

Практическая работа № 1

Исследование морфологической изменчивости в природных популяциях на примере нематоды *Hysterothylacium incurva* – паразита меч-рыбы - методом регрессионного анализа.

При создании теории естественного и искусственного отбора Дарвин много внимания уделял исследованию индивидуальной изменчивости, считая её основным материалом для отбора. В современных исследованиях изменчивости, особенно внутри- и межвидовой, применяются достаточно строгие методы анализа. Одним из них является регрессионный анализ.

У животных в процессе онтогенеза рост органов часто непропорционален росту тела и сам по себе подвержен отбору. Неправильный учёт относительных размеров количественных признаков может приводить к ошибкам в диагностике.

Файл ***incurva.xls*** содержит данные промеров внутренних органов 90 экз. нематод *Hysterothylacium incurva*, от ранних личинок третьей стадии до крупных половозрелых особей. Эти черви обитают в кишечнике рыб, питающихся кальмарами: меч-рыбы (*Xiphias gladius*), марлинов (*Tetrapturus* spp.) и змеевидных макрелей (*Gempylus* spp.).

Регрессионный анализ

С чем работает регрессионный анализ. Задачей регрессионного анализа является построение формулы некоторой линии, адекватным образом проходящей через облако точек на графике. "Адекватность" подразумевает согласие с некоторым критерием, *моделью* или *минимизирующей функцией* (*model, loss function*). Обычно (и по умолчанию) таким критерием является *метод наименьших квадратов* (*least square*). Его использование подробно освещается в специальной литературе.

Разновидности регрессионного анализа. Также по умолчанию обычно принимается, что линия регрессии - это геометрическое место средних *X* при заданном *Y*, однако существуют методы регрессии (*непараметрическая регрессия*), описывающие положения не средних, а медиан.

Различают *простую* (для описания связи между двумя случайными величинами) и *множественную* (одна случайная величина зависит от нескольких других), а также *линейную* и *нелинейную* регрессии. Линейная регрессия применяется тогда, когда через облако точек можно провести прямую линию. При использовании множественной регрессии возможны неточности метода, связанные с тем, что независимые переменные сильно скоррелированы друг с другом. Например, вес человека можно считать функцией от роста и окружности талии, но обе этих величины тесно связаны. В таких случаях желательно применение *гребневой регрессии* (другое название - *ридж-регрессия*, от англ. *ridge regression*).

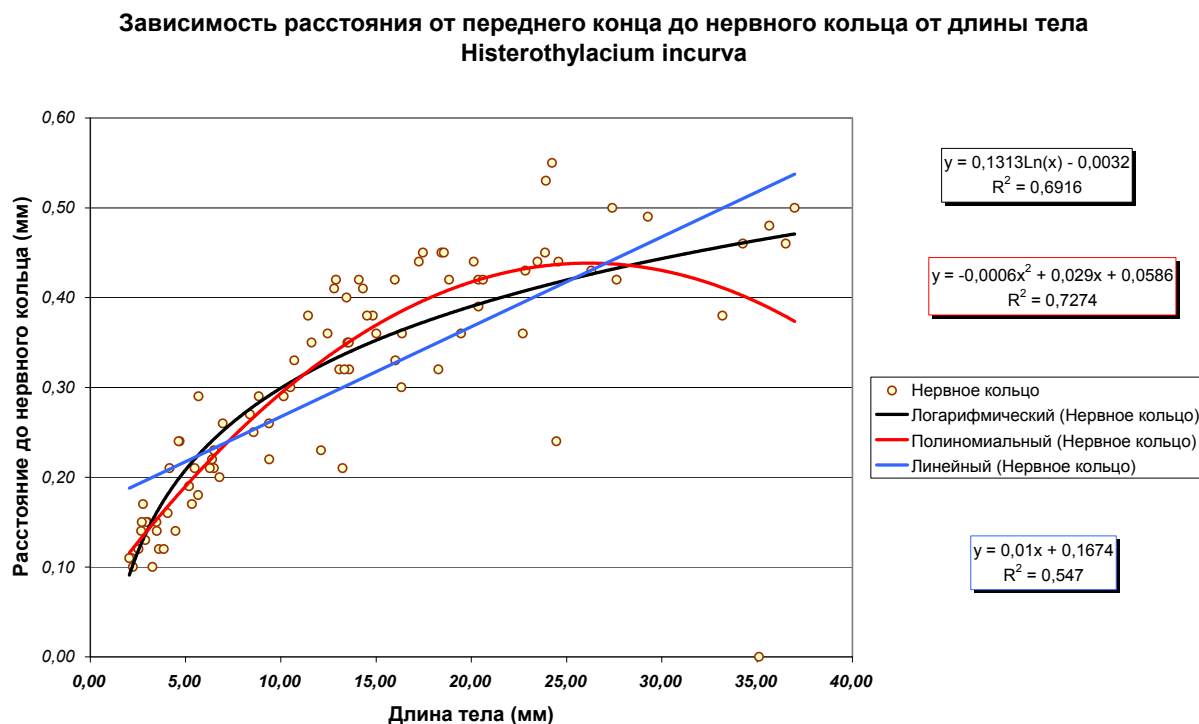
Отдельно следует рассматривать пошаговую регрессию. По существу, метод пошаговой регрессии применяется для удаления малоинформативных показателей, которые сильно коррелируют с более значимыми переменными.

