Bio Data Science

The definitive guide

Prof. Dr. Jochen Kruppa-Scheetz

2025-10-26

Table of contents

Preface

"Life is difficult." — M. Scott Peck's, The Road Less Traveled

This cookbook contains an extensive collection of recipes from the fields of biostatistics, biometrics, statistics, bioinformatics and R programming. These will enable you to gain in-depth knowledge of these subject areas for data analysis, eliminating the need to attend my courses. Feel free to browse through the book to see if anything interests you, and then put together your own menu. I am continuously updating the book and adding to it all the time. In addition to the written content, there are also explanatory YouTube videos available. Either way, I am delighted that you are interested in learning something new, whether through personal interest or exam preparation. Either way, I recommend that you just take a look around. Don't be put off by the sheer volume of material – it just happened that way.

A German version of this OpenBook Bio Data Science is available. The page can be found at the following link: https://jkruppa.github.io/. However, it is much more unstructured. It is more like a cookbook than this book, which has a continuous storyline.

Part I

From science to data by models

Last modified on 27. October 2025 at 13:16:42

> *"I'd like to solve the puzzle." — Wheel of Fortune*

Here comes the preface text

1 What is science?

Last modified on 28. October 2025 at 15:38:26

> *"A quote." — Dan Meyer*

1.1 General background

Idea of hypotheses

Science is guessing and falsification

1.2 Theoretical background

1.3 R packages used

1.4 Data

1.5 Alternatives

Further tutorials and R packages on XXX

1.6 Glossary

term what does it mean.

1.7 The meaning of "Data Science" in this chapter.

- itemize with max. 5-6 words

1.8 Summary

References

2 What is data?

Last modified on 28. October 2025 at 15:38:28

> *"The limits of my language mean the limits of my world." — Ludwig Wittgenstein*

2.1 General background

2.2 Theoretical background

2.3 R packages used

2.4 Data

```r
jump_weight_tbl <- tibble(x = c(0.6, 1, 2.3, 3.5, 5.2, 7.1, 8.4, 9.2, 10),
                          y = 0.15*x^3 - 2.2*x^2 + 8.8*x + 3.2 + rnorm(9, 0, 0.5)) |>
  mutate_all(round, 1) |>
  rename(weight_mg = x, jump_length_cm = y)
```

Table 2.1: foo.

weight_mg	jump_length_cm
0.6	8.4
1.0	9.3
2.3	13.6
3.5	13.3
5.2	10.2
7.1	8.4
8.4	10.8
9.2	14.5
10.0	21.3

Equation 2.1

$$y = 0.15 \cdot x^3 - 2.2 \cdot x^2 + 8.8 \cdot x + 3.2 \tag{2.1}$$

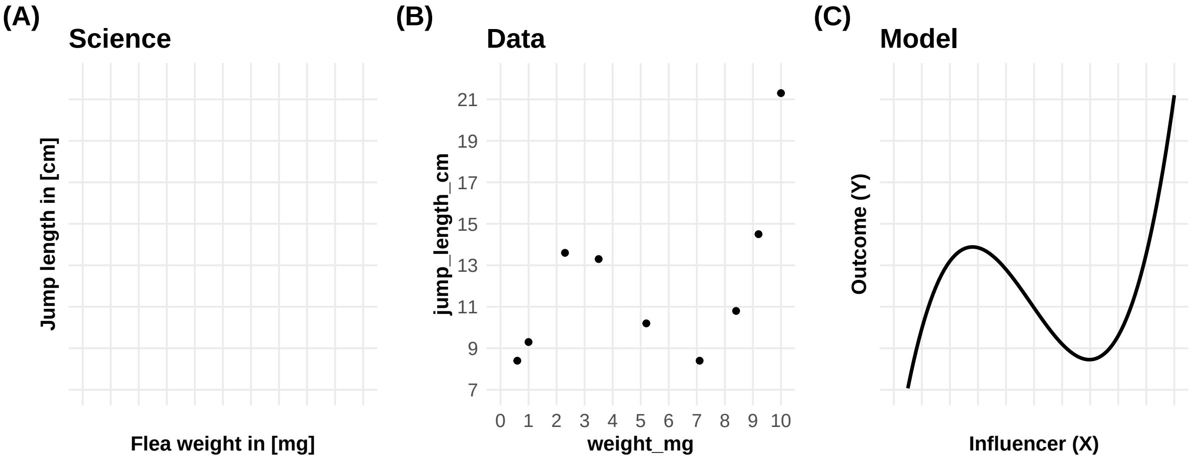

Figure 2.1: foo.

