость, но анализировать её намного сложнее).

Link functions:

Для нормального распределения – не нужна (в модели выбирается identity link).

Бинарная переменная (биномиальное распределение) – logit link. $f(Y) = \log(Y/(1-Y))$, где Y – вероятность 1.

Дискретная переменная (распределение Пуассона) – log link, $f(Y) = \log(Y)$

В результате получается ЛИНЕЙНАЯ модель, в которой нужно оценить коэффициенты при X (предикторах). Оценка коэффициентов итеративным путем.

Предикторов может быть много.

Тестирование гипотез о коэффициентах -

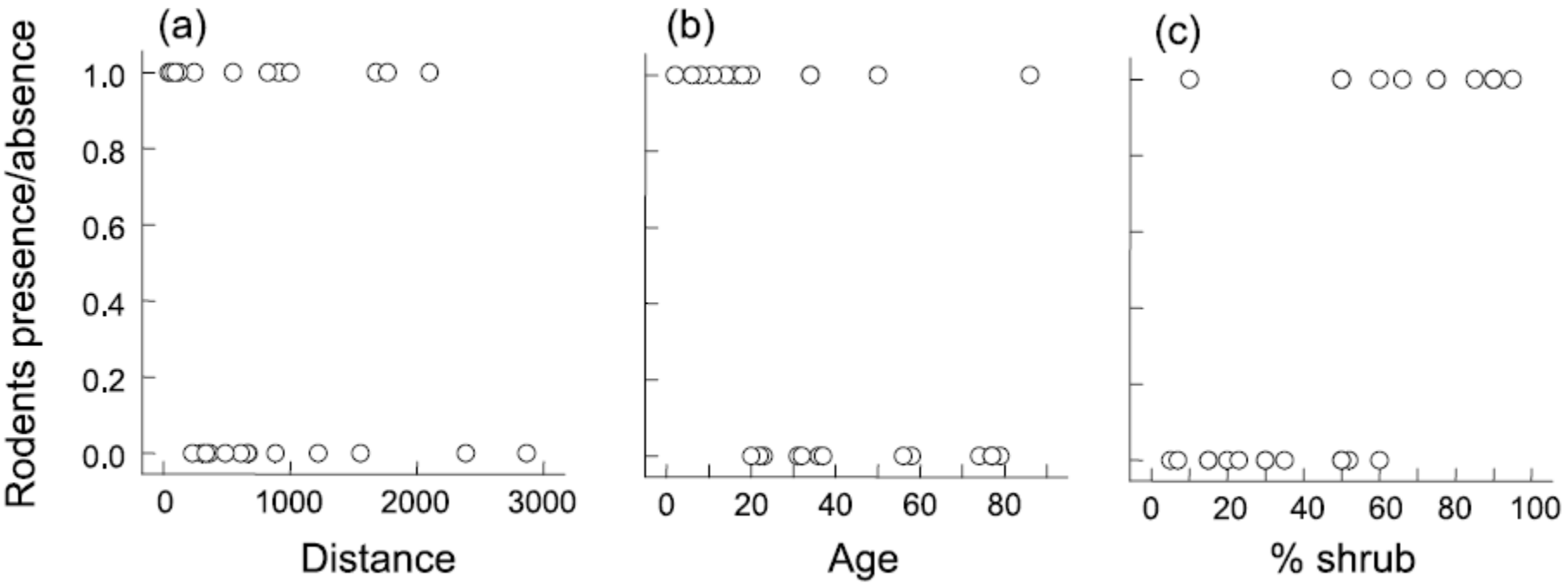
H_0: $\beta = 0$
H_1: $\beta \neq 0$

1. **Тест Вальда** (Wald statistics) – аналог т-теста, основанный на методах максимального правдоподобия.

$$\frac{b_1}{s_{b_1}}$$

2. **Сравнение качества** полной модели и редуцированной (без соответствующего предиктора) на основе residuals (как в линейных моделях) – likelihood ratio test = **G^2 test, статистика хи-квадрат** (тоже методами максимального правдоподобия)

Generalized linear model



В этой модели проверялось влияние расстояния до другого подходящего местообитания, возраста, доли кустарников на присутствие грызунов на данной территории.

