# ОБРАЗОВАНИЕ ДЛЯ УСТОЙЧИВОГО РАЗВИТИЯ

Научная статья
УДК 57.081.23
DOI: 10.18384/2712-7621-2025-3-164-185

## АЛГОРИТМ ПРИМЕНЕНИЯ ЛОГИСТИЧЕСКОЙ МОДЕЛИ ДЛЯ АНАЛИЗА БИНАРНЫХ ДАННЫХ В ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

**© CC BY *Захаров К. В.[1], Коновалов А. М.[2], Ломсков М. А.[3]***

*[1] Московская государственная академия ветеринарной медицины и биотехнологии – МВА имени К. И. Скрябина*
*г. Москва, Российская Федерация*
*e-mail: k.v.zaharov@gmail.com, ORCID: 0000-0002-1620-3895*

*[2] Московская государственная академия ветеринарной медицины и биотехнологии – МВА имени К. И. Скрябина*
*г. Москва, Российская Федерация*
*e-mail: zoolog82@mail.ru, ORCID: 0000-0002-4050-0259*

*[3] Московская государственная академия ветеринарной медицины и биотехнологии – МВА имени К. И. Скрябина*
*г. Москва, Российская Федерация*
*e-mail: lomskovma@mail.ru, ORCID: 0000-0001-6579-0048*

Поступила в редакцию 13.06.2025
После доработки 04.08.2025
Принята к публикации 25.08.2025

***Аннотация***

**Цель.** Разработать алгоритм применения *GLM* (*Generalized Linear Model*) для бинарных данных на всех этапах анализа, доступный для обычных пользователей без специальных знаний в области программирования.

**Процедура и методы**. В качестве основы выбран широко распространённый алгоритм статистического анализа, включающий 3 основных этапа: формулирование исследовательской гипотезы и сбор материала, его исследование, а также подгонку и проверку статистической модели. Для примера выбраны литературные данные о лесотаксационных характеристиках насаждений дуба черешчатого (*Quercus robur L.*) на пробных площадях в 15 экорегионах. Смоделировано среднее значение диаметра ствола дуба в зависимости от возраста, высоты, числа стволов, а также географического положения пробной площади. Поскольку характер зависимой переменной не позволяет выбрать классический регрессионный анализ, с помощью логистической регрессии смоделирована вероятность превышения порогового значения диаметра ствола в 30 см. В оценке качества модели использованы такие показатели как разница девиансов, т. е. отличия между зависимой переменной и её предсказанным значением (остатки) и вычисляемые для бинарных моделей их взвешенные значения (т. н. квантильные остатки), непараметрические тесты в сравнении вложенных моделей и потенцирование параметров модели.

**Результаты.** Показана зависимость диаметра ствола от возраста деревьев и числа стволов, а также широты расположения пробной площади. Существующий алгоритм регрессионного анализа дополнен процедурой анализа мощности и оценкой предсказания зависимой переменной с использованием каппы-коэффициента и *ROC*-кривых. Предложенный алгоритм позволяет смоделировать переменные, не отвечающие требованиям линейной регрессии, сравнить полученные модели, оценить качество прогноза и размер необходимой выборки.

**Теоретическая и/или практическая значимость**. Работа имеет методическую направленность. Показано, что за исключением оценки минимального объёма выборки, для работы с логистическими моделями достаточно функций, встроенных в ядро среды R и установки некоторых пакетов.

***Ключевые слова:*** GLM, бинарные данные, количественные методы, методика экологического исследования, RStudio, дуб черешчатый (*Quercus robur L.*)

***Для цитирования:***
Захаров К. В., Коновалов А. М., Ломсков М. А. Алгоритм применения логистической модели для анализа бинарных данных в экологических исследованиях // Географическая среда и живые системы. 2025. № 3. С. 164–185. DOI: 10.18384/2712-7621-2025-3-164-185

Original Article

# THE ALGORITHM OF USING THE LOGISTIC MODEL FOR BINARY DATA ANALIZIS IN ECOLOGY INVESTIGATIONS

**© CC BY K. Zakharov[1], A. Konovalov[2], M. Lomskov[3]**

*[1] Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K. I. Skryabin Moscow, Russian Federaton*
*e-mail: k.v.zaharov@gmail.com, ORCID: 0000-0002-1620-3895*

*[2] Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K. I. Skryabin Moscow, Russian Federaton*
*e-mail:zoolog82@mail.ru, ORCID: 0000-0002-4050-0259*

*[3] Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K. I. Skryabin Moscow, Russian Federaton*
*e-mail:lomskovma@mail.ru, ORCID: 0000-0001-6579-0048*

Received 13.06.2025
Revised 04.08.2025
Accepted 25.08.2025

***Abstract***
**Aim.** The *GLM* (*Generalized Linear Model*) algorithm has been created for working with binary data on all analysis steps, which would be possible for all users without special IT skills. **Methodology.** We used wide distributed algorithm statistically analysis which includes three main steps: biology hypnotizes formulation and data collecting, investigation of data, fitting and checking finally models. As data example were choose forest taxation features of Oak petiolate (*Quercus robur L.*) from 15 ecoregions. The average trunk diameter has been simulated according to age, high, trunk number and geographically location of sample area. Because features of dependent variable didn't allow choosing classical regression analysis we simulated of probability the excess of 30-cm threshold of trunk diameter with logistic regression using. For quality assessment of models we used different deviances, i.e. difference between depended variable and predicted values (residuals), calculated specific quantile residuals for *GLM*, nonparametric tests for nested models and model parameters potentiating. **Results.** There was showed the dependence trunk diameter on age and trunk number, as well as latitude of sample area. The existing algorithm has been supplemented by power test and assessment of prediction independent variable by kappa-coefficient and *ROC*-curves. The new algorithm allows simulating variables which don't fit to demands of lineal regression, to comparison models, to assessment the quality of prognosis and a volume of minimal sample. **Research implications**. This work has a methodic direction. There has been showed for logistic models creation there are enough functions from RStudio core excluding the minimal sample assessment.

***Keywords:*** GLM, binary data, quantitative methods, ecology investigation methods, RStudio, Oak petiolate (*Quercus robur L.*)

***For citation:***
Zakharov K., Konovalov A., Lomskov M. The algorithm of using the logistic model for binary dataanalizis in ecology investigations. In*: Geographical Environment and Living Systems*, 2025, no. 3, pp. 164–185. DOI: 10.18384/2712-7621-2025-3-164-185

## ВВЕДЕНИЕ

Регрессионный анализ, один из наиболее распространённых статистических методов [11], прочно вошёл в инструментарий экологов, как теоретиков, так и практиков. Использование регрессионного анализа позволяет оценить влияние одного признака на другой, т. е. предполагается, что зависимая переменная представляет собой функцию одной или нескольких независимых переменных [3]. Результатом такого анализа становится математическая модель изучаемой системы, которая с определёнными допущениями даёт описание этой системы и позволяет построить прогноз [5]. Внимание к математическим моделям в экологии не случайно, поскольку такие методы позволяют описать сложные природные [11; 12] или социально-природные системы [1; 20].

Наиболее простая линейная регрессия подразумевает, что отношения между зависимой и независимой переменными напоминают линейную функцию и используют общую линейную модель (*General Lineal Model* или LM). Это т. н. параметрический метод, который, однако, предъявляет к данным определённые требования. LM применима к интервальным непрерывным данным, которые не всегда удаётся получить в экологических исследованиях [12]. Нередко приходится иметь дело с разнообразными индексами, шкальными данными (например, оценка степени рекреационного нарушения в баллах), долями, счётными (число встреч) или бинарными данными (самец, самка). Применить общую линейную регрессию к таким данным нельзя. В этом случае возможны два пути. Первый вариант – перейти к аппроксимирующей функции непрерывного аргумента или же использовать непараметрические тесты, к каковым и относят обобщённые линейные модели *GLM* (Generalized *Linear Model* или *GLM*), семейство непараметрических линейных моделей, где зависимость между переменными сохраняет линейный характер, но данные не отвечают требованиям LM [12].

Работа с *GLM* несколько отличается от привычных линейных моделей, поэтому лучше использовать специальное программное обеспечение. Мы рассмотрим регрессионный анализ в *RStudio*[1] – далее R – свободно распространяемой и весьма популярной в научном мире среде разработки [8; 13].

Конечно, наша статья далеко не первая, накопилась значительная библиография по использованию *GLM*[2]. Однако работа с такими источниками может быть сопряжена с затратами времени или же требовать специальной подготовки, в т. ч. понимания кодов, часто довольно громоздких. Краткие же руководства, например, заметки в интернете, отрывочны и обычно не содержат некоторые важные этапы анализа [4].

По этой причине мы поставили задачу *разработать алгоритм применения GLM для бинарных данных на всех этапах анализа, доступный для обычных пользователей без специальных знаний в области программирования.*

Важно понимать, что моделирование – это именно многоэтапный процесс, причём каждый из этапов должен быть проведён корректно [22].

## ОПИСАНИЕ ЛОГИСТИЧЕСКОЙ МОДЕЛИ

Семейство *GLM* состоит из нескольких моделей, которые позволяют аппроксимировать данные, соответствующие, например, распределению Пуассона или биномиальному распределению. *GLM* были предложены, в первую очередь, именно для исследований живой природы, где данные часто не соответствуют требованиям *LM* [5; 6]. Принципиальное отличие *GLM* от *LM* заключается в использовании другого алгоритма – метода максимального правдоподобия (ММП) – метода, максимизирующего вероятность нахождения неизвестных параметров, а не метода наименьших квадратов (МНК). В целом

[1] R Core Team.R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. 2023 [Электронный ресурс]. URL: https://www.R-project.org (дата обращения: 19.06.2025).

[2] См. список монографий на сайте http://www.statsci.org/glm/bibliog.html (дата обращения: 19.06.2025).

ММП заключается в поиске функции с такими параметрами, которые с наибольшей вероятностью соответствуют неизвестным параметрам [10]. В данной работе мы ограничимся моделированием бинарных данных, т. е. данных, принимающих лишь два значения, например, выжил или нет.

В общем, методы регрессии направлены на поиск параметров, связывающих зависимую переменную *y*, и независимую переменную, или предиктор, *x*. Проблема заключается в том, что в случае с бинарными данными *y* принимает только два значения, а не лежит в области действительных чисел. Следовательно, необходим какой-то оригинальный подход, и на помощь приходит теория вероятности.

Для дальнейших рассуждений введём понятие «испытание», а именно некий опыт, который может закончиться одним из двух элементарных и взаимно противоположных событий, например, выживет популяция или нет. Такие события получили условные названия «успех» (*p*) или «неуспех» (*q*) и связаны отношением *p+q=1*. Значения как *p,* так и *q* находятся в диапазоне от 0 до 1. Логистическая регрессия моделирует именно вероятность успешного события *p,* когда *y*=1. Конечно, смоделировать параметр *p* непросто, однако очевидно, если вероятность успеха мала то *p* близко к 0, если же высока, то к 1.

