GLM Model Selection and Validation

Cóilín Minto, Olga Lyashevska

Marine and Freshwater Research Centre
Atlantic Technological University
Galway, Ireland

July 15th 2022

Ollscoil
Teicneolaíochta
an Atlantaigh

Atlantic
Technological
University

Outline

1. Model Selection

2. Model Validation

3. Principles for improved statistical ecology

All models are wrong but some are useful

- Statistical models are mathematical approximations to reality that represent the important features of data for the task at hand.

All models are wrong but some are useful

- Statistical models are mathematical approximations to reality that represent the important features of data for the task at hand.
- The purpose of a statistical model effects how it is developed: prediction versus interpretation (association not cause-effect)

All models are wrong but some are useful

- Statistical models are mathematical approximations to reality that represent the important features of data for the task at hand.
- The purpose of a statistical model effects how it is developed: prediction versus interpretation (association not cause-effect)
- There are numerous predictors that could be chosen. How do we choose a statistical model? Statistical models are based on underlying theory, or from an understanding of the biological features, and are built with this knowledge in mind.

Criteria for model selection

An adequate statistical model balances two criteria:

- **Accuracy:** The model should accurately describe both the systematic (how the mean response changes as the explanatory variables change) and the random (the variation of the data about the mean) component;

Criteria for model selection

An adequate statistical model balances two criteria:

- **Accuracy:** The model should accurately describe both the systematic (how the mean response changes as the explanatory variables change) and the random (the variation of the data about the mean) component;
- **Parsimony:** The model should be as simple as possible. The simplest accurate model is the preferred model. Complex models may fit the given data well but usually do not generalize well to other data sets (over-fitting);

Why model selection?

1. The deviance of the model (reciprocally the likelihood and the R^2) always decreases (increases) with the inclusion of more predictors – no matter whether they are significant or not.

Why model selection?

1. The deviance of the model (reciprocally the likelihood and the R^2) always decreases (increases) with the inclusion of more predictors – no matter whether they are significant or not.
2. The excess of predictors lead to a larger variability in the estimation of the model which results in lower precision.

Why model selection?

1. The deviance of the model (reciprocally the likelihood and the R^2) always decreases (increases) with the inclusion of more predictors – no matter whether they are significant or not.
2. The excess of predictors lead to a larger variability in the estimation of the model which results in lower precision.
3. Multicollinearity may hide significant variables, change the sign of them, and result in an increase of the variability of the estimation.

Comparing 2 nested models

$$\mu_A = \beta_0 + \ldots + \ldots + \beta_3 \tag{1}$$

$$\mu_B = \beta_0 + \beta_1 + \beta_2 + \beta_3 \tag{2}$$

Comparing 2 nested models

$$\mu_A = \beta_0 + \ldots + \ldots + \beta_3 \quad (1)$$

$$\mu_B = \beta_0 + \beta_1 + \beta_2 + \beta_3 \quad (2)$$

Model (1) is nested in Model (2), since Model (1) is a special case of Model (2) obtained by setting β_1 and β_2 to 0. In comparing these models, we wish to know whether the more complex model is necessary.

Model selection

- We can look at AICs:

  ```
  AIC(mod1)
  ```

  Lower AIC values gives better model.
- Compare 2 nested models

  ```
  anova(mod1, mod2)
  ```
- Stepwise selection through *MASS::stepAIC*

  ```
  MASS::stepAIC(mod2, trace = FALSE)
  ```
- Automated multi-model selection through *MuMIn::dredge* optional choices of other model properties

  ```
  MuMIn::dredge()
  ```

Outline

Diagnostics

The estimation and inference from the regression model depends on several assumptions. There are three categories:

- **Error**: We assume that $\epsilon \sim N(0, \sigma^2)$ or in other words, that the errors are normally distributed with mean zero and equal variance σ^2.
- **Model:** We assume that the model is correct
- **Unusual observations:** can have a dramatic effect and need to be detected!

Checking error assumptions

- independent for different data values (name?, example?);
- have mean zero;
- have constant variance (name? example?);
- follow a Normal distribution.

Checking error assumptions

- independent for different data values (name?, example?);
- have mean zero;
- have constant variance (name? example?);
- follow a Normal distribution.

Which assumption is the most important?

Relative importance of assumptions

- Independence of the errors (zero autocorrelation) is the most important and also most difficult to accommodate when fails (observations are close in time or space);

Relative importance of assumptions

- Independence of the errors (zero autocorrelation) is the most important and also most difficult to accommodate when fails (observations are close in time or space);
- Constant variance (homoscedasticity) is intermediate, in that nonconstant variance (heteroscedasticity) can have a substantial effect on inferences, but can also be accommodated in many situations.

Relative importance of assumptions

- Independence of the errors (zero autocorrelation) is the most important and also most difficult to accommodate when fails (observations are close in time or space);
- Constant variance (homoscedasticity) is intermediate, in that nonconstant variance (heteroscedasticity) can have a substantial effect on inferences, but can also be accommodated in many situations.
- Normal distribution is the least important assumption, especially for large sample size. Normal distribution of errors is needed for inference. Some models depend more on normality than others.