Достоверна только связь с)

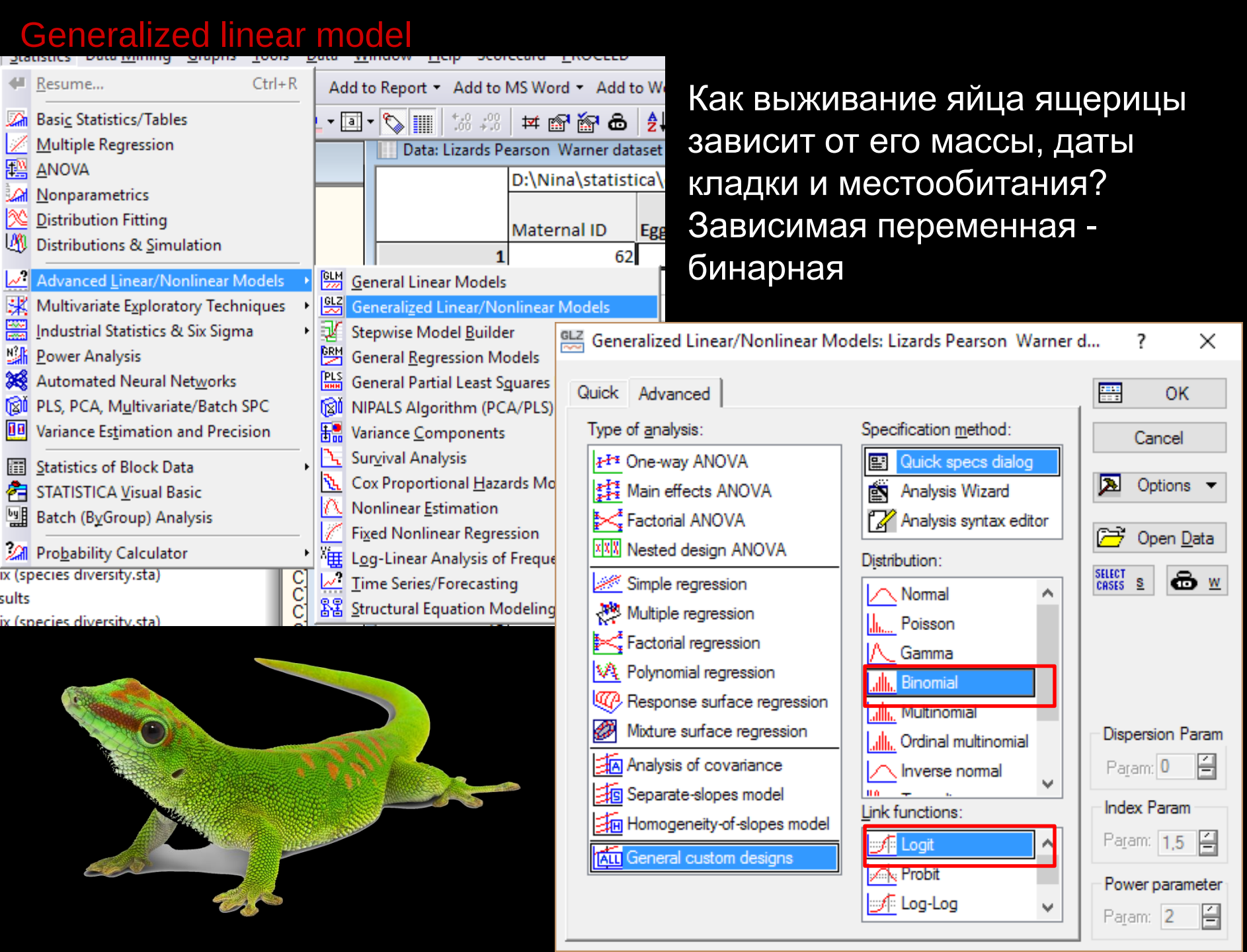
Информационные критерии –
Позволяют выбрать **лучшую модель** из нескольких моделей (например, изо всех возможных сочетаний предикторов).

Лучшая модель та, которая:
✓наилучшим образом описывает изменчивость зависимой переменной;
✓включает наименьшее число факторов.

Критерий Акаике (AIC) – позволяет выбрать лучшую модель: у лучшей модели значение AIC минимально.