Рассмотрим такой показатель как отношение шансов, который можно выразить как отношение успеха и неуспеха *p/q* или *p/(1-p)*[3].

Таким образом, можно от бинарных данных перейти к значениям вероятностей, которые занимают диапазон (0, 1). Если же отношение шансов логарифмировать по основанию *e*, то можно получить значения в диапазоне (-∞, ∞). Выражение ln{p/(1-p)} называется также *логит*; заметим, что logit (0.5) = 0, logit (0) → -∞, а logit (1) → ∞. Такая процедура логистической трансформации и дала название методу логистической регрессии [6; 11].

Уравнение логистической регрессии имеет вид:

$$logit(p_i) = ln\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \quad (1)$$

*где:*

$p_i$ – вероятность успешного исхода i-го события, когда *y*=1,

$x_{ki}$ – значение k-ой независимой переменной для i-го наблюдения,

$\beta_0$ – значение *y* при нулевом значении всех независимых переменных,

$\beta_k$ – параметр, или регрессионный коэффициент для k-й независимой переменной.

Если *logit* $(p_i)$ обозначить как $\eta_i$, то:

$$\eta_i = \sum_{i=0}^{k} \beta_j \, x_{ji} \quad (2)$$

где $x_{0i} = 1$, тогда получаем:

$$p_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \quad (3)$$

Таким образом, логистическое преобразование позволяет смоделировать бинарные переменные.

## ПОСТРОЕНИЕ СТАТИСТИЧЕСКОЙ МОДЕЛИ

Работа со статистическими моделями в экологических исследованиях включает три основных этапа[4] [22]:

1. формулирование исследовательской гипотезы и сбор данных;
2. исследование данных;
3. подгонку и проверку статистической модели.

*Первый этап* требует от исследователя понимания изучаемого вопроса, соответствующей теоретической подготовки и работы с литературой.

Ниже рассмотрим влияние долготы и широты места произрастания деревьев на диаметр ствола. Очевидно, что бесперспективно изучать влияние диаметра ствола на географическую долготу.

[3] Бослаф С. Статистика для всех. М.: ДМК Пресс, 2017. 586 с.

[4] Smith C., Warren M. GLMs in R for Ecology. Independently published, 2019. 79 p. [Электронный ресурс]. URL: https://irep.ntu.ac.uk/id/eprint/37478/1/14596_Smith.pdf. (дата обращения: 12.02.2025).

*Второй этап* – исследование данных, включает следующую последовательность поиска и оценки:

- выбросов в переменных;
- нормальности и однородности зависимой переменной;
- превышения нулевых ответов в зависимой переменной;
- мультиколлинеарности независимых переменных;
- отношений между зависимой и независимыми переменными;
- взаимной независимости наблюдений, когда одно наблюдение не оказывает влияния на другое.

*Третий этап* состоит в подгонке и проверке модели, а также её использования для оценки значимости предикторов или же для прогноза [3; 12; 16].

Рассмотрим, как построить логистическую модель в R в соответствии с протоколом [22]. При этом могут быть использованы как бинарные, так и не бинарные данные, которые впоследствии будут перекодированы. Мы расширили протокол исследования и включили процедуру анализа мощности.

В качестве источника данных мы выбрали дендрологические материалы [17]. Данные получены из 15 экорегионов Евразии, включающих хвойные и широколиственные леса, и содержат лесотаксационные характеристики хвойных и лиственных насаждений. Для примера мы выбрали дуб черешчатый (*Quercus rubur L.*) и экспортировали данные в таблицу d. Попытаемся смоделировать средние значения диаметра ствола (*DBH*) дуба на пробных площадях в разных регионах. Оценим, могут ли влиять на диаметр ствола такие переменные, как средние значения возраста (*Tree_age*), средняя высота деревьев (*Htree*), число деревьев на гектар (*Tree_number*), широта (*Latitude*) и долгота (*Longitude*) местоположения пробной площади, а также экорегион (*Ecoregion*) в соответствии с классификацией биомов из работы [18]. Это в основном непрерывные переменные, которые могут принимать любое значение в пределах некоего диапазона. Число деревьев на гектар это счётная, а *Ecoregion* – категориальная или факторная переменная, где регионы зашифрованы сочетаниями цифр, поэтому через команду *as.factor()* укажем, что *Ecoregion* – это именно факторная переменная.

```
d$Ecoregion<- as.factor(d$Ecoregion)
```

Знак $ разделяет название таблицы и колонки в этой таблице.

Использование функции *is.factor()* подтверждает результат.

```
is.factor(d$Ecoregion)
[1] TRUE
```

На втором этапе мы исследуем данные в соответствии с протоколом моделирования [22] и обоснуем целесообразность применения к ним непараметрических методов.

В соответствии с приведённым выше протоколом исследования *оценим выбросы*, т. е. те наблюдения, значения которых сильно отличаются от других[5]. В *R* выбросы определяются как значения в 1,5 раза превышающие межквартильный размах [6; 12]. На рис. 1 показаны «ящики с усами» для четырёх переменных, построенных с использованием команды *boxplot()*. Значения географических координат (широты, долготы) и факторную переменную *Ecoregion* оценивать на выбросы вряд ли целесообразно. Как мы видим на рис. 1, выбросы имеют лишь переменные: диаметр ствола (*DBH*) и возраст деревьев (*Tree_age*).

Как поступить с выбросами? Сразу удалить их из набора данных вряд ли разумно, поскольку выбросы могут содержать интересные денные[6]. Каждый раз нужно стремиться понять причину выбросов, в нашем случае это наибольшие значения диаметра ствола и возраста. Очевидно, что такие деревья встречаются нечасто, поэтому удалять выбросы из набора данных мы не будем. На всех диаграммах медиана, т. е. чёрная линия, пересекающая серый квадрат, ограничивающий значения верхнего и нижнего квартилей сдвинута от срединного положения. Распределение, вероятно, будет отличаться от нормального.

[5] Бослаф С. Статистика для всех. М.: ДМКПресс, 2017. 586 с.

[6] Smith C., Warren M. GLMs in R for Ecology. Independently published, 2019. 79 p. [Электронный ресурс]. URL: https://irep.ntu.ac.uk/id/eprint/37478/1/14596_Smith.pdf. (дата обращения: 12.02.2025).

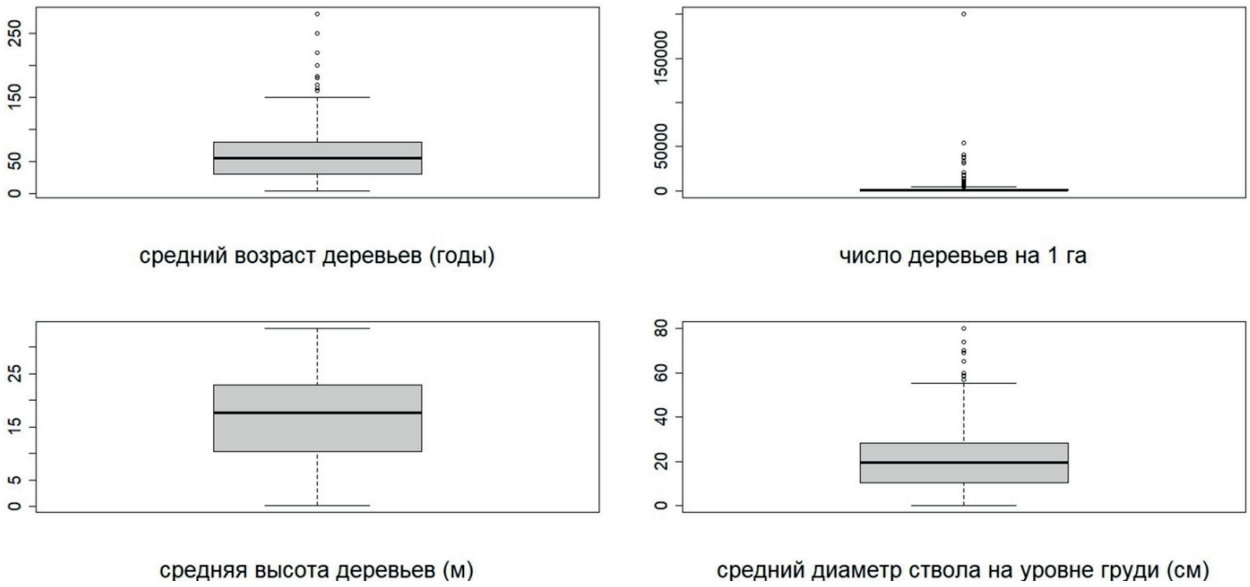


**Рис. 1 / Fig. 1.** Диаграммы «ящики с усами» лесотаксационных характеристик дуба черешчатого на пробных площадях / Boxplot diagrams of forest taxation characteristics for Oak petiolate on sampling areas

*Источник:* составлено авторами

Диаграммы на рис. 1 позволяют предположить, что *распределение зависимой переменной* отличается от нормального, но лучше использовать специальные тесты. Обычно применяют тест Шапиро-Уилка или тест Колмогорова-Смирнова из пакета *nortest*[7], это соответственно: *shapiro.test(d$DBH)* с результатом p-value = 4,884e$^{-12}$ и *lillie.test(d$DBH)* с p-value = 6,605e$^{-09}$. Нулевая гипотеза подразумевает нормальное распределение данных, а значение p<0,05 отвергает это предположение.

Таким образом, выбор *GLM* вполне оправдан.

Однородность в регрессионном анализе проверяется с использованием остатков модели, что мы сделаем после подгонки моделей.

Под однородностью понимается скоррелированность между независимыми переменными, которая негативно влияет на результат анализа и может оказаться значительной проблемой, поэтому необходима *проверка на мультиколлинеарность* [22]. Корреляция со значениями коэффициента более 0,75 считается сильной.

В нашем случае можно предположить корреляцию между возрастом дубов, числом стволов и высотой деревьев. Все 3 переменные не отвечают требованиям нормальности, поэтому рассчитаем непараметрический коэффициент корреляции Спирмена с помощью функции *cor.test()*:

```
cor.test(d$Tree_age, d$Tree_number,
method ='spearman', exact = FALSE)
```

```
        Spearman's rank correlation rho

data:  d$Tree_age and d$Tree_number
S = 14470217, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8764902
```

Результаты показывают, что между переменными *Tree_age* и *Tree_number* существует сильная отрицательная и достоверная корреляция rho = -0,87, p-value <0,05. Сильная корреляция показана и для сочетаний с переменной *Height* (высота деревьев). Поэтому использовать переменные *Tree_age*, *Tree_number* и *Height* в одной модели нельзя.