Violation of assumptions

Violation of assumptions can lead to bias in:

- regression coefficients
- standard errors
- confidence intervals
- significance tests

Violation of assumptions

Violation of assumptions can lead to bias in:

- regression coefficients
- standard errors
- confidence intervals
- significance tests

The quality of inference depends on how well the errors ϵ_i conform to our assumptions.

Violation of assumptions

Violation of assumptions can lead to bias in:

- regression coefficients
- standard errors
- confidence intervals
- significance tests

The quality of inference depends on how well the errors ϵ_i conform to our assumptions. We do not observe the errors. So how do we check these assumptions?

Violation of assumptions

Violation of assumptions can lead to bias in:

- regression coefficients
- standard errors
- confidence intervals
- significance tests

The quality of inference depends on how well the errors ϵ_i conform to our assumptions. We do not observe the errors. So how do we check these assumptions? The closest we can get to the errors are residuals r_i.

Residuals

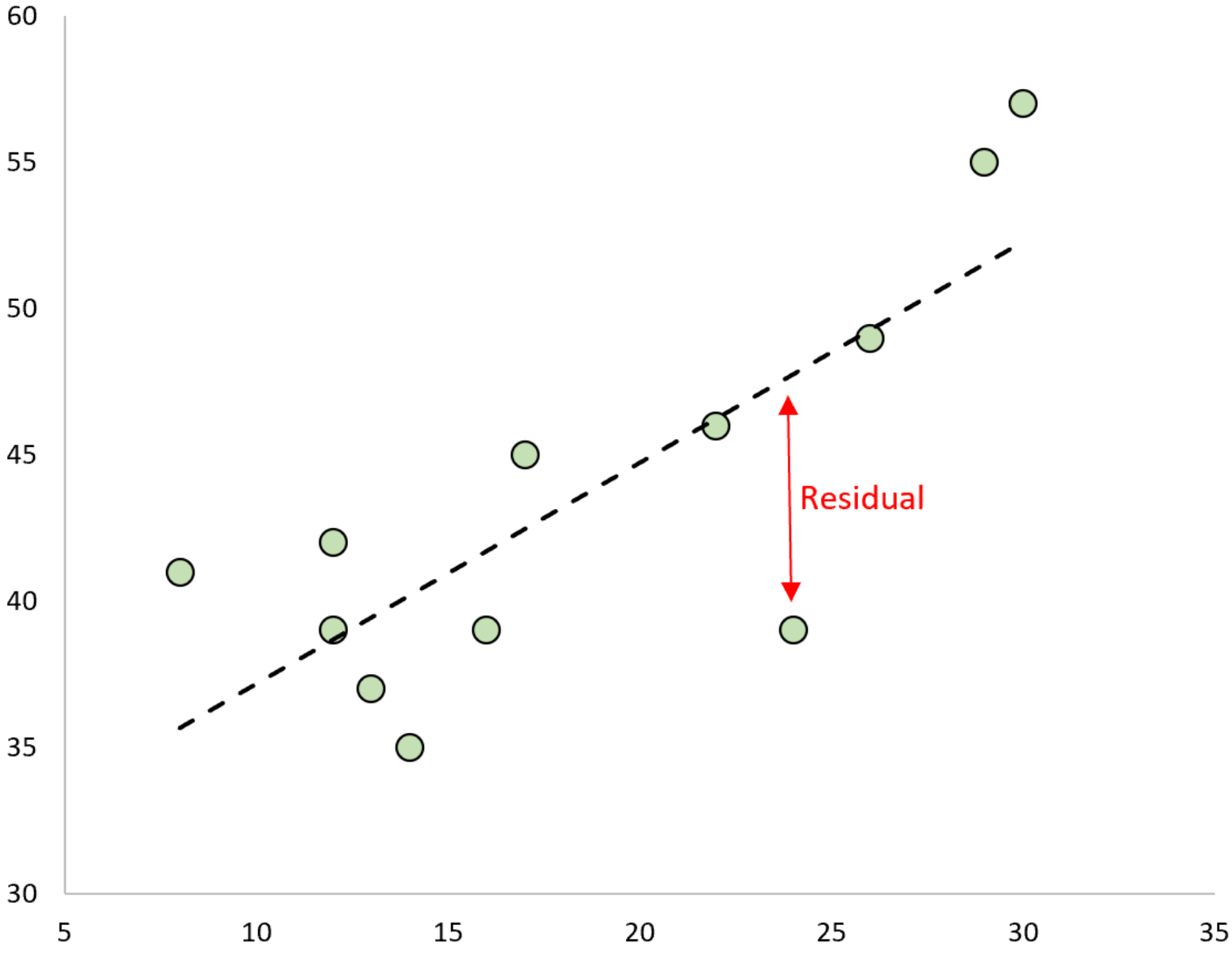
What exactly are residuals?