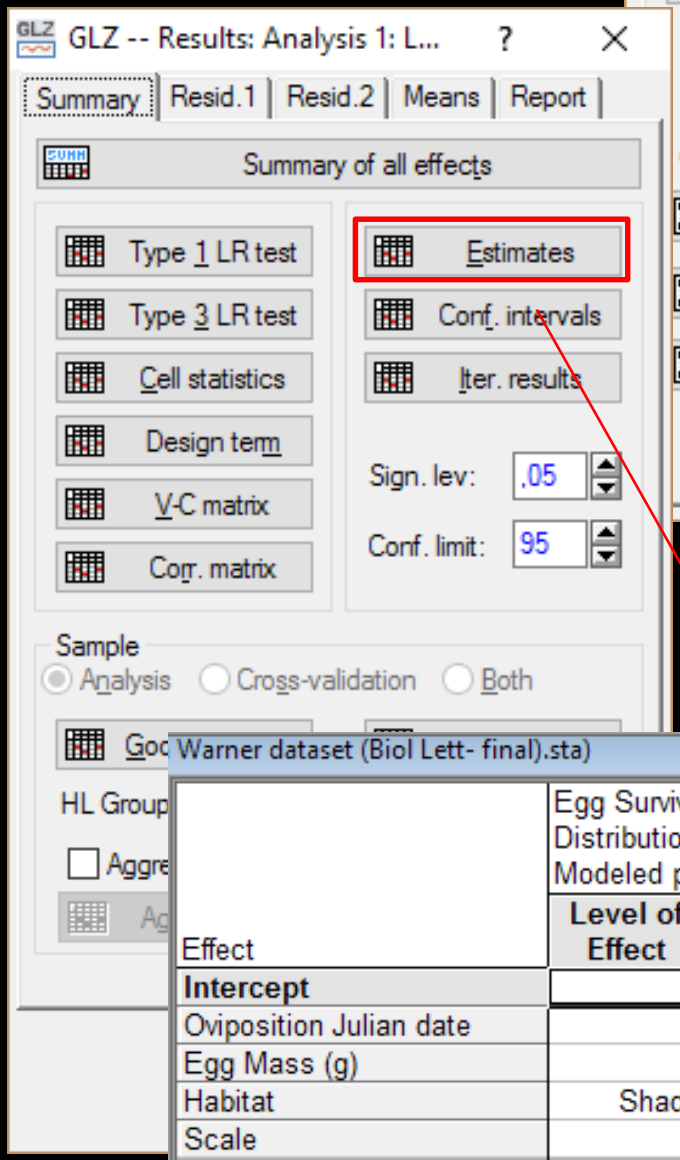
Generalized linear model

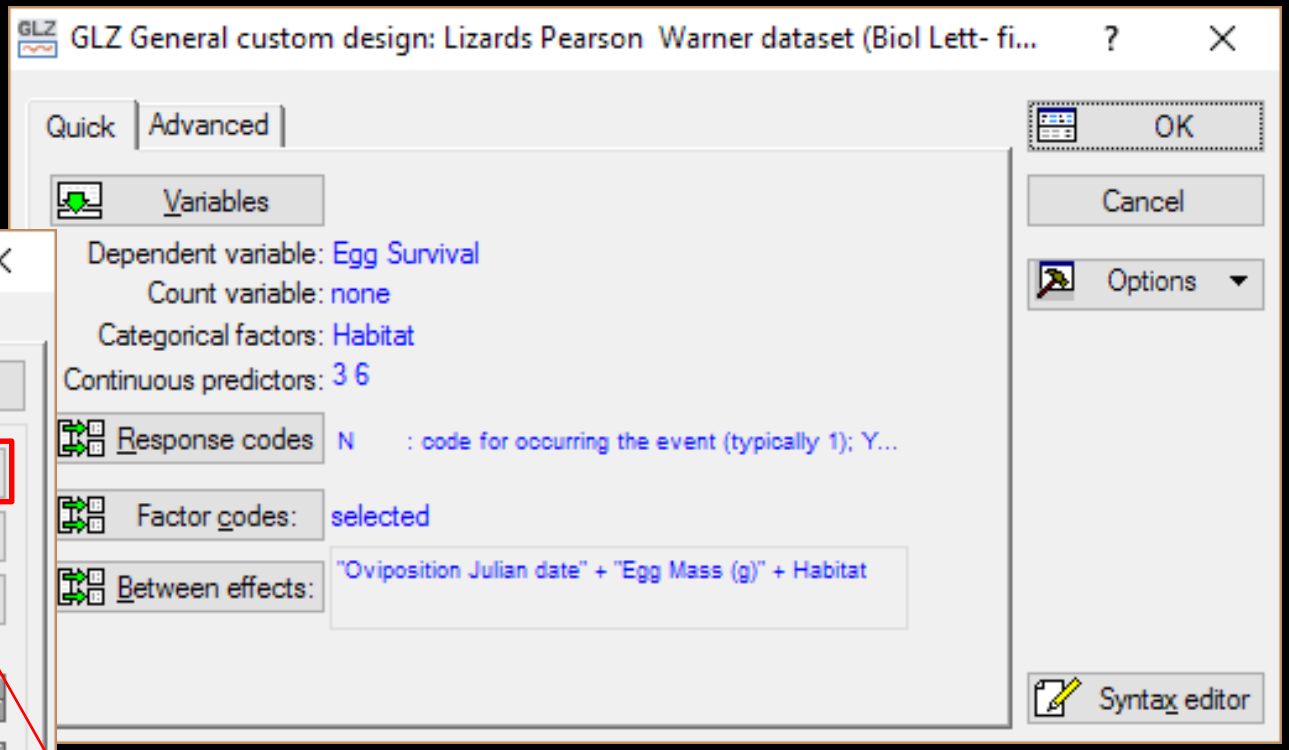
Как выживание яйца ящерицы зависит от его массы, даты кладки и местообитания? Зависимая переменная - бинарная



Generalized linear model

Выбор переменных





В статью – оценки коэффициентов, их ошибки, значения статистики Вальда и *p*

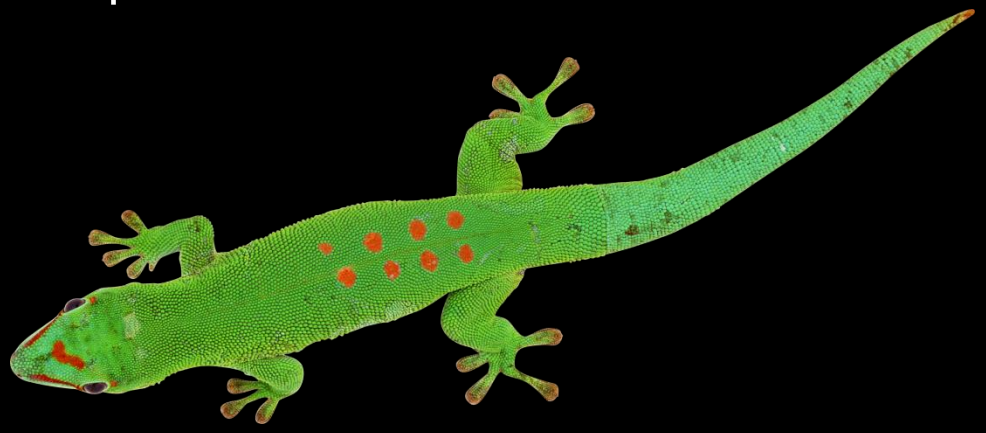
Warner dataset (Biol Lett- final).sta)

Egg Survival - Parameter estimates (Lizards Pearson Warner dataset (Biol Lett- final).sta)
Distribution : BINOMIAL, Link function: LOGIT
Modeled probability that Egg Survival = N

Effect	Level of Effect	Column	Estimate	Standard Error	Wald Stat.	Lower CL 95,0%	Upper CL 95,0%	p
Intercept		1	115,9495	243,8906	0,226020	-362,067	593,9664	0,634490
Oviposition Julian date		2	-0,0086	0,0184	0,218952	-0,045	0,0274	0,639840
Egg Mass (g)		3	-20,1080	10,3138	3,800994	-40,323	0,1067	0,051222
Habitat	Shade	4	-0,0099	0,2580	0,001463	-0,515	0,4957	0,969484
Scale			1,0000	0,0000		1,000	1,0000	

Generalized linear model

Likelihood ratio test – альтернатива статистике Вальда. Предпочтителен для малых выборок! Имеет большую мощность.