Какие зависимости предположить? Построим диаграммы рассеяния (рис. 2) и посмотрим на корреляцию между переменными графически, с помощью функции *plot(x~y)*, где *y* – переменная *DBH*, *x* – переменные *Tree_age*, *Tree_number*, *Latitude* и *Longitude*, тогда как категориальная переменная *Ecoregion* не отображена.

[7] Gross J. Ligges _nortest: Tests for Normality_. R package version 1.0-4. 2015: [Электронный ресурс]. URL: https://CRAN.R-project.org/package=nortest (дата обращения: 06.03.2025).

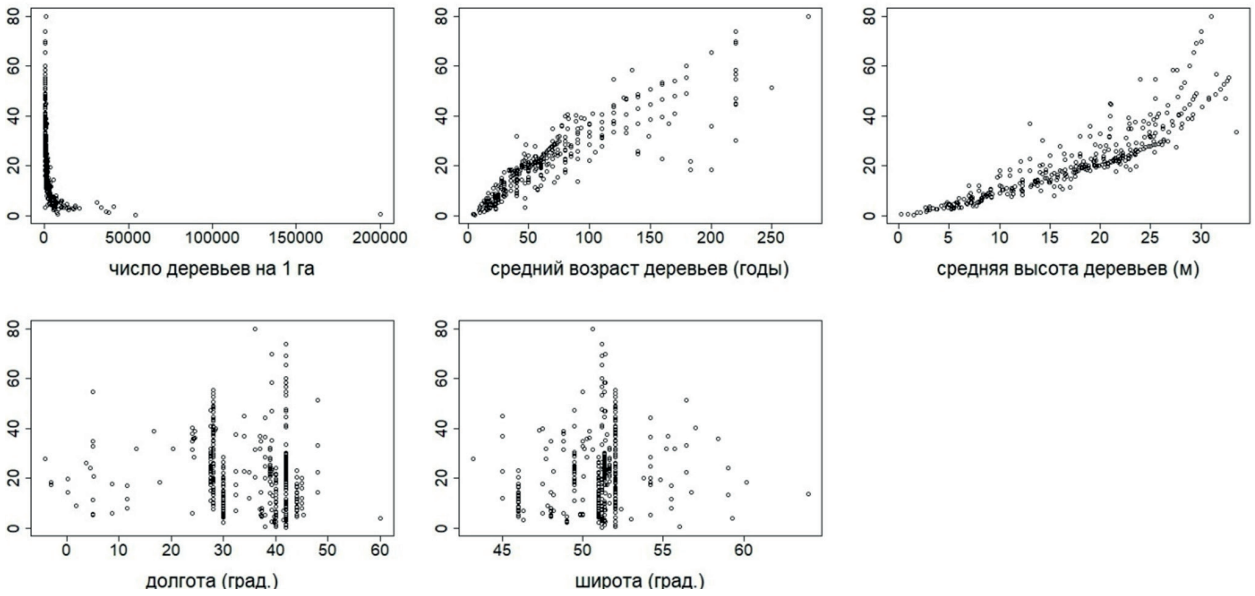


**Рис. 2 / Fig. 2.** Диаграммы рассеяния, где по *Oy* отложен диаметр ствола, а по *Ox* другие количественные переменные / The scatter plots, where *Ox* represents trunk diameter, *Oy* represents other quantity variables

*Источник:* составлено авторами

Отчётливо прослеживается связь между диаметром ствола, возрастом и высотой деревьев, а также, возможно, числом деревьев на гектар. Мы смоделировали такие зависимости, используя *LM* регрессию, однако полученные модели не отвечают предъявляемым требованиям, в т. ч. требованию нормального распределения остатков, поэтому обратим внимание на непараметрические методы.

На рисунках 1 и 3 отчётливо видно, что диаметр ствола дубов преимущественно не превышает 30 см, бóльшие значения диаметра встречаются реже. Попробуем понять, что же влияет на вероятность превышения порогового показателя в 30 см, для чего используем *GLM*.

## ПОДГОНКА И ПРОВЕРКА МОДЕЛЕЙ

Добавим в таблицу данных *d* колонку с новой бинарной переменной *dbh*, кодирующей диаметр ствола меньше (обозначено как 0) и больше (1) 30 см:

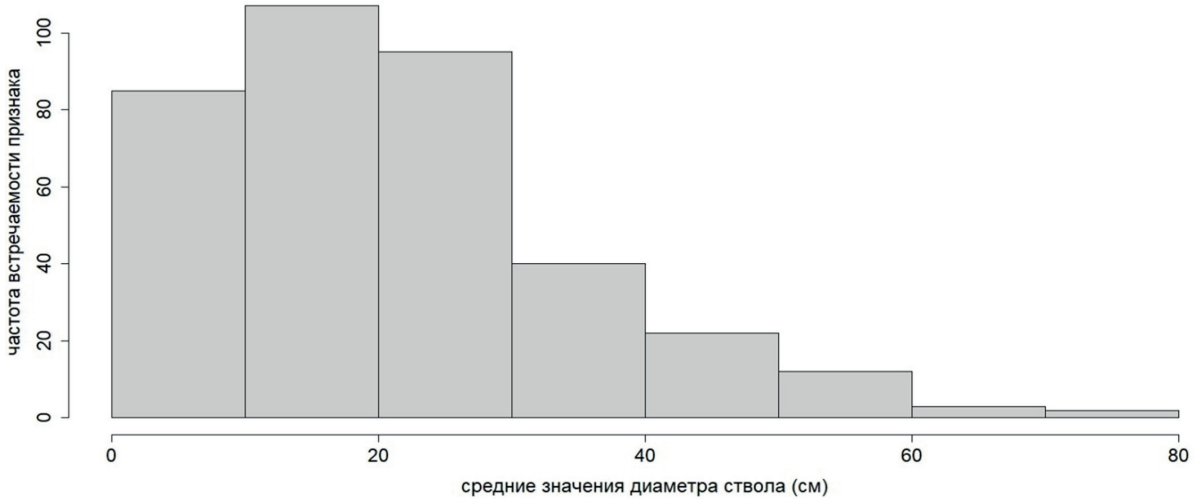


**Рис. 3 / Fig. 3.** Гистограмма средних значений диаметра ствола дуба / The histogram of average *Oak* trunk diameter

*Источник:* составлено авторами

*d$dbh[d$DBH < 30] <- 0*
*d$dbh[d$DBH > 30] <- 1*

Бинарная переменная *dbh* будет зависимой, тогда как другие переменные – независимыми предикторами (регрессорами).

После проверки данных можно переходить к третьему этапу протокола [22], а именно подгонке и проверке модели.

При подгонке регрессионных моделей можно создать:

- насыщенную модель, включающую все возможные предикторы;
- предполагаемую модель, которая содержит наиболее вероятные предикторы;
- нулевую модель, которая не содержит предикторов.

Подогнать насыщенную модель со всеми предикторами мы не можем, поскольку столкнулись с проблемой мультиколлинеарности. Поэтому для начала построим 3 модели с некоррелирующими переменными. Для этого воспользуемся функцией *glm()* и создадим модель *d1*, где моделируется переменная *dbh*, справа от знака ~ независимые предикторы, а именно средний возраст деревьев, широта, долгота и экорегион. Источник данных – таблица *d*, распределение биномиальное, а *logit* – связующая функция.

Код в *R* для модели *d1* выглядит следующим образом:

*d1<- glm(dbh ~ Tree_age + Latitude + Longitude + Ecoregion, data = d, family=binomial(link="logit"))*

Подробную информацию о модели можно получить с использованием функции *summary(d1)*.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.073e+00  2.218e+03   0.002   0.9985
Tree_age        6.806e-02  9.271e-03   7.341 2.12e-13 ***
Latitude       -5.137e-01  2.110e-01  -2.435   0.0149 *
Longitude      -7.435e-02  8.388e-02  -0.886   0.3754
Ecoregion80405  1.842e+01  2.218e+03   0.008   0.9934
Ecoregion80406 -7.260e+00  4.536e+03  -0.002   0.9987
Ecoregion80409  5.416e-01  3.570e+03   0.000   0.9999
Ecoregion80412  1.770e+01  2.218e+03   0.008   0.9936
Ecoregion80416  1.055e+01  2.218e+03   0.005   0.9962
Ecoregion80419  1.803e+01  2.218e+03   0.008   0.9935
Ecoregion80421  1.557e+00  3.569e+03   0.000   0.9997
Ecoregion80431  3.327e+01  4.536e+03   0.007   0.9941
Ecoregion80436  2.104e+01  2.218e+03   0.009   0.9924
Ecoregion80445  1.630e+01  2.218e+03   0.007   0.9941
Ecoregion80504  2.006e+01  2.218e+03   0.009   0.9928
Ecoregion80608  6.717e+00  2.874e+03   0.002   0.9981
Ecoregion80611  1.226e+01  4.536e+03   0.003   0.9978
Ecoregion80814  1.639e+01  2.218e+03   0.007   0.9941
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 381.81  on 365  degrees of freedom
Residual deviance: 136.67  on 348  degrees of freedom
AIC: 172.67

Number of Fisher Scoring iterations: 16
```

Колонка *Coefficients* содержит список независимых переменных, и поскольку предиктор *Ecoregion* мы задали как факторную переменную, то выведен полный список природных регионов. Колонка *Estimate* – это $\beta$-коэффициенты или параметры модели, *Std.Error* – стандартные ошибки.

Колонка *z value* содержит результат теста Вальда о достоверности отличий коэффициентов от 0, предполагая, что нулевая гипотеза $H_0$:$\beta_j$=0; альтернативная гипотеза $H_a$: $\beta_j \neq 0$. Тест Вальда рассчитывается по формуле: *Estimate/Std.Error*, статистическую значимость теста Вальда показывает *Pr(>|z|);* если p-значение <0,05 мы отверга-

ем нулевую гипотезу и принимаем альтернативную о значимости предиктора. Чем меньше значения стандартной ошибки, и чем больше *z value*, тем лучше, что полезно при сравнении моделей. Как видим, в модели *d1 Pr(>|z|)<0,05* для коэффициентов *Tree_age* и *Latitude.*

Уделим внимание и такому важному показателю как разница девиансов. Единица девианса – это разница между переменной *y* и её значением, предсказанным моделью, т. н. остатки. Значения *Null deviance* показывают остатки «нулевой» модели без предикторов, а *Residual deviance* – модели с предикторами. По этой причине, чем меньше *Residual deviance* по сравнению с *Null deviance*, тем лучше [6; 10]. Разницу девиансов можно рассчитать по формуле:

$$D = 1 - \frac{\text{Residual deviance}}{\text{Null deviance}} \quad (4)$$

Для модели *d1* этот показатель составил 0,64, что неплохо. Считается, что модель можно принимать во внимание если разница девиансов более 0,4, если же она превышает 0,75, то модель весьма удачна. Посмотрим и на информационный критерий Акаике (*AIC*), который разработан для сравнения моделей: чем меньше значения *AIC*, тем лучше [3; 5].

Поскольку предикторы *Ecoregion* и *Longitude* не значимы, построим сокращенную модель d2 с двумя предикторами:

