## Регрессионный анализ, часть 2. Домашнее задание, вариант 5.

В.В.Решетникова

### Введение

В этой работе мы проверим, можно ли определить возраст моллюсков на основании данных морфометрии, иными словами, существует ли линейная зависимость числа колец на спиле раковины от этих факторов. Это поможет упростить метод подсчёта возраста моллюсков, потому что точный подсчёт колец на спиле их раковины - очень долгий и трудоёмкий процесс [1, 2].

### Методы

В исследовании Nash et al. (1994) [2] были изучены такие морфометрические признаки моллюсков *Haliotis rubra* (на 1310 экземплярах), как длина раковины (length), её диаметр (diameter), вес (shell weight), число колец (rings), а также общий вес моллюска (whole weight) и другие в зависимости от их гендера (sex).

Подбор линейной модели зависимости числа колец на спиле раковины от различных факторов был осуществлён с помощью метода наименьших квадратов [3]. Для тестирования статистической значимости зависимости между исследованными признаками использовался дисперсионный анализ [3], а для оценки условий применимости модели - построение графика зависимости стандартизованных остатков от предсказанных значений [4]. Корректировка модели проведена при помощи преобразования её в модель, основанную на распределении Пуассона (так как зависимая переменная - это счётная величина) [5].

Расчеты проведены в R [6, 7] с использованием функций lm() и glm() из пакета stats [7]; ggplot() и fortify() из пакета ggplot2 [8]; Anova(), residualPlot() и qqPlot() из пакета car [9]; groupp\_by(), do(), ungroup() из пакета dplyr [10] и glm.nb() из пакета MASS [11].

### Результаты и обсуждение

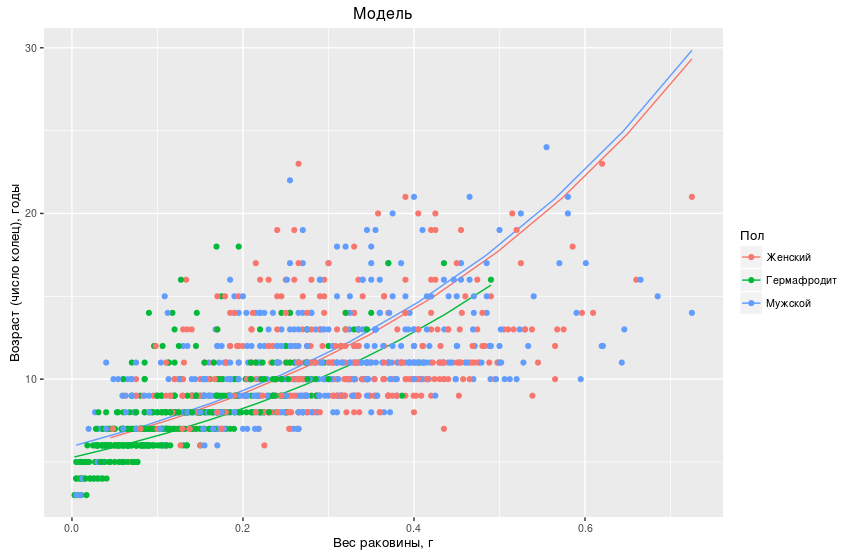
Первая подобранная модель оказалась невалидной из-за гетероскедастичности, поэтому была подобрана модель, основанная на распределении Пуассона (так как зависимая переменная - это счётная величина). Оказалось, что возраст моллюсков (число колец на раковине) *Haliotis rubra* зависит от гендера, диаметра и веса раковины, а также от общего веса моллюска (см. табл. 1). Эта зависимость описывается уравнением rings = 1.62 + 0.02sex + 1.47diameter - 0.55whole-weight + 2.22shell-weight (см. диаг. 1).

**Таблица 1.** Таблица коэффициентов регрессии с результатами z-тестов.

|  | Estimate | z value | Pr(>) |
| --- | --- | --- | --- |
| sex | 0.02 | 0.85 | < 0.001 |
| diameter | 1.47 | 5.79 | < 0.001 |
| whole\_weight | -0.55 | -8.55 | < 0.001 |
| shell\_weight | 2.22 | 10.7 | < 0.001 |

*Согласно z-тесту, существует значимая ( < 0.001) зависимость возраста моллюсков (число колец на раковине) от гендера, диаметра и веса раковины, а также от общего веса моллюска.*

**Диаграмма 1.** Модель, основанная на распределении Пуассона.



*На диграмме показана зависимость возраста моллюсков (число колец на раковине) от веса раковины. Красная, голубая и зелёная линии - линии регрессии для женского, мужского гендера и гермафодитов, соответственно. По оси ох - вес раковины, г; по оси оу - возраст (число колец), годы. Красные точки - данные моллюсков женского гендера, голубые - мужского, зелёные - гермафродитов.*

### Выводы

Наше исследование показало, что для определения возраста халиотисов можно использовать данные об их гендере и морфометрических характеристиках (диаметр и вес раковины, а также общий вес моллюсков). Это поможет в значительной степени облегчить оценку их возраста и избежать трудоёмкого подсчёта колец на спиле раковины.

### Список литературы

[1] Yoneyama S. (2010). Formation of Shell Growth Rings in the Abalone, *Haliotis gigantea* from Izu-Oshima. Aquaculture Science Vol. 39 (1991) No. 2 P 181-188. Retrieved October 24, 2016, from <https://www.jstage.jst.go.jp/article/aquaculturesci1953/39/2/39_2_181/_article>

[2] Nash, W.J., Sellers, T.L., Talbot S.R., Cawthorn A.J., Ford W.B., 1994. The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. Sea Fisheries Division, Technical Report No. 48

[3] Brain S., Hotorn T. (2010). Handbook of Statistical Analyses Using R. CRC Press.

[4] Quinn, G. G. P., & Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge University Press.

[5] Zuur, A.F. et al. (2009). Mixed effects models and extensions in ecology with R. Statistics for biology and health. Springer, New York, NY.

[6] RStudio (2012). RStudio: Integrated development environment for R (Version 0.99.903). Boston, MA. URL <http://www.rstudio.org/>

[7] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

[8] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[9] John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>

[10] Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. <https://CRAN.R-project.org/package=dplyr>

[11] Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0