Нелинейный регрессионный анализ допускает наличие не одной, а нескольких типов функций, приблизительно одинаково точно описывающих облако точек. В таком случае необходим критерий, по которому можно выделять наиболее предпочтительные типы функций. Обычно в этом качестве выступает *корреляционное отношение R^2 (R-square)* и уточненный R^2 (*adjusted R-square*). **Обычный способ исследования - рассчитать регрессию для нескольких функций и выбрать ту, у которой R^2 наибольший.**

Регрессионный анализ обычно рассчитывается в специальных статистических программах, но простейшие результаты можно получить в ЭТ Excel.

Задание 1. Получение кривых регрессии.

Сделайте *точечный* график зависимости некоторой количественной переменной (напр. **Нервное кольцо**) от длины тела *на отдельном листе*. В верхнем меню *Диаграмма* выберите опцию *Добавить линию тренда...* В появившемся диалоговом окне укажите подходящий тип графика, во вкладке **Параметры** проставьте галочки в окнах *Показывать уравнение на диаграмме* и *Поместить на диаграмме величину достоверности аппроксимации (R^2)*¹. Желательно повторить эту операцию с разными функциями. Результат может быть примерно таким:



¹ Студенты часто забывают сразу выполнить эту операцию, но её можно сделать и позже, вызвав диалоговое окно двойным щелчком по линии регрессии.

Как видно из данного графика, линейное уравнение показывает сравнительно малый R^2 , следовательно, должно быть отклонено. Более простое двучленное логарифмическое уравнение $y = 0,1313\ln(x) - 0,0032$ имеет почти тот же R^2 , что и трёхчленное степенное, и, вследствие своей простоты, является более предпочтительным².

Задание 2. Исследование остатков.

График даёт самое грубое представление о том, как сильно варьирует исследованный признак. Поскольку точки кривой регрессии обычно рассматриваются как среднее значение Y при данном значении X , зачастую требуется оценить, насколько размерный признак стабилен, т.е., отклоняется от среднего значения. Разность между конкретным значением и его теоретической средней, рассчитанной по кривой регрессии $y_{ост} = (y_i - y_{cp})$, называется *остатком*. Остаток, делённый на среднюю, будем называть относительным остатком.

Ш	С	Ш	С	Р
Длина тела	личинка	самец	самка	пол
	0,40			
13,07	0,22			
12,11	0,25			
12,90	0,25			
8,85	0,44			
8,38	0,21			
9,40	0,33			
8,58	0,17			
10,15	0,13			
10,71	0,09			
6,49	0,14			
6,40	0,11			
5,34	0,08			
5,19	0,11			
5,66	0,16			
6,78	0,27			
6,49	0,30			
6,96	0,13			
5,69	0,20			
5,48	0,15			
3,92	0,11			
4,69	0,12			
4,48	0,12			
2,68	0,07			
3,47	0,08			
2,18	0,06			
3,60	0,07			
2,89	0,09			
2,95	0,11			
2,95	0,05			
4,16	0,09			
4,07	0,13			
2,77	0,11			
2,54	0,08			
2,71	0,07			
2,24	0,08			
3,27	0,10			
3,49	0,13			
3,86	0,11			
4,84	0,10			
20,37	0,12			
20,61	0,20			
19,46	0,16			
18,84		0,15		
18,56		0,17		
18,27				
15,02		0,21		
16,36		0,14		
13,25		0,15		
14,83		0,14		
14,10		0,11		
13,45		0,17		
13,57		0,16		
12,46		0,13		
10,51		0,13		
11,62		0,18		
11,44		0,16		
22,71		0,13		
27,40		0,24		
22,84		0,17		
15,99			0,48	
18,44			0,48	
17,25			0,35	
17,47			0,68	

Остаток, делённый на среднюю, будем называть относительным остатком.