```
d2 < - glm(dbh ~ Tree_age + Latitude,
data = d, family=binomial(link="logit"))
summary(d2)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.292554   4.574202   0.064    0.949
Tree_age     0.060529   0.007262   8.335   <2e-16 ***
Latitude    -0.125560   0.092093  -1.363    0.173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 381.81  on 365  degrees of freedom
Residual deviance: 168.32  on 363  degrees of freedom
AIC: 174.32

Number of Fisher Scoring iterations: 6
```

Предиктор *Tree_age* статистически значим, *Pr(> |z|) > 0,05*, в сравнении с моделью *d1* стандартная ошибка уменьшилась с 0,009 до 0,007, показатель *z* увеличился с 7,34 до 8,33; разница девиансов составила 0,55, *AIC* незначительно увеличился. Главный недостаток модели *d2* – меньшая разница девиансов. Мы получили т. н. «вложенные» модели, поскольку модель *d2* это модель *d1* с сокращённым числом предикторов. Необходимо понять, есть ли достоверные различия между вложенными моделями, для чего оценивается достоверность различий между девиансами с помощью теста хи-квадрат, функция *anova()*.

```
anova(d1, d2, test = 'Chisq')
```

```
Analysis of Deviance Table

Model 1: dbh ~ Tree_age + Latitude + Longitude + Ecoregion
Model 2: dbh ~ Tree_age + Latitude
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       348     136.66
2       363     168.32 -15  -31.651 0.007181 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Значения Pr(>Chi)<0,05, следовательно, модель *d2* объясняет данные хуже, чем *d1*. Функция *anova2()* из пакета *glmtoolbox*[8] использует другие алгоритмы и полезна при сравнении моделей. Для анализа вложенных моделей со многими предикторами можно использовать специальные функции, например *stepCriterion()* из пакета *glmtoolbox*. При использовании этой команды на основе различных критериев и тестов будет выбрана лучшая модель.

Понять, какие предикторы действительно значимы можно, если построить модели с каждым из них. Сведём результаты подгонки моделей с каждым из предикторов в таблицу 1.

*Таблица 1 / Table 1*

**Результаты моделирования влияния переменных / Results of simulating variables influence**

| Предиктор | Pr(>\|z\|) | Разница девиансов | Информационный критерий Акаике AIC |
|---|---|---|---|
| Tree_age | *** | 0.55 | 174.34 |
| Tree_number | *** | 0.52 | 182.75 |
| Height | *** | 0.53 | 183.65 |
| Longitude | ** | 0.02 | 378.14 |
| Latitude | · | 0.01 | 382.79 |
| Ecoregion | · | 0.10 | 331.31 |

*Примечание:* *p<0,05; **p<0,01; ***p<0,0001; · p>0,05

*Источник:* составлено авторами

Статистически значимы предикторы *Tree_age*, *Tree_number*, *Height* а также *Longitude*, хотя разница девиансов для последней переменной не велика.

Как показали результаты предварительных тестов, между предикторами *Tree_age*, *Tree_number* и *Height* установлена сильная корреляция. Для оценки мультиколлинеарности в *LM* и *GLM* разработана специальная функция *vif()* (*Variance Inflation Factor*) из пакета *car*. Аналогичный результат покажет функция *gvif()* из пакета *glmtoolbox*, поскольку для *GLM* рассчитывается *GVIF* (*generalized variance-inflation factors*). Применительно к модели *d1* мы получим следующий результат:

```
> vif(d1)
                GVIF Df GVIF^(1/(2*Df))
Tree_age    1.639226  1        1.280322
Latitude    4.172914  1        2.042771
Longitude 11.221700  1        3.349881
Ecoregion 61.568452 14        1.158526
```

Считается, что показатель *GVIF≥10* показывает значительную мультиколлинеарность, но проблема мультиколлинеарности существует уже при показателях 3 [15]. Мы можем видеть, что *VIF* превышает 3 для переменных *Ecoregion*, *Latitude* и *Longitude*, что вполне объяснимо.

Результаты сравнения с использованием дисперсионного анализа показали, что сокращенная модель *d3* с одним предиктором *Tree_age* объясняет наши данные так же хорошо, как и модель *d2* с независимыми переменными *Tree_age* и *Latitude*. Для краткости изложения код d3 мы приводить не стали. Использование функции *anova2* (*d1, d3, test='lr'*)с тестом отношения правдоподобий для сравнения моделей *d1* и *d3* показало отсутствие между ними достоверных различий, поскольку *Pr(Chisq) >0,05*

```
  Likelihood-ratio test

Model 1 :  dbh ~ Tree_age + Latitude + Longitude + Ecoregion
Model 2 :  dbh ~ Tree_age

         Chi    df  Pr(Chisq>)
1 vs 2 -33.674   0           1
```

Их трёх моделей *d1, d2* и *d3* выберем наиболее простую модель *d3*.

Смоделируем влияние других значимых переменных.

*d4<-glm(dbh ~ Tree_number + Latitude, data = d, family=binomial(link="logit"))*
*summary(d4)*

[8] Vanegas L., Rondyn L., Paula G. glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models_. R package version 01.10.2004: [Электронный ресурс]. URL: https://clck.ru/3R8pwW (дата обращения: 03.03.2025).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 16.877362   5.222938   3.231  0.00123 **
Tree_number -0.009044   0.001221  -7.409 1.28e-13 ***
Latitude    -0.254273   0.096789  -2.627  0.00861 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 373.24  on 358  degrees of freedom
Residual deviance: 171.61  on 356  degrees of freedom
  (7 пропущенных наблюдений удалены)
AIC: 177.61

Number of Fisher Scoring iterations: 11
```

Коэффициенты значимы, разница девиансов − 0,54, поэтому модель *d4* можно принять во внимание.

Модель *d5* содержит лишь один значимый предиктор:

*d5<-glm(dbh ~ Height, data = d, family=binomial(link="logit"))*

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.46780    1.39362  -8.229  < 2e-16 ***
Height        0.47061    0.05908   7.965 1.65e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 381.32  on 364  degrees of freedom
Residual deviance: 179.65  on 363  degrees of freedom
  (1 пропущенное наблюдение удалено)
AIC: 183.65

Number of Fisher Scoring iterations: 7
```

Разница девиансов модели *d5* составила 0,53.

Таким образом, мы выбрали 3 модели с 4 значимыми предикторами:

1. модель *d3* с предиктором *Tree_age,*
2. модель *d4* с предикторами *Tree_number* и *Latitude*
3. модель *d5* с предиктором *Height.*

Интерпретация полученных коэффициентов представляет существенный интерес для исследователя, поскольку отвечает на вопрос «что говорят полученные коэффициенты об изучаемом вопросе?» и состоит из оценки отношения между зависимой и независимой переменными и определения единицы изменения независимой переменной [10]. Вспомним, что мы моделируем влияние независимых переменных на вероятность превышения среднего значения диаметра ствола показателя в 30 см. Для примера выберем модель *d4*, поскольку она содержит два значимых предиктора, *Tree_number* или среднее число экземпляров дуба, и *Latitude,* который измеряется в градусах широты. Выведем коэффициенты модели:

*coef(d4)*

*(Intercept) Tree_number Latitude*

*16.877362448 -0.009044065 -0.254273213*

Запишем регрессионное уравнение логита ($p_i$):

$$logit(p_i) = 16{,}87 - 0{,}009 \times Tree\_number - 0{,}25 \times \mathrm{L}atitude$$

Например, для пробной площади с 800 экз. дуба, расположенной на широте 45° логит составит:

$$logit(p_{800}) = 16{,}87 - 0{,}009 \times 800 - 0{,}25 \times 45 = -1.58$$

а для пробной площади со 100 экз. на широте50°:

$$logit(p_{100}) = 16{,}87 - 0{,}009 \times 100 - 0{,}25 \times 50 = 3.47$$

Вспомним, что *логит* – это натуральный логарифм отношения вероятностей, поэтому для нахождения вероятности используем формулу (3) и потенцируем полученные результаты с использованием функции *exp()*.

```
exp(-1.58)/(1+exp(-1.58))
[1] 0.1707955
exp(3.47)/(1+exp(3.47))
[1] 0.969822
```

Таким образом, в соответствии с моделью *d4* вероятность превышения диаметра ствола дубов средней 30 см на пробной площади с 800 экз. на широте 45° составляет 0,17 или 17%, а на пробной площади со 100 экз., хотя и расположенной на 5° севернее, уже 0,97 или 97%.

Изобразить линию регрессионной зависимости можно с помощью команды *visreg()* (*Visualization of regression functions*) из пакета *visreg* [2]. Для простоты используем модель *d5* с одной независимой переменной:

```
visreg(d5, ylim = c(0,1), scale = 'response')
```

Аргумент *scale='response'* добавляется в код *glm*-моделей для построения функции в масштабе ответа, для бинарных данных это значения вероятностей от 0 до 1. На рисунке 4 изображена функция связи между зависимой и независимой (средним значением высоты дубов на пробной площади) переменными в границах 95%-го доверительного интервала.

Схожий график может построить команда *binreg_plot()* из пакета vcd [14].

