Appendix

Derivation of BLUP

The material in this section is not required but it is additional material for those who are interested in knowing more of the background and of the sources that have led to the methods presented earlier in these course notes.

In chapter 3 we have assumed the solution for the mixed linear effects model using BLUP as a given fact without deriving them. So far we were given the recipe to use mixed model equations to produce estimates of fixed effects and predictions of random effects.

There are many derivations and explanations about BLUP. But most of them including the original work by Henderson are not easy to understand. I found the derivation given by [Schaeffer, 2019] to be in the same spirit as the derivation that we have used for the fixed linear effects model. In the chapter about prediction theory (`http://animalbiosciences.uoguelph.ca/~lrs/ABModels/NOTES/predict.pdf`) in the `Notes` section of [Schaeffer, 2019] the BLUP solutions and the mixed model equations are derived in an understandable way.

At this point, we are going to replicate the complete derivation. We rather try to provide some additional explanations which might help in understanding the given derivation. Sections 1 and 2 of the chapter on prediction theory given an introduction and specify the mixed linear effects model. The introduction starts with a definition of the term `prediction`. It has to be noted here that the distinction between `estimation` for fixed effects and `prediction` for random effects is much sharper and much stricter in the English language than it is e.g. in German. The mixed linear effects model is called `General Linear Mixed Model` in the cited reference. But those terms mean the same model which is given by the model equation and by the specified expectations and variance-covariance matrices.

In section 3 some general facts about different predictors are given. These facts are used as an explanation of why the given predictand used in section 4 where BLUP is derived. The term `predictand` is defined as the function of the unknown parameters. The `predictor` is the linear function of the data that is used to predict the predictand. The reason why the predictand is a linear function of the unknown parameters is similar to what was described about estimability in chapter 4 about estimability of linear functions of parameters. The estimation and prediction problems often lead to over-determined systems of linear equations where the unknowns can be expressed as linear combinations of the data. The linear factors with which the data vector is multiplied usually involves some generalized inverses of a matrix. Since these generalized inverses are not unique and because many solutions do exist for the over-determined systems, only predictands are useful which are invariant that means which do not depend on the choice of a specific solution to the system of equations that arise in the prediction problem. Prediction and Estimation Theory has shown that there exist linear functions which are invariant to the choice of the equation solutions. Such functions are called estimable functions and in the context of mixed linear effects models they are written as

$$K^Tb + M^Tu$$

The above shown linear function of the unknown parameter which is to be predicted by a linear function L^Ty of the data y together with the properties of

- unbiasedness and
- minimum error variance

lead to the BLUP solutions for the estimates $\hat{b}$ for the fixed effects and the predictions $\hat{u}$ for the random effects. These correspond to

$$\hat{b} = (X^TV^{-1}X)^{-}X^TV^{-1}y$$

and

$$\hat{u} = GZ^TV^{-1}(y - X\hat{b})$$

In sections 5 and 6 the variance of the predictors and the variance of the prediction error are shown. Section 7 then shows how the mixed model equations produce the same results as found in section 4.

Bibliography

M Adibuzzaman, P DeLaurentis, J Hill, and B D Benneyworth. Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database. *AMIA Annu Symp Proc*, 2017:384–392, 2017. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pubmed/29854102.

J Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, (36):192–236, 1974. URL http://www.jstor.org/stable/2984812.

Peter Buehlmann and Martin Maechler. Computational Statistics Course, 2014.

Bartosz Czech, Magdalena Fraszczak, Magda Mielczarek, and Joanna Szyda. Identification and annotation of breed-specific single nucleotide polymorphisms in Bos taurus genomes. *PLoS ONE*, 13(6):1–9, 2018. ISSN 15498328. doi: 10.1109/TCSI.2014.2341116. URL https://doi.org/10.1371/journal.pone.0198419.

Alois Essl. *Statistische Mehoden in der Tierproduktion*. Österreichischer Agrarverlag, Wien, 1987. ISBN 3-7040-0859-1.

D. S. Falconer and Trudy F. C. Mackay. *Introduction to Quantitative Genetics*. Addison Wesley Longman Limited, Essex, 4 edition, 1996. ISBN 0582-24302-5.

Rohan L Fernando, Hao Cheng, and Dorian J Garrick. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. *Genetics Selection Evolution*, 48(1):80, dec 2016. ISSN 1297-9686. doi: 10.1186/s12711-016-0260-7. URL http://gsejournal.biomedcentral.com/articles/10.1186/s12711-016-0260-7.

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, nov 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596. URL https://doi.org/10.1109/TPAMI.1984.4767596.

Daniel Gianola, Gustavo De Los Campos, William G. Hill, Eduardo Manfredi, and Rohan Fernando. Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363, 2009. ISSN 00166731. doi: 10.1534/genetics.109.103952.

Sasha Issenberg. How Obama's Team Used Big Data to Rally Voters. *MIT Technology Review*, 116(1):38–49, 2013.

David J. Lilja. *Linear regression using R : an introduction to data modeling*. University of Minnesota Libraries Publishing, Minneapolis, 2016. ISBN 9781946135001. URL https://open.umn.edu/opentextbooks/textbooks/linear-regression-using-r-an-introduction-to-data-modeling.

John R Mashey. Big Data ... and the Next Wave of InfraStress, 1998. URL http://static.usenix.org/event/usenix99/invited{_}talks/mashey.pdf.

T H E Meuwissen, B J Hayes, and M E Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157:1819–1829, 2001.

Sameer D. Pant, Flavio S. Schenkel, Chris P. Verschoor, and Niel A. Karrow. Use of breed-specific single nucleotide polymorphisms to discriminate between Holstein and Jersey dairy cattle breeds. *Animal Biotechnology*, 23(1):1–10, 2012. ISSN 10495398. doi: 10.1080/10495398.2012.636224. URL https://doi.org/10.1080/10495398.2012.636224.

R Core Team. R: A Language and Environment for Statistical Computing, 2018. URL https://www.r-project.org/.

L. R. Schaeffer. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123(4):218–223, 2006. ISSN 09312668. doi: 10.1111/j.1439-0388.2006.00595.x.

L R Schaeffer. Animal Models, 2019. URL http://animalbiosciences.uoguelph.ca/{~}lrs/ABModels/.

S R Searle. *Linear Models*. John Wiley & Sons, New York, wiley clas edition, 1971. ISBN 0-471-18499-3.

G. E. Seidel, Jr. Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms. *Reproduction Fertility and Development*, 22(1):138–144, 2010. ISSN 1031-3613. doi: 10.1071/RD09220. URL https://doi.org/10.1071/RD09220.

P.M. VanRaden. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008. ISSN 00220302. doi: 10.3168/jds.2007-0980. URL http://dx.doi.org/10.3168/jds.2007-0980.

Contributors Wikipedia. Big data - Wikipedia, 2019. URL https://en.wikipedia.org/w/index.php?title=Big{_}data{&}oldid=881938730.