Residuals

What exactly are residuals?

Regardless the structure and complexity of the model the raw residuals r_i are simply the differences between the data y_i and the fitted value y_i

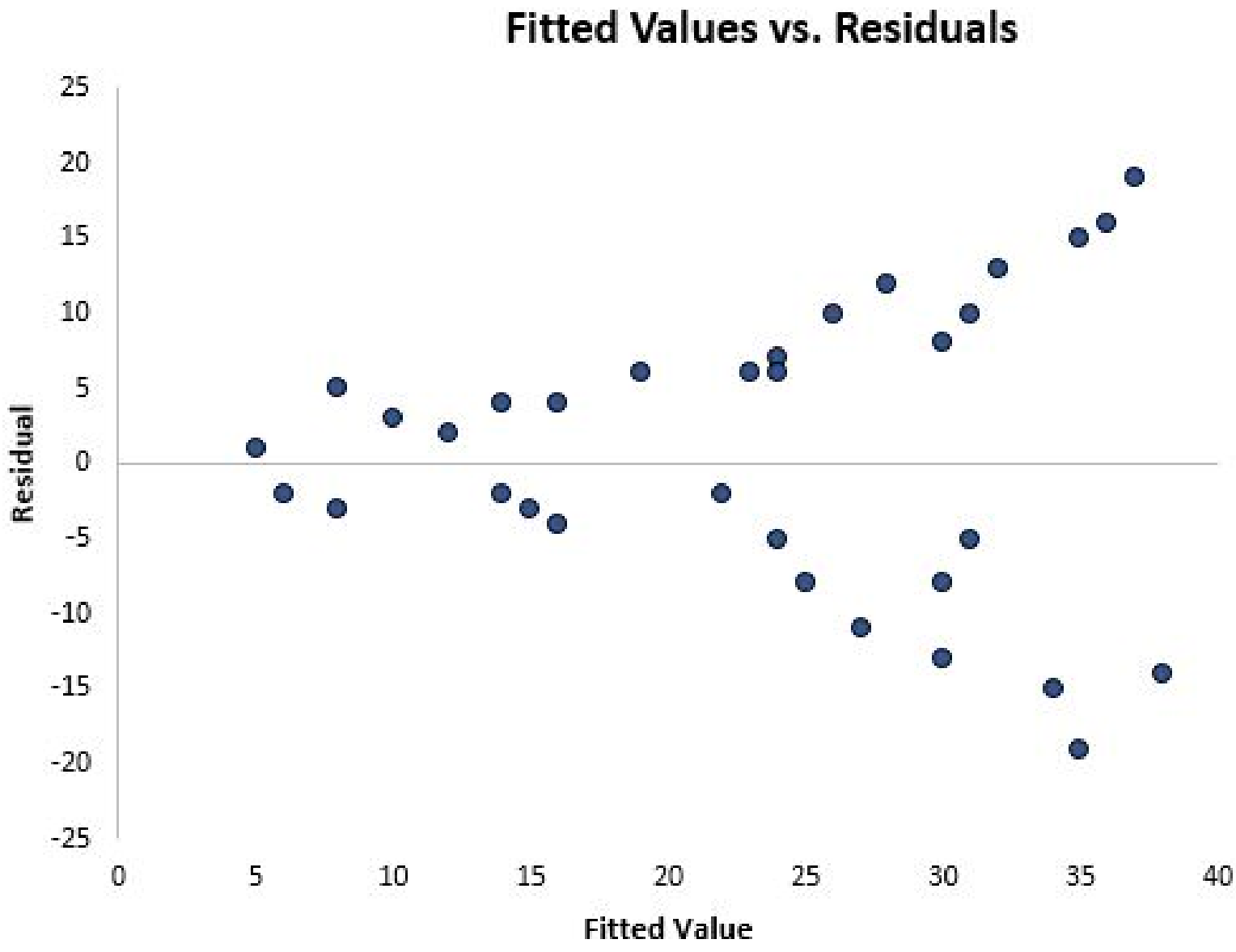
Residuals



Heteroscedasticity

- Plot the residuals r_i against the fitted values $y_i - r_{ij}$
- If the variance is constant, the vertical spread of the points will be about the same.
- Nonconstant variance is revealed as a pattern in the spread of the residuals.

Heteroscedasticity



Outline

Four principles for improved statistical ecology

ISEC 2022, Session 18. Statistical Theory. Gordana Popovic

1. First define a focused research question, then plan sampling and analysis to answer it;
2. Develop a model that accounts for the distribution and characteristics (dependencies) of your data;
3. Emphasise effect sizes to replace statistical significance (p-values) with ecological relevance;
4. Report you methods and finding in sufficient detail so that your research is valid and reproducible;

References

- Cleasby, I. R., and Nakagawa, S. (2011). Neglected biological patterns in the residuals: A behavioural ecologist's guide to co-operating with heteroscedasticity. Behavioral Ecology and Sociobiology, 65(12), 2361–2372. http://www.jstor.org/stable/41414703