Из подогнанных моделей мы выбрали модель *d4*, которую и проверим на соответствие требованиям, предъявляемым к линейным моделям. Немалое внимание уделяется остаткам модели, в т. ч. их графическому отображению [3; 6]. Разница между наблюдаемыми и вычисленными

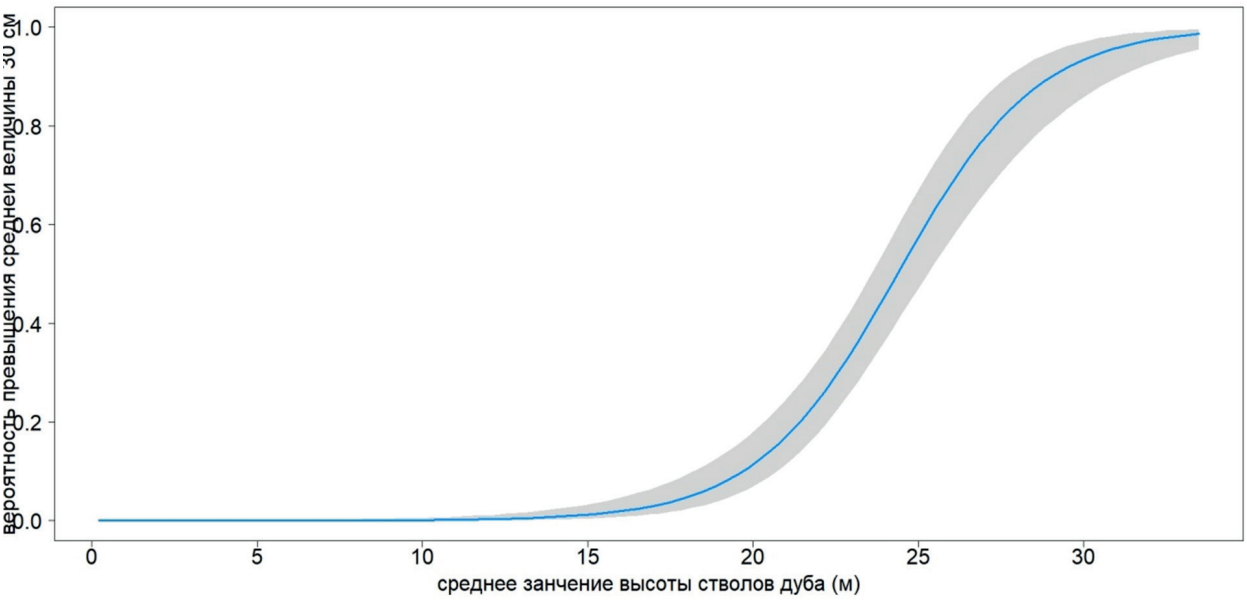


**Рис. 4 / Fig. 4.** Функция связи – голубая линия, серая заливка показывает доверительный интервал в 95% / Connection function is the blue line, the border of 95% confidence interval has grey color

*Источник:* составлено авторами

значениями зависимой переменной получили название сырых остатков ($r_i$), которые обычно используются при оценке *LM*, поскольку они отвечают требованиям нормальности. В *GLM* дисперсия остатков не постоянна и изменяется для разных $y_i$, поэтому обычно используются взвешенные значения остатков. Для бинарных моделей это т. н. квантильные остатки ($r_q$) [3; 7], которые рассчитываются при нахождении эквивалента стандартного отклонения для каждого ответа [6; 7]. Рассчитать $r_q$ можно посредством команды *qresid()* из пакета *statmod* [9]. Квантильные остатки для модели *d4* отвечают требованиям нормального распределения, что подтверждает тест Колмогорова-Смирнова: *shapiro.test(qresid(d4)), p-value=0,1.*

Воспользуемся командой *simulateResiduals()* из пакета *DHARMa*[9] и построим графики (рис. 5), которые содержат результаты нескольких тестов. Код программы для модели d4: *plot(simulateResiduals(d4))*.

Графики слева, или *Q-Q*-графики, показывают соотношение теоретически ожидаемых и расчетных значений квантильных остатков, расположение которых вдоль прямой линии подтверждает соответствие распределения нормальному. Проверка выбросов (*outliertest*) не показала значительных результатов (*p=0,10*). Тест дисперсий (*dispersiontest*) не показал значительных отклонений (*p=0,30*). Правый график демонстрирует различия между ожидаемыми и расчётными значениями квантильных остатков. Остатки должны быть распределены равномерно по квантилям, обозначенным горизонтальной штриховой линией, что подтверждают ровные сплошные линии. В противном случае линии изогнуты и выделены красным, что видно на графике для модели *d1*, которая не соответствует предъявляемым требованиям.

Принципиально иное направление проверки моделей – это оценка качества прогноза, т. е. оценка того, насколько хорошо полученная модель предсказывает значение зависимой переменной. Для этого выборку разбивают на 2 части – обучающую и тестовую; первую используют для подгонки модели, а вторую – для проверки. Разбить выборку на 2 части несложно используя команду *sample.split()* из пакета *caTools*[10], которая разделяет данные не случайно, а таким образом, чтобы в обучающей и тестовой выборках сохранилось соотношение положительных и отрицательных ответов.

```
split<-sample.split(d, SplitRatio = 0.8)
trset = d[split, ]
testset = d[!split, ]
```

Мы разделили выборку на 2 неравные части в соотношении 0,8:0,2 (параметр *SplitRatio=0,8*). В результате мы получили 2 новых таблицы *trset* и *testset*, которые сохраняют примерное соотношение положительных и отрицательных ответов как 8:2.

```
table(trset$dbh)/sum(table(trset$dbh))
0 1
0.7852113 0.2147887

table(testset$dbh)/sum(table(testset$dbh))
0 1
0.7804878 0.2195122
```

Подгоним модель *trset5* с одной независимой переменной *Height* на обучающей выборке *trset*, аналогично модели *d5*:

```
trset5 <- glm(dbh ~ Height, data = trset,
family=binomial(link="logit"))
summary(trset5)
```

```
Call:
glm(formula = dbh ~ Height, family = binomial(link = "logit"),
    data = trset)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.96463    1.50985  -7.262 3.81e-13 ***
Height        0.44784    0.06385   7.014 2.32e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 295.01  on 282  degrees of freedom
Residual deviance: 144.09  on 281  degrees of freedom
  (1 пропущенное наблюдение удалено)
AIC: 148.09

Number of Fisher Scoring iterations: 7
```

Переменная *Height* значима и значения коэффициентов близки к таковым модели *d5*.

[9] Hartig F. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models_. R package version 0.4.6. 2022 [Электронный ресурс]. URL: https://clck.ru/3R8yum (дата обращения: 26.10.2024).

[10] Tuszynski J. caTools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.2. 2021 [Электронный ресурс]. URL: https://clck.ru/3R8yuA (дата обращения: 26.10.2024).

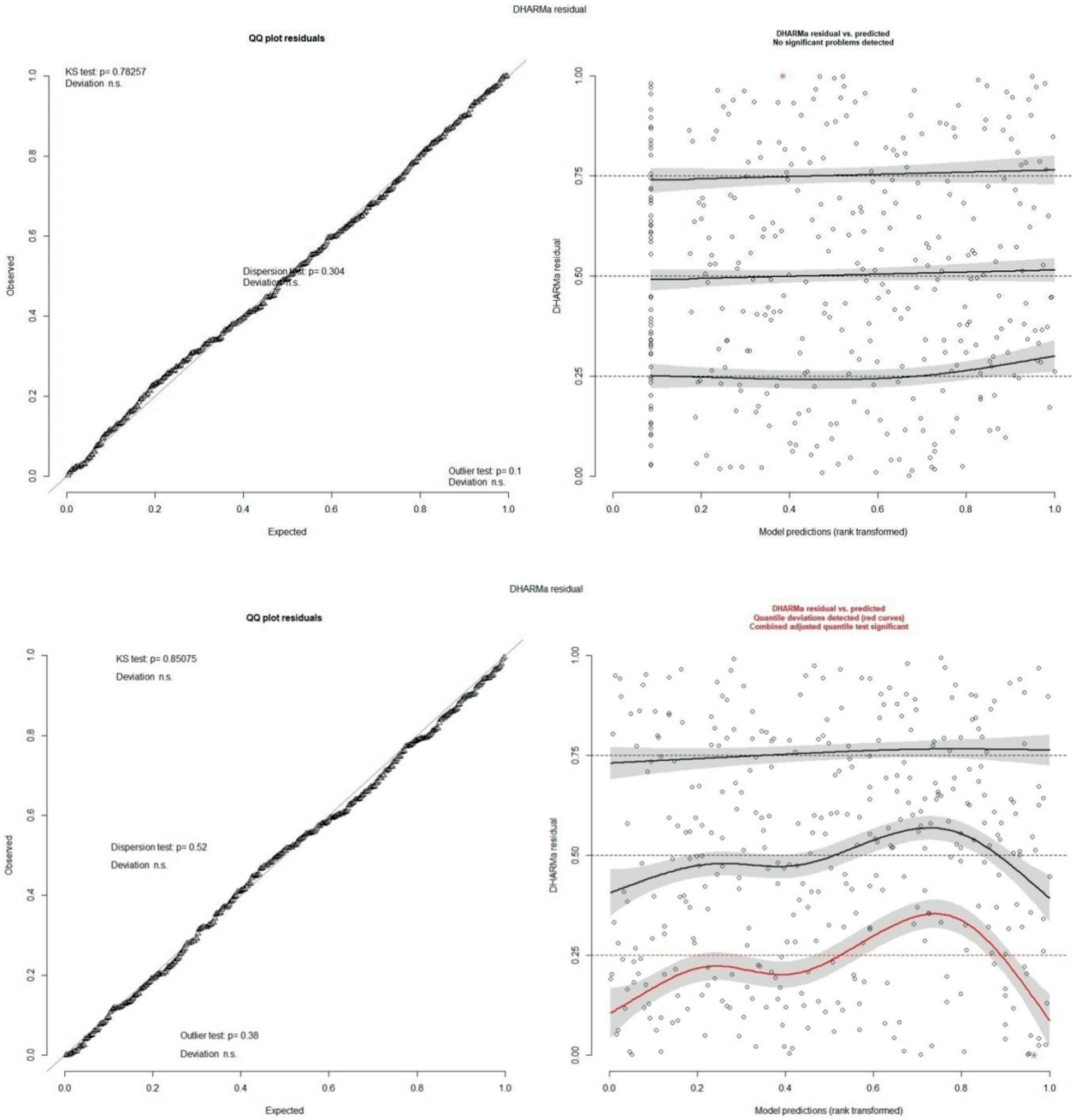


**Рис. 5 / Fig. 5.** Q-Q – графики и графики распределения квантильных остатков для модели d4 (вверху) и модели d1 (внизу) / Q-Q plots and quintile residuals distribution plots for model d4 (upper) and d1 (down)

*Источник:* составлено авторами

Проверим, насколько же хорошо модель *trset5* предскажет зависимую переменную *y* в тестовой выборке *testset*. Для этого используем команду *predict()*, которая рассчитывает результаты применения модели *trset5* к набору данных *testset*, параметр *type="response"* задаёт характерный для бинарных данных альтернативный ответ:

*predict_result < -predict(trset5, testset, type = 'response')*

В результате получен числовой вектор со значениями вероятностей успешных событий *y* для каждого значения *x*. Проверим, насколько отличается число истинных и ложных ответов после применения модели *trset5* от реального соотношения ответов в наборе данных *testset*, для чего полезна функция *table()*, которая создаёт таблицу сопряжённости. Значение аргу-

мента *predict_result* определяет положительные ответы для вероятности более 0,5.

```
table(testset$dbh, predict_result > 0.5)
FALSE TRUE
0 62 2
1 7 11
```

Мы получили таблицу частот или сопряжённости, где значениям 0 и 1 в строках соответствуют истинные и ложные ответы в столбцах. В колонке *FALSE* содержаться негативные ответы, а в колонке *TRUE* – положительные.