GLZ -- Results: Analysis 1: L...

Summary | Resid.1 | Resid.2 | Means | Report

Summary of all effects

Type 1 LR test | Estimates

Type 3 LR test | Conf. intervals

Cell statistics | Iter. results

Design term

V-C matrix

Corr. matrix

Sign. lev: ,05

Conf. limit: 95

Sample

Analysis | Cross-validation | Both

Goodness of fit | Raw data

HL Groups: 10

Overdispersion

Aggregation

Pearson Chi ²

Deviance

Aggreg. data

Modify | Close

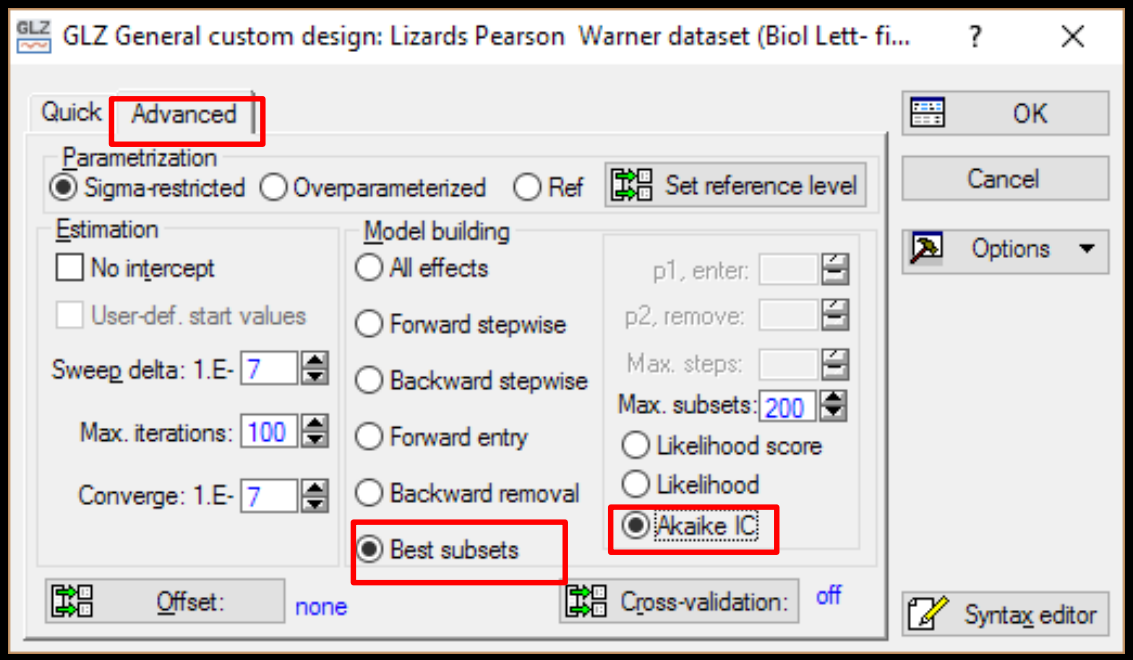
By Group | Options

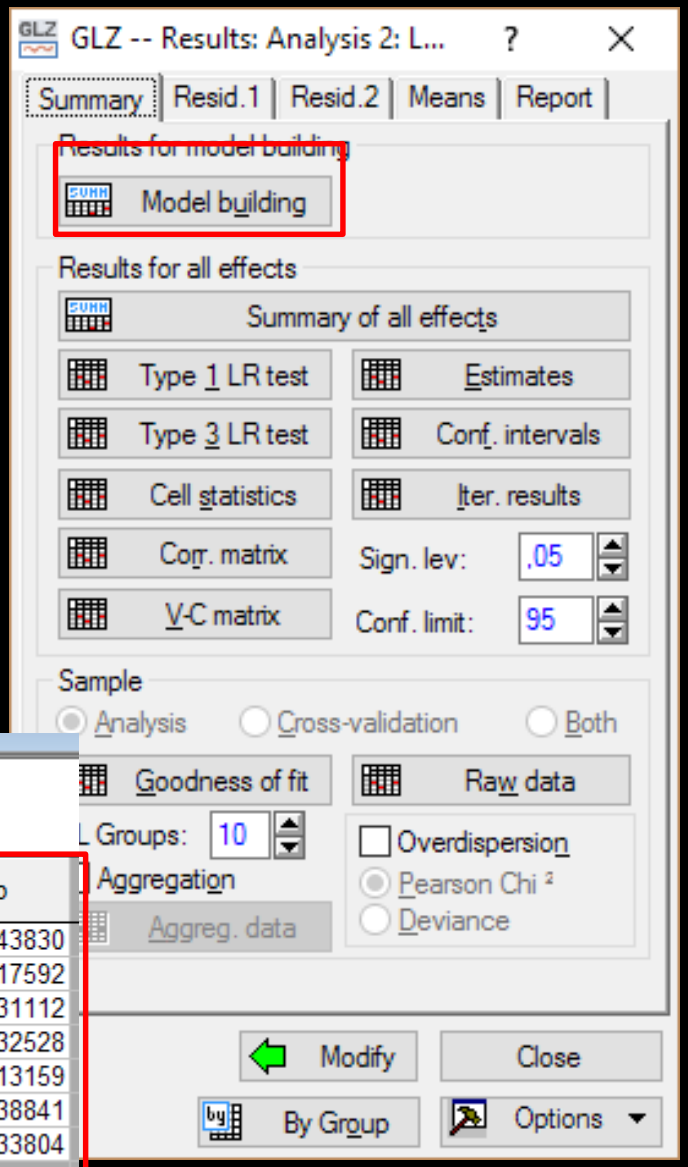
Egg Survival - Likelihood Type 1 Test (Lizards Pears
Distribution : BINOMIAL, Link function: LOGIT
Modeled probability that Egg Survival = N

Effect	Degr. of Freedom	Log-Likelihd	Chi-Square	p
Intercept	1	-51,7499		
Oviposition Julian date	1	-51,7469	0,005887	0,938841
Egg Mass (g)	1	-49,6094	4,275175	0,038673
Habitat	1	-49,6086	0,001464	0,969480

Generalized linear model

Критерий Акаике и выбор лучшей модели





Egg Survival - Model building results (Lizards Pearson Warner dataset (Biol Lett- final).sta)
Distribution : BINOMIAL, Link function: LOGIT
Modeled probability that Egg Survival = N (Analysis sample)

	Var. 1	Var. 2	Var. 3	Degr. of Freedom	AIC	L.Ratio Chi?	p
1	Egg Mass (g)			1	103,436662	4,063113	0,043830
2	Oviposition Julian date	Egg Mass (g)		2	105,218713	4,281062	0,117592
3	Egg Mass (g)	Habitat		2	105,436362	4,063412	0,131112