2.5 Alternatives

Further tutorials and R packages on XXX

2.6 Glossary

term what does it mean.

2.7 The meaning of "Data Science" in this chapter.

- itemize with max. 5-6 words

2.8 Summary

References

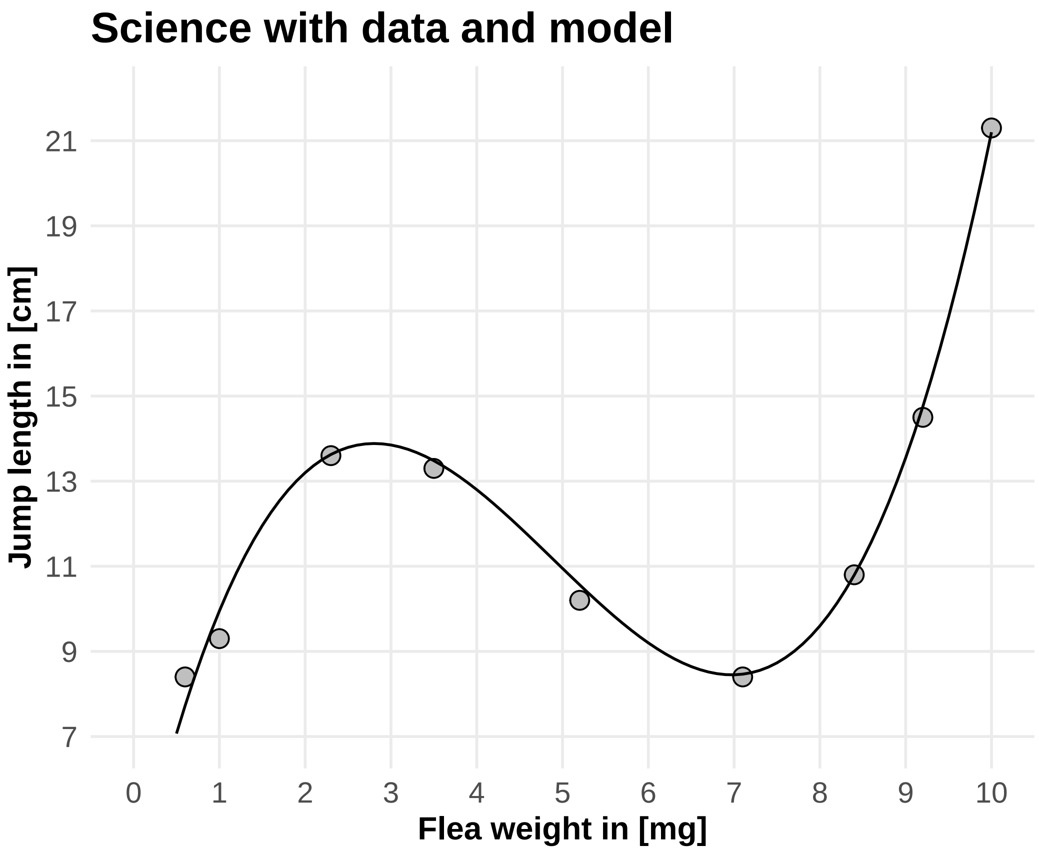

Figure 2.2: foo.

3 What is a model?

Last modified on 28. October 2025 at 15:37:34

> *"A quote." — Dan Meyer*

3.1 General background

Table 3.1: Table of terms used in the statistical modelling. The terms in bold are used here. Depending on the scientific background, the usage of these terms can vary widely.