Результаты не столь однозначны, поэтому следует оценить достоверность различий между ответами в модели *trset5* и набором данных *testset*. Для этого применяются меры согласия, самый простой из которых – оценка процента согласия, т. е. соотношения верных и неверных ответов, нередко применяется тест хи-квадрат. К сожалению, указанные методы не лишены недостатков, поскольку совпадения могут быть случайны[11].

Подобные случайности исключает специально разработанная мера согласия, а именно каппа Коэна или каппа-коэффициент, который оценивает согласованность результатов и рассчитывается как диагональная сумма частот в таблице сопряжённости. Значения каппы изменяются в диапазоне от -1 до +1, где 0 означает случайное совпадение, +1 – выражает полную согласованность, а -1 – полную несогласованность.

Рассчитать каппу можно с помощью команды *Kappa()* из пакета *vcd* [14].

```
t<-table(testset$dbh, predict_result > 0.5)
Kappa(t)
 value ASE z Pr(>|z|)
Unweighted 0.6442 0.1075 5.99 2.093e-09
Weighted 0.6442 0.1075 5.99 2.093e-09
```

Каппа составила 0,64, что принято как хороший уровень согласованности. Ниже границы 0,4 уровень слабый, выше 0,8 – очень сильный. Приведены взвешенное *Weighted* (для порядковых переменных) и не взвешенное *Unweighted* значения, стандартные ошибки (ASE), *z*-статистика (z = value/ASE); $p < 0,05$, т. е. полученные результаты достоверны. Отметим, что при повторном разделении данных на обучающую и тестовые выборки результаты могут несколько отличаться.

Для проверки качества прогноза можно использовать и графические методы, а именно *ROC*-кривые (*Receiver Operating Characteristic*) (рис. 6). ROC-кривые можно построить с использованием пакета *ROCR* [19], где команда *prediction()* создаёт объект с результатами прогнозирования, а команда *performance()* оценивает прогноз. Вновь оценим модель *trset5*, результаты применения которой содержит числовой вектор *predict_result*.

Код построения *ROC*-кривой:

```
ROCRpred5 = prediction(predict_result,
testset$dbh)
ROCRperf5 = performance(ROCRpred5,
"tpr", "fpr")
plot(ROCRperf5, colorize=TRUE, print.
cutoffs.at = seq(0,1,0.1), text.adj =
c(-0.2,1.7))
```

При построении *ROC*-кривых вновь используется матрица ошибок, а кривая выражает компромисс между чувствительностью (*sensitivity*) или вероятностью предсказать положительные ответы, когда они действительно положительны и специфичностью (*specificity*) или вероятностью предсказать отрицательные результаты, когда они действительно отрицательны. Поэтому, чем выше качество модели, тем больше значения по *Oy* и меньше значения по оси *Ox*, а сама кривая ближе к точке с координатой (0;1). Модель *trset5* с предиктором *Height* предсказывает переменную *dbh* достаточно хорошо и площадь под *ROC*-кривой близка к площади всего графика.

Таким образом, полученные результаты показали, что вероятность превышения диаметром ствола порога в 30 см зависит от числа деревьев на площади, средних возраста, высоты и в меньшей степени от широты, на которой расположена пробная площадь. Переменная долгота и такой, казалось бы, очевидный показатель как экорегион оказались незначимы.

## АНАЛИЗ МОЩНОСТИ

При проведении научного исследования следует определиться с минимальным

[11] Бослаф С. Статистика для всех. М.: ДМК Пресс, 2017. 586 с.

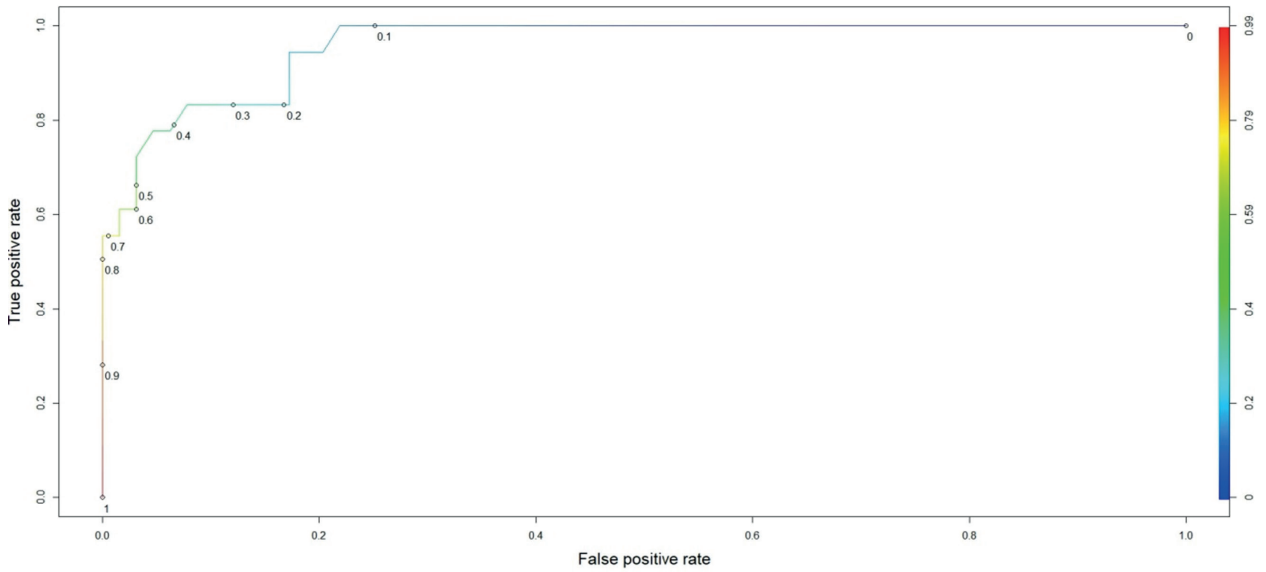


**Рис. 6 / Fig. 6.** ROC-кривая для модели trset5 / ROC-curve for trset5 model

*Источник:* составлено авторами

объёмом данных, необходимым для корректного результата. Авторы не всегда уделяют внимание этому вопросу, между тем недостаточная выборка не позволит сделать достоверный статистический вывод, тогда как получение излишних данных нередко связано со значительными затратами ресурсов. Это особенно актуально для экологических исследований и сбора материала в поле.

Размер выборки тесно связан с таким понятием, как мощность. Мощность – это вероятность не совершить статистическую ошибку II рода (β), приняв нулевую гипотезу при верной альтернативной. Обычно исследователи оценивают только вероятность ошибки I рода (α), когда отвергается верная нулевая гипотеза. Мощность рассчитывается как *(1–β)*, а порог мощности принят в 80 или 90%. Если уровень значимости или $\alpha$ – это вероятность нахождения несуществующей закономерности, то мощность – это вероятность обнаружения существующей закономерности[12].

На мощность влияют: вероятность ошибки I рода, различия в результате между группами, размер выборки и выбранный критерий. Поэтому, задавая минимальную мощность, теоретически можно определить минимальный размер выборки.

Для оценки размера выборки в *R* разработан пакет *pwr*, однако для *GLM* специальных функций не предусмотрено. В интернете можно найти коды для расчета мощности в *R*[13] или же использовать онлайн-калькуляторы[14], но сначала скажем пару слова о теоретической базе расчётов. В основе рассуждений лежит одновыборочный двусторонний *Z*-тест проверки гипотез, который оценивает, действительно ли различается частота бинарных ответов в 2 выборках ($p$ и $p_0$) и учитывает объём выборки $n$[15] [21].

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \times p(1-p)}{(p-p_0)^2} \quad (5)$$

Нулевая гипотеза $H_0: p = p_0$, альтернативная гипотеза $H_a: p \neq p_0$. Для нулевой гипотезы $p_0 = 0{,}5$, т. е. вероятность успеш-

[12] Кабаков Р. И. R в действии. Анализ и визуализация данных в программе R / пер. А. Киселёва. М.: ДМК Пресс. 2023. 768 с.

[13] Calculate Sample Size Needed to Test 1 Proportion: 1-Sample, 2-Sided Equality [Электронный ресурс]. URL: https://clck.ru/3R8zt5 (дата обращения: 12.02.2025).

[14] Comparing Two Proportions – Sample Size [Электронный ресурс]. URL: https://select-statistics.co.uk/calculators/sample-size-calculator-two-proportions/ (дата обращения: 12.02.2025).

[15] Smith C., Warren M. GLMs in R for Ecology. Independently published, 2019.79 p. URL: https://irep.ntu.ac.uk/id/eprint/37478/1/14596_Smith.pdf (дата обращения: 12.02.2025).

ных и неуспешных событий одинакова, тогда как вероятность успешных событий для альтернативной гипотезы мы можем узнать лишь из собранных данных. *Z*-значение стандартного нормального распределения, $\alpha$ и $\beta$ – вероятность ошибки I и II рода.

Определим среднее значение переменной *dbh:*

```
mean(d$dbh)
[1] 0.215847
```

Вероятность положительных ответов в нашем наборе данных $p=0,2$. Приведём код для расчета размера выборки[1], для сравнения двух выборок с $p=0,2$ и $p_0=0,5$. Уровень значимости или вероятность ошибки I рода $\alpha=0,01$ и вероятности ошибки II рода $\beta = 0,1$.

```
p=0.2
p0=0.5
alpha=0.01
beta=0.10
```

Перепишем формулу (5) на языке *R* и рассчитаем критическое значение для теста с заданным уровнем $\alpha$ и $\beta$. *Z*-значение задаётся с использованием команды *qnorm()*.

```
n=p*(1-p)*((qnorm(1-alpha/2)+qnorm
(1-beta))/(p-p0))^2
[1] 26.45224
```

Минимальный объём выборки ($n$) для заданных условий составил 27 наблюдений, поэтому мы можем использовать таблицу *d* для подгонки *GLM*. При меньших различиях между $p$ и $p_0$ это число увеличится.

Следует сказать несколько слов об оформлении полученных материалов. Для простоты понимания мы приводим данные в том виде, в котором они изображены в *R*, однако это вряд ли удобно для публикации.

