SS2: Foundations in Statistics

SS2.2: Challenges with data and regression analysis

Professor Caroline Brophy

School of Computer Science and Statistics
Trinity College Dublin

LegumeLegacy Event 2, Dublin, Ireland

November 2023

LegumeLegacy funding

This project has received funding from the European Union's Horizon 2021 doctoral network programme under the Marie Skłodowska-Curie grant agreement No. 101072579.

Summary of SS2: Foundations in Statistics II

- SS2.1: Regression Analysis
  - Simple linear regression and One-way ANOVA models
  - Two-way ANOVA models
  - Multiple regression models
- SS2.2: Challenges with data and regression analysis (this session)
- SS2.3: The design and implementation of experiments and the management of data

SS2.2: Challenges with data and regression analysis

SS2.2.1: Warning signs with regression modelling

The importance of assessing model assumptions

Any statistical model comes with assumptions.

When you are modelling data:

- It is your responsibility to know what the model assumptions are.
- It is your responsibility to ensure that the model assumptions are reasonably satisfied.

If your model assumptions are not satisfied, it may mean that your inference is not valid. Your science is too important and with high impact potential to let this happen!

How to detect issues in a multiple regression model

- State the model assumptions
- Assess each of them one by one! Some of this might involve questioning the data collection process, while some of it may involve examining the model residuals.
- We looked at model assumptions in SS2.1, how should these be assessed?

1. $E[\epsilon_i] = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
3. ϵ_i are independent.
4. $\epsilon_i \sim N(0, \sigma^2)$.

What to do when your model may not satisfy assumptions

A starting point for how a regression model may be improved is by:

- Including new predictors.
- Transforming a predictor.
- Transforming the response.

It may also be necessary to:

- Adapt the model to allow for more flexible or different model assumptions.
- Consider a different or more complex modelling framework.

Remember the route to solving a problem may not be unique.

SS2.2.2: General challenges with data

The importance of data source and data screening

- Rubbish in, rubbish out...

- Machines and humans make errors!

Common types of challenges with data

- Missing values (what *type* of missingness is it?)
- Hierarchical data, e.g., data collected over space and time or multiple responses collected on individuals.
- Non-Gaussian responses
- Multivariate data

SS2.2.3: Becoming confident with your data and analyses

Words of advice...

- Before you start: question why you are analysing the data, clearly define what your hypotheses are and identify the purpose of the analysis. Make sure these aspects align!
- Use your common sense!
- Critically assess your data (its source, what has been recorded, potential biases, what population it comes from, screen the data), and your statistical analysis choices and implementation!
- Test your model assumptions!
- Trust your instinct!

SS2.2.4: Discussion (after the lab session)

Repeated measures analysis

The regression models that we looked at in SS2.1 all assumed independence of the Y_i's and ϵ_i's.

$$\text{Var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

If the same experimental unit is measured multiple times, what needs to change? Suppose there are three measurements recorded over time on each experimental unit:

$$\text{Var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \dots & 0 & 0 & 0 \\ \sigma_{12} & \sigma_{22}^2 & \sigma_{23} & \dots & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_{33}^2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ 0 & 0 & 0 & \dots & \sigma_{12} & \sigma_{22}^2 & \sigma_{23} \\ 0 & 0 & 0 & \dots & \sigma_{13} & \sigma_{23} & \sigma_{33}^2 \end{bmatrix}$$

Non-normally distributed response data

What can you do when you have a non-Gaussian distributed response?

- Explore transformations.
- But note that when you transform, it can increase complexity in the interpretation of the model.
- Use the GLM modelling framework with non-Gaussian distributions.

Non-constant variance

- If a residual versus predicted plot indicates that there is changing variance, what should you do?
- Transformations may work (but can increase complexity in the interpretation of the model).
- Explore weighted regression options.

SS2.2.5: Recommended reading

Some suggested reading material

Books

OpenIntro Statistics, by Diez, Çetinkaya-Rundel and Barr.

Applied Linear Regression, by Sandford and Weisberg.

Regression Analysis by Example, by Chatterjee and Hadi.

The Statistical Sleuth, by Ramsey and Schafer.

Statistics for experimenters, by Box, Hunter and Hunter

Papers

Hector, von Felten and Schmid (2010) Analysis of variance with unbalanced data: an update for ecology & evolution. Journal of Animal Ecology, 79, 308–316.

Zuur,Ieno and Elphick (2009) A protocol for data exploration to avoid common statistical problems. Methods in Ecology & Evolution, 1, 3–14.

Box and Cox (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26, 211–52.