Application	Symbol	Name	Description
	y	**outcome**, response, endpoint, dependent variable	*The right-hand side (abbr. RHS) of the model. Describing the values measured in an experiment or study.*
	x	**influencer**, influential variable, risk factor, fixed effect, independent variable	*The left-hand side (abbr. LHS) of the model. Describing the influential variables in an experiment or study.*
	z	**random effect**	*A factor that provides a description of an grouping variable, which is not part of the controlled experimental setting.*
Explanation	x	**explanator**, explanatory variable	*The influencer is used to describe or explain the outcome.*
Prediction	x	**predictor**, predictive variable	*The influencer is used to predict the outcome.*
Main effect	x	**focal explanator**, **focal predictor**, focal variable	*In a model with multiple influencers, the focal variable is the variable of primary interest.*
Continuous x	c	**covariate**, covariable	*The influencer is a numeric variable with continuous values.*
Categorical x	f_A	**factor A**, factorial variable, categorical variable	*The influencer is discrete, functioning as a grouping variable, such as an experimental group or a treatment.*

Application	Symbol	Name	Description
Factor f_A	$A.1$ to $A.j$	**levels**, groups, treatment groups	*The discrete groups included in one factor A.*

A sentence why we use y and not x for mean and other stuff.

3.2 Theoretical background

3.3 R packages used

3.4 Data

3.5 Alternatives

Further tutorials and R packages on XXX

3.6 Glossary

term what does it mean.

3.7 The meaning of “Data Science” in this chapter.

- itemize with max. 5-6 words

3.8 Summary

References

4 Data examples

Last modified on 06. November 2025 at 17:07:41

> *"A quote."* — *Dan Meyer*

4.1 General background

4.2 Theoretical background

4.3 R packages used

4.4 Data

Based on standard methods in flea research and experimental entomology, as well as a similar published experiment, the following three types of food for adult cat fleas are possible, representing different nutritional conditions:

Blood: This is the natural and optimal food source for adult fleas.Often, defibrinated or anticoagulated animal blood (e.g. bovine blood, rabbit blood) is used for this purpose, which is offered in special in vitro feeding systems (e.g. through a membrane).Expectation: Fleas that are optimally nourished should show the greatest jumping distance.

Sugar water: Often serves as a 'control feed' or as a feed that provides energy (sugar) but lacks essential nutrients (such as proteins from blood).Expectation: Jumping distance could be reduced due to the lack of blood and thus the proteins important for reproduction, which could impair physiological fitness.

Ketchup - a nutrient-poor or unsuitable food: This option is used to simulate poor, incomplete or stressful nutritional conditions.In a similar documented experiment, ketchup (a combination of sugar, vinegar and minimal other substances, but no blood) was used as the third feed.Expectation: It can be assumed that fleas will show the shortest jumping distance under these conditions, as they lack both essential nutrients and the necessary energy.

4.5 Alternatives

Further tutorials and R packages on XXX

4.6 Glossary

term what does it mean.

4.7 The meaning of "Data Science" in this chapter.

- itemize with max. 5-6 words

4.8 Summary

References

Part II

Template Preface

Last modified on 25. October 2025 at 20:17:25

> *"What problem have you solved, ever, that was worth solving where you knew all the given information in advance? No problem worth solving is like that. In the real world, you have a surplus of information and you have to filter it, or you don't have sufficient information and you have to go find some." — Dan Meyer in Math class needs a makeover*

Here comes the preface text

5 Template chapter

Last modified on 27. October 2025 at 14:15:06

> *"A quote." — Dan Meyer*

5.1 General background

5.2 Theoretical background

5.3 R packages used

5.4 Data

5.5 Alternatives

Further tutorials and R packages on XXX

5.6 Glossary

term what does it mean.

5.7 The meaning of "Data Science" in this chapter.

- itemize with max. 5-6 words

5.8 Summary

References