При выведении результатов целесообразно использовать специальные команды, например *stargazer()* из одноимённого пакета[16]. Функция *stargazer(model, type='text')* выводит параметры модели в формате *ASCII; type = 'html' в виде кода html.*

Код и результат его выполнения для модели *d4:*

*stargazer(d4, type = 'text').*

| | *Dependent variable:* |
|---|---|
| | dbh |
| Tree_number | -0.009*** |
| | (0.001) |
| Latitude | -0.254*** |
| | (0.097) |
| Constant | 16.877*** |
| | (5.223) |
| Observations | 359 |
| LogLikelihood | -85.804 |
| AkaikeInf. Crit. | 177.609 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

## ЗАКЛЮЧЕНИЕ

В результате моделирования переменной *dbh* мы получили несколько моделей, в т. ч. наиболее качественную модель *d2*. Наш анализ показал, работа с *GLM* имеет свои особенности, отличные от *LM* это использование разницы девиансов и квантильных остатков в оценке качества модели, использование непараметрических тестов в сравнении вложенных моделей, потенцирования параметров модели. Все команды, использованные в моделировании переменной *dbh*, сведены в таблице 2.

Мы рассмотрели лишь некоторые команды и пакеты, обзор всех возможных функций для работы с логистическими моделями выходит за рамки нашей работы.

Принципиальная схема анализа данных приведена в работе Р. И. Кабакова[17]; исследователи часто используют протокол, предложенный в работе *«A Protocol for data exploration to avoid common statistical problems»* [22], мы же уточнили этот весьма общий алгоритм применительно к логистической регрессии (рис. 7) и добавили

[16] Hlavac M. stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. 2022. https://CRAN.R-project.org/package=stargazer.

[17] Кабаков Р. И. R в действии. Анализ и визуализация данных в программе R. М.: ДМК Пресс. 2023. 768 с.

*Таблица 2 / Table 2*

**Использованные на разных этапах моделирования переменной *dbh* команды и пакеты *R* / Functions and packages R were used for different steps of dbh variable modeling**

| Этапы исследования | | Команды | Пакеты |
|---|---|---|---|
| Получение данных | Анализ мощности | Код с использованием функции qnorm() | ядро R |
| Исследование данных | Определение выбросов | boxplot() | ядроR |
| | Проверка нормальности распределения | shapiro.test() | ядро R |
| | | lillie.test() | nortest |
| | Оценка мультиколлинеарности | cor.test(), plot() | ядро R |
| | | vif() | car |
| Подгонка и проверка модели | Подгонка модели | glm(y~x, family=binomial(link="logit")) | ядро R |
| | Оценка достоверности различий вложенных моделей | anova(model1, model2, test = "Chisq") | ядро R |
| | | anova2() | glmtoolbox |
| | Потенциирование параметров модели | exp() | ядро R |
| | | parameters() | parameters |
| | Построение линии регрессии | visreg() | visreg |
| | | binreg_plot() | vcd |
| | Получение квантильных остатков | qresid() | statmod |
| | Построение графика распределения квантильных остатков | simulateResiduals() | DHARMA |
| | | | |
| Оценка прогностической способности модели | Расщепление выборки | split() | caTools |
| | Расчёт результатов применения модели | predict() | ядро R |
| | Построение таблиц сопряжённости | test() | ядро R |
| | Расчет Каппа-коэффициента | Kappa() | vcd |
| | Создание объекта с результатами прогноза для построения ROC-кривых | prediction() | ROCR |
| | Оценка прогноза | performance() | ROCR |
| | Построение ROC-кривых | plot() | ядро R |

*Источник:* составлено авторами

некоторые важные этапы. Наш вариант включает 10 этапов, ошибка на каждом из них может привести к неверному результату. Однако ошибки в расчётах можно исправить, важно понять, какие этапы имеют принципиальное значение.

Во-первых, исследователь может неверно выбрать независимые переменные и впоследствии получить модели, где все предикторы будут незначимы. Такой результат всё равно может представлять интерес и заслуживать публикации, однако авторы всё же стремятся получить значимые результаты. Дать простой совет здесь сложно, поскольку многое зависит от знакомства исследователя с темой работы и знания литературы.

Во-вторых, оценка минимального объёма данных. На этом этапе исследователь может принять неверное решение, поскольку для такой оценки необходимо знать соотношение отрицательных и положительных ответов, т.е. иметь уже собранные данные. В наших данных разница между p и $p_0$ значительна, поэтому объём выборки невелик, однако, чем меньше разница между p и $p_0$, тем большая выборка потребуется. При значениях $p=0,4$ и $p_0=0,5$, $\alpha=0,05$ и $\beta=0,2$, мы получим $n=189$. Думаем, на эту цифру можно ориентиро-

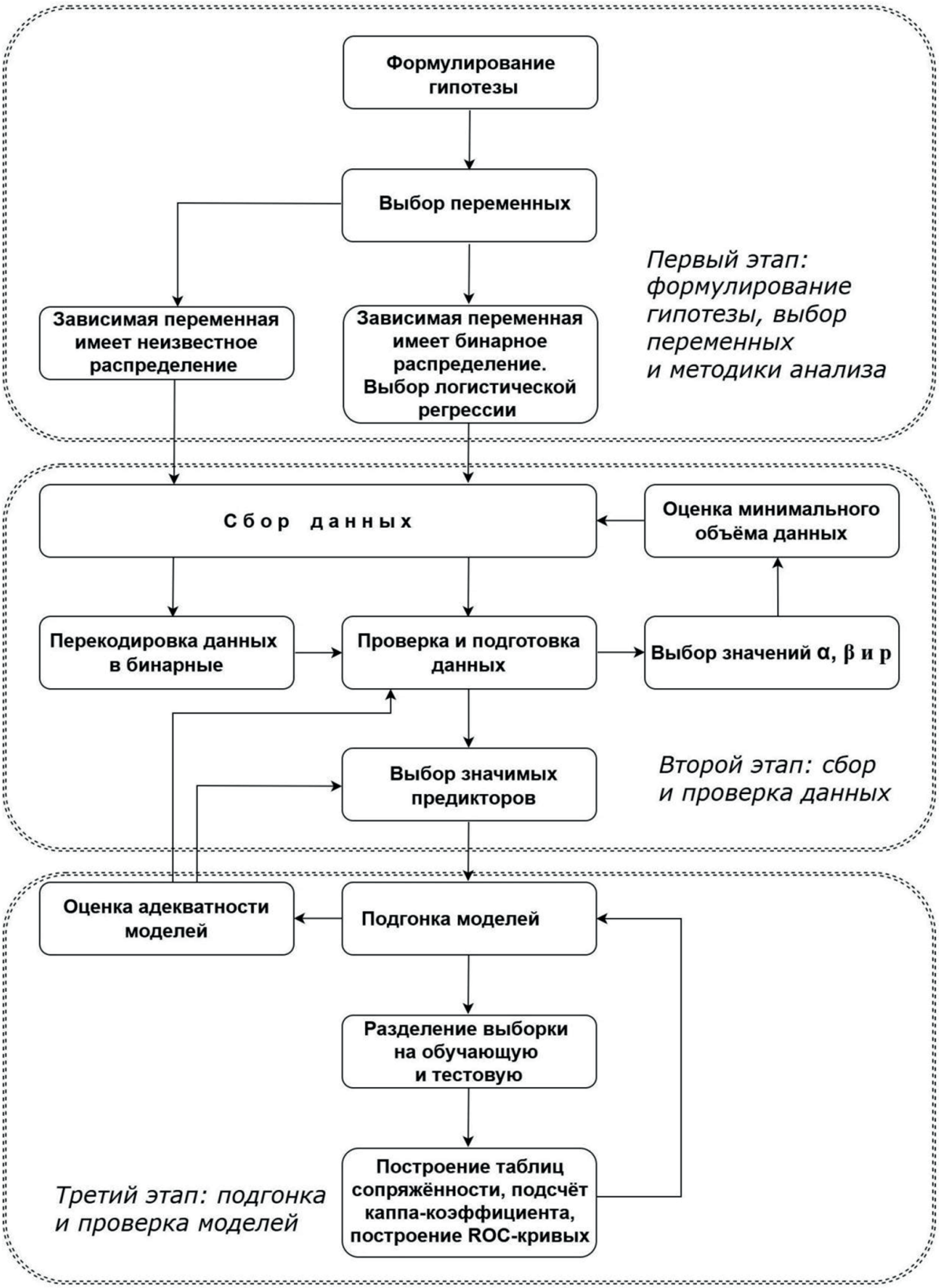


**Рис. 7 / Fig. 7.** Схема построения и проверки логистической модели / The algorithm of the logistic model creation and checking

*Источник:* составлено автором

ваться при планировании исследования. Следует учитывать и последующее разделение выборки на обучающую и тестовую, когда неизбежно происходит уменьшение объёма данных, что целесообразно предусмотреть заранее.

В-третьих, очень ответственный этап до начала подгонки моделей – это проверка и подготовка данных, поскольку от него зависит не только конечный результат, но и расчёт значений *p*. Неслучайно именно проверка данных отнимает столько времени и сил [3; 22]. На этом этапе существенной может оказаться проблема выбросов, следует понять их причину. Это объекты со значительными показателями, как деревья с большим диаметром ствола, которых обычно немного, либо же ошибки в сборе материала[18]. Решение об удалении выбросов зависит от исследователя и носит субъективный характер, которое, однако, существенно влияет на дальнейшую работу.

В-четвёртых, это разделение выборки на обучающую и тестовую части, поскольку результаты каждый раз могут отличаться, особенно если объём данных не очень велик.

Наконец, предложенный нами алгоритм анализа однозначно подтверждает известный принцип: первоначально следует выбрать цель исследования и сформулировать гипотезу, а затем уже приступать к сбору материала. В нашем примере мы использовали готовые данные, однако если бы мы планировали исследование с нуля, то важнейший вопрос, на который должен ответить исследователь – какие именно данные собирать? Зачастую исследователи наудачу собирают материал и лишь впоследствии пытаются понять, что же с ним делать. В экологических работах это случается нередко, поскольку полевые работы кажутся не только более важными, но и наиболее интересными, особенно для молодых исследователей. Конечно, это неверный путь, ведь данных может не хватить, либо выбранные предикторы будут не значимы.

Надеемся, что предложенный нами алгоритм упростит применение логистических моделей в научной работе.

[18] Smith C., Warren M. GLMs in R for Ecology [Электронный ресурс]. URL: https://irep.ntu.ac.uk/id/eprint/37478/1/14596_Smith.pdf. (дата обращения: 12.02.2025).