Введите формулу регрессии³ и раскопируйте её для всех значений y_{cp} . В следующей колонке проставьте абсолютные значения остатков, затем - относительные. Рассчитайте дисперсию, минимум и максимум для относительных остатков.

Задание 3. Исследование половых различий.

Хвост у самок нематод данного отряда относительно длиннее, чем у самцов. В ЭТ Excell это можно отобразить следующим образом:

- выделите колонки **Длина тела**, **Длина хвоста** и **Пол**, скопируйте их в новый лист и ранжируйте по колонке **Пол**;

- данные из колонки **Длина хвоста** разнесите на три колонки, которые озаглавьте **Личинки**, **самцы**, **самки** как на рисунке слева;

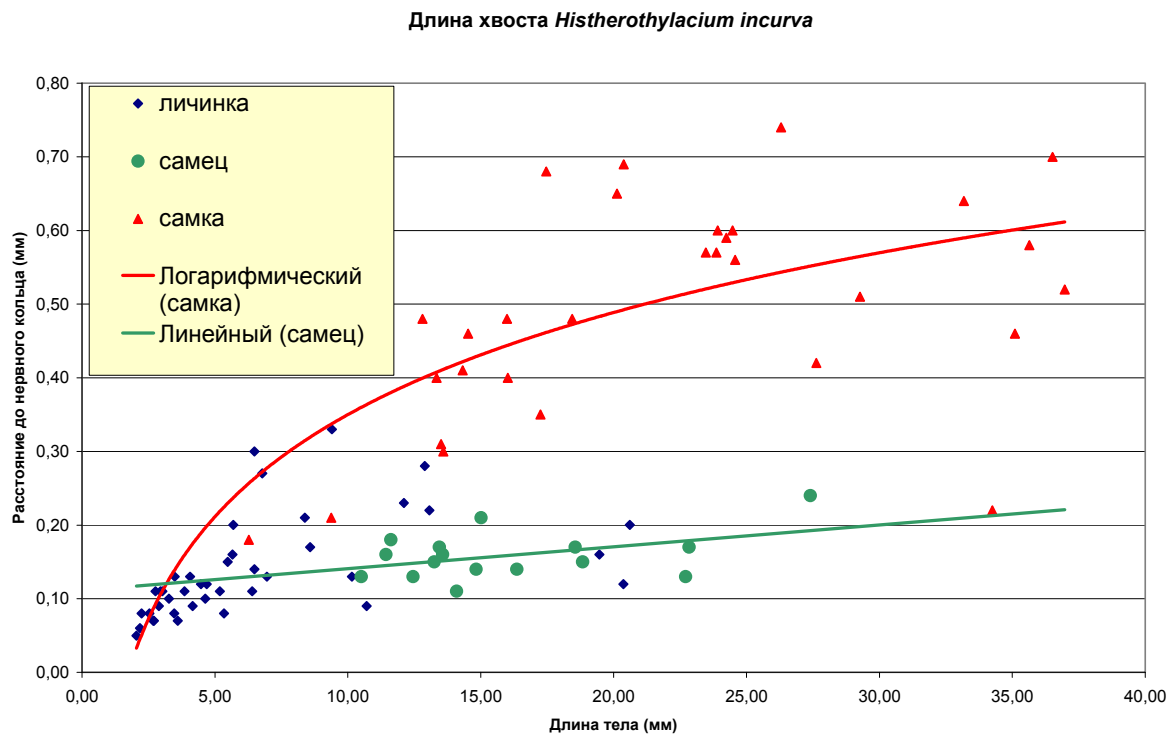
- постройте точечный график зависимости размеров хвоста от длины тела, выделив четыре колонки (без колонки **Пол**)

- щелчком ко точке определённого типа постройте линии регрессии для самцов и для самок.

Конечный график может выглядеть так:

² У самых разных животных нелинейный рост органов чаще всего описывается логарифмическим уравнением этого типа.

³ Её параметры можно получить копированием выделенного текста уравнения регрессии непосредственно с графика. Вместо x вставьте адрес соответствующей ячейки и добавьте перед логарифмом знак умножения.



Сравните дисперсии относительных остатков у самцов и самок.