4	Oviposition Julian date	Egg Mass (g)	Habitat	3	107,217249	4,282526	0,232528
5	Habitat			1	107,364633	0,135142	0,713159
6	Oviposition Julian date			1	107,493888	0,005887	0,938841
7	Oviposition Julian date	Habitat		2	109,362798	0,136977	0,933804

В лучшую модель вошла только масса яйца. В статью – вся табличка

Generalized linear model

Для публикаций

В методы: использовали обобщённые линейные модели для оценки влияния А на Б. Значимость эффектов оценивалась на основе статистики Вальда (или теста отношения правдоподобия, Likelihood ratio test). Выбор лучшей модели осуществлялся с помощью информационного критерия Акаике.

В результаты:

✓Можно привести просто коэффициенты, их ошибки, статистику Вальда (или хи-квадрат из Likelihood ratio test) и p.

✓Если проводим процедуру выбора лучшей модели, приводим табличку с моделями, значениями минимального AIC и для каждой модели – разницы ее AIC с минимальным; df; p не обязательно.

Требования к выборкам для проведения GLZ

1. Нельзя брать какую попало зависимую переменную, её распределение должно быть известно. (проще всего с бинарными и нормально распределенными)
2. Следует исключить предикторы, сильно коррелирующие с другими.
3. Размер выборок: как и для АНОВы, в группах не должно быть меньше 10 измерений. Чем больше количественных переменных, тем больше выборка.

К практическому занятию

MDS – Cars
Cluster – Cars
GLZ - Crabs