## ЛИТЕРАТУРА


1. Замятина Н. Ю., Котов Е. А., Гончаров Р. В., Бурцева Е. А., Гребенец В. И., Медведков А. А., Молодцова В. А., Клюева В. П., Кульчицкий Ю. В., Миронова Б. А., Никитин Б. В., Пилясов А. Н., Поляченко А. Е., Потураева А. В., Стрелецкий Д. А., Шамало И. А. Оценка потенциала жизнестойкости городов Российской Арктики // Вестник Московского университета. Серия 5: География. 2022. № 5. С. 52-65.
2. Breheny P., Burchett W. Visualization of Regression Models Using visreg // The R Journal. 2017. Vol. 9. P. 56–71. DOI: 10.32614/RJ-2017-046
3. Corlatti L. Regression models, fantastic beasts, and where to find them: a simple tutorial for ecologists using R // Bioinformatics and Biology Insights. 2021. Vol. 15. P. 1–19. DOI: 10.1177/11779322211051522
4. Chow S., Shao J., Wang H. Sample Size Calculations in Clinical Research. NY: Basel, 2008. 358 p
5. Collet D. Modelling Binary Data. Taylor & Francis Group, 2003. 408 p.
6. Dunn P. K., Smyth G. K. Generalized Linear Models with Examples in R // Springer Texts in Statistics. 2018. 562 p. DOI: 10.1007/978-1-4419-0118-7_4
7. Dunn P K., Smyth G K. Randomized quantile residuals // Journal of Computational and Graphical Statistics. 1996. Vol. 5. P. 236–244. DOI: 10.2307/1390802
8. Fox J., Weisberg S. An R Companion to Applied Regression. Sage, 2019. 608 p.
9. Hosmer D. W., Lemeshow Jr. S., Sturdivant R. X. Applied Logistic Regression. Canada: Wiley, 2013. 528 p.
10. Jakaitiene A. Nonlinear Regression Models // Encyclopedia of bioinformatics and computational biology. 2019. Vol. 1. P. 731–737. DOI: 10.1016/B978-0-12-809633-8.20361-0
11. Jørgensen B. Generalized Linear Models // Encycpedialo of Environmetrics / A. H. El-Shaarawi, W. W. Piegorsch, eds. Chichester: Wiley. 2013. P. 1152–1159.
12. Giner G., Smyth G. K. statmod: probability calculations for the inverse Gaussian distribution // The R Journal. 2016. Vol. 8. P. 339–351. DOI: 10.32614/RJ-2016-024
13. Logan M. Biostatistical Design and Analysis Using R. A Practical Guide. Wiley-Blackwell, 2010. 546 p.

14. Meyer D., Zeileis A., Hornik K. The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd. // Journal of Statistical Software. 2006. Vol. 17. P. 1–48. DOI: 10.18637/jss.v017.i03
15. O'Brien R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors // Qual Quant. 2007. Vol 41. P 673–690. DOI: 10.1007/s11135-006-9018-6
16. Olson D. M. et al. Terrestrial Ecoregions of the World: A New Map of Life on Earth // BioScience. 2001. Vol. 51. No. 11. P. 933–938. DOI: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2
17. Schepaschenko D., Shvidenko A., Usoltsev V., Lakyda P., et al. A database of forest biomass structure for Eurasia // Scientific Data. 2017. DOI: 10.1594/PANGAEA.871491
18. Schmettow M. New Statistics for Design Researchers: A Bayesian Workflow in Tidy R. Springer Nature, 2021. 471 p.
19. Sing T., Sander O., Beerenwinkel N., Lengauer T. ROCR: visualizing classifier performance in R // Bioinformatics. 2005. Vol. 21. P. 7881. DOI: 10.1093/bioinformatics/bti623
20. Zakharov K., Mizgajski A. Socioeconomic and political settings for the land development decreasing urban green. Inside view from Moscow // Land Use Policy. 2024. Vol. 141. P. 107153. DOI: 10.1016/j.landusepol.2024.107153
21. Zhang Wenjun. SampSizeCal: The platform-independent computational tool for sample sizes in the paradigm of new statistics // Network Biology. 2024. Vol. 14. P. 100–155.
22. Zuur A. F., Ieno E. N., Elphick C. S. A Protocol for data exploration to avoid common statistical problems // Methods in ecology and evolution. 2010. Vol. 1. P. 3–14. DOI: 10.1111/j.2041-210X.2009.00001.x

## REFERENCES

1. Zamyatina N. Yu., Kotov E. A., Goncharov R. V., Burceva A.V., Grebenets V.I., Medvedkov A.A., Molodtsova V.A., Klyueva V.P., Kulchitskii Yu. V., Mironova B.A., Nikitin B.V., Pilyasov A.N., Polyachenko A.E., Poturaeva A.V., Streletskii D.A., Shamalo I.A.. [Resilience potential of the Russian Arctic cities]. In: *Vestnik Moskovskogo universiteta. Seriya 5: Geografiya* [Bulletin of Moscow University. Series 5: Geography], 2022, no. 5, pp. 52–65.
2. Breheny P., Burchett W. Visualization of Regression Models Using visreg. In: *The R Journal,* 2017, vol. 9, pp. 56–71. DOI: 10.32614/RJ-2017-046
3. Corlatti L. Regression models, fantastic beasts, and where to find them: a simple tutorial for ecologists using R. In: *Bioinformatics and Biology Insights,* 2021, vol. 15, pp. 1–19. DOI: 10.1177/11779322211051522
4. Chow S., Shao J., Wang H. *Sample Size Calculations in Clinical Research*. NY: Basel, 2008. 358 p.
5. Collet D. *Modelling Binary Data*. Taylor & Francis Group, 2003. 408 p.
6. Dunn P. K., Smyth G. K. Generalized Linear Models with Examples in R. In: *Springer Texts in Statistics,* 2018. 562 p. DOI: 10.1007/978-1-4419-0118-7_4
7. Dunn P K., Smyth G K. Randomized quantile residuals. In: *Journal of Computational and Graphical Statistics*, 1996, vol. 5, pp. 236–244. DOI: 10.2307/1390802
8. Fox J., Weisberg S. *An R Companion to Applied Regression*. Sage, 2019. 608 p.
9. Hosmer D. W., Lemeshow Jr. S., Sturdivant R. X. *Applied Logistic Regression*. Canada: Wiley, 2013. 528 p.
10. Jakaitiene A. Nonlinear Regression Models. In: *Encyclopedia of bioinformatics and computational biology*, 2019, vol. 1, pp. 731–737. DOI: 10.1016/B978-0-12-809633-8.20361-0
11. Jørgensen B. Generalized Linear Models. In: El-Shaarawi A. H., Piegorsch W. W., eds. *Encycpedialo of Environmetrics*. Chichester: Wiley, 2013. P. 1152–1159.
12. Giner G., Smyth G. K. statmod: probability calculations for the inverse Gaussian distribution. In: *The R Journal*, 2016, vol. 8, pp. 339–351. DOI: 10.32614/RJ-2016-024
13. Logan M. *Biostatistical Design and Analysis Using R. A Practical Guide*. Wiley-Blackwell, 2010. 546 p.
14. Meyer D., Zeileis A., Hornik K. The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd. In: *Journal of Statistical Software*, 2006, vol. 17, pp. 1–48. DOI: 10.18637/jss.v017.i03
15. O'Brien R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. In: *Qual Quant,* 2007, vol 41, pp. 673–690. DOI: 10.1007/s11135-006-9018-6
16. Olson D. M. et al. Terrestrial Ecoregions of the World: A New Map of Life on Earth. In: *BioScience*, 2001, vol. 51, no. 11, pp. 933–938. DOI: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2
17. Schepaschenko D., Shvidenko A., Usoltsev V., Lakyda P., et al. A database of forest biomass structure for Eurasia. In: *Scientific Data*, 2017. DOI: 10.1594/PANGAEA.871491

18. Schmettow M. *New Statistics for Design Researchers: A Bayesian Workflow in Tidy R.* Springer Nature, 2021. 471 p.
19. Sing T., Sander O., Beerenwinkel N., Lengauer T. ROCR: visualizing classifier performance in R. In: *Bioinformatics*, 2005, vol. 21, p. 7881. DOI: 10.1093/bioinformatics/bti623
20. Zakharov K., Mizgajski A. Socioeconomic and political settings for the land development decreasing urban green. Inside view from Moscow. In: *Land Use Policy,* 2024, vol. 141, pp. 107153. DOI: 10.1016/j.landusepol.2024.107153
21. Zhang Wenjun. SampSizeCal: The platform-independent computational tool for sample sizes in the paradigm of new statistics. In: *Network Biology*, 2024, vol. 14, pp. 100–155.
22. Zuur A. F., Ieno E. N., Elphick C. S. A Protocol for data exploration to avoid common statistical problems. In: *Methods in ecology and evolution*, 2010, vol. 1, pp. 3–14. DOI: 10.1111/j.2041-210X.2009.00001.x

## ИНФОРМАЦИЯ ОБ АВТОРАХ

*Захаров Константин Валентинович* (г. Москва) – кандидат биологических наук, доцент кафедры зоологии, экологии и охраны природы имени А. Г. Банникова факультета биотехнологии и экологии Московской государственной академии ветеринарной медицины и биотехнологии – МВА имени К. И. Скрябина;
e-mail: k.v.zaharov@gmail.com, ORCID: 0000-0002-1620-3895

*Коновалов Александр Михайлович* (г. Москва) – кандидат сельскохозяйственных наук, заведующий кафедрой зоологии, экологии и охраны природы имени А. Г. Банникова факультета биотехнологии и экологии Московской государственной академии ветеринарной медицины и биотехнологии – МВА имени К. И. Скрябина;
e-mail: zoolog82@mail.ru; ORCID: 0000-0002-4050-0259

*Ломсков Михаил Александрович* (г. Москва) – кандидат биологических наук, доцент кафедры зоологии, экологии и охраны природы имени А. Г. Банникова факультета биотехнологии и экологии Московской государственной академии ветеринарной медицины и биотехнологии – МВА имени К. И. Скрябина;
e-mail: lomskovma@mail.ru, ORCID: 0000-0001-6579-0048

## INFORMATION ABOUT THE AUTHORS

*Konstantin V. Zakharov* (Moscow) – PhD (Biology), Assoc. Prof., Department of Zoology, Ecology and Nature Protection named after A. G. Bannikov, Faculty of Biotechnology and Ecology, Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K. I. Skryabin;
e-mail: k.v.zaharov@gmail.com, ORCID: 0000-0002-1620-3895

*Alexandr M. Konovalov* (Moscow) – PhD (Agricultural), Departmentally Head, Department of Zoology, Ecology, and Nature Conservation named after A. G. Bannikov, Faculty of Biotechnology and Ecology, Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K. I. Skryabin;
e-mail: zoolog82@mail.ru; ORCID: 0000-0002-4050-0259

*Mikhail A. Lomskov* (Moscow) – PhD (Biology), Assoc. Prof., Department of zoology, ecology and nature protection named after A. G. Bannikov, Faculty of Biotechnology and Ecology, Moscow State Academy of Veterinary Medicine and Biotechnology – MVA named after K. I. Skryabin;
e-mail: lomskovma@mail.ru, ORCID: 0000-0001-6579-0048