Upcoming challenges in genetic evaluation from a statistician's perspective

Stephen D. Kachman

University of Nebraska Lincoln

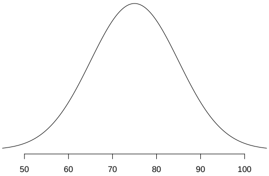
Genetic Prediction Workshop
December 5, 2018

Where we were

- Pedigree information
- Phenotype information
  - Linear traits
  - Threshold traits

Linear Traits

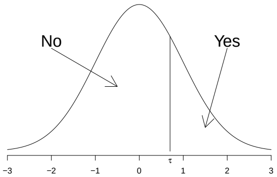
- Assume the data follow a nice bell shaped curve



- Add in fixed effects
- Add in random effects
- Process using linear mixed model machinery

Threshold Traits

- Assume the underlying data follow a nice bell shaped curve



- Add in fixed effects
- Add in random effects
- Process using modified linear mixed model machinery

- Powerful and very flexible
- Very large class of models
- Handle large amounts of data very efficiently
- It just plain works

- Powerful and very flexible
- Very large class of models
- Handle large amounts of data very efficiently
- It just plain works
- The number of dependent variables, fixed and random effects were small
- Sparse system of equations that needed to be solved

Genomics

What has changed?

- Dramatic increase in the number of effects
- Effect covariates were missing for most animals
- System of equations are no longer sparse

- Marker effects (Bell shaped curve again)
    - Model as individual effects
        - Potentially add a chance of the effect being zero
    - Model as the cumulative effect
- Missing covariates
    - Normal theory

$$\mathbf{g}_m|\mathbf{g}_o \dot{\sim} \mathrm{N}(\mathbf{C}_{mo}\mathbf{V}_o^{-1}\mathbf{g}_{\mathrm{o}}, \mathbf{V}_m - \mathbf{C}_{mo}\mathbf{V}_o^{-1}\mathbf{C}_{om})$$

    - Linear predictor

$$\begin{aligned} \widehat{\mathbf{g}}_m &= \mathbf{C}_{mo}\mathbf{V}_o^{-1}\mathbf{g}_{\mathrm{o}} \\ \mathrm{var}(\mathbf{g}_m - \widehat{\mathbf{g}}_m) &= \mathbf{V}_m - \mathbf{C}_{mo}\mathbf{V}_o^{-1}\mathbf{C}_{om} \end{aligned}$$

Two camps

- single step GBLUP
  - Linear mixed model
- single step Bayes
  - Posterior means using MCMC sampling

Two camps

- single step GBLUP
  - Linear mixed model
- single step Bayes
  - Posterior means using MCMC sampling
- Which to choose?
  - Their similarities are much greater than their differences
  - Major differences revolve around what compromises are made

Two camps

- single step GBLUP
  - Linear mixed model
- single step Bayes
  - Posterior means using MCMC sampling
- Which to choose?
  - Their similarities are much greater than their differences
  - Major differences revolve around what compromises are made
  - single step GBLUP
    - Considerable experience working with linear mixed models
  - single step Bayes
    - opens up a broader class of models

- Started with
  - Few dependent variables and effects (many levels)
  - Covariate values were known
  - Linear mixed models based on the normal distribution

- Started with
  - Few dependent variables and effects (many levels)
  - Covariate values were known
  - Linear mixed models based on the normal distribution
- Added Genomics
  - Number of effects greatly increased
  - Many unknown covariate values
  - System of equations were no longer sparse

- Started with
  - Few dependent variables and effects (many levels)
  - Covariate values were known
  - Linear mixed models based on the normal distribution
- Added Genomics
  - Number of effects greatly increased
  - Many unknown covariate values
  - System of equations were no longer sparse
- What other types of data may we see in the future?

Traceability

- The ability to track meat back to its source opens up a number of possibilities
    - What other information is collected at each of the time points
    - Management variables
    - Health information
    - Caracas information

- Errors in variables models
  - Covariates are replaced by proxies due to variation in what is recorded
- Missing covariates
  - Considerable variation in the amount of data recorded

- Errors in variables models
  - Covariates are replaced by proxies due to variation in what is recorded
- Missing covariates
  - Considerable variation in the amount of data recorded
- Hierarchical models where **X** and **Z** are no longer assumed to be known

High throughput data

- Microbiome data
  - Fecal sample
  - Counts for various taxa (some of which are identified to a given taxonomic level others placed in operational taxanomic units)
  - Composition is determined by both environmental factors along with host genetics
  - Interested in selecting animals that:
    - ★ Harbor communities that are resistant to harboring pathogenic taxa
    - ★ Harbor communities that improve feed efficiency

- Individual taxa could be modeled as zero inflated counts (negative binomial)
- Greatly increasing the number of dependent variables
- Interested in communities as the the functional unit
    - Latent variable model
      Genetic and environmental factors operate through unobserved latent variables to influence both community structure and production traits

Epistatic effects

- While modeling interactions between loci is conceptually straight forward
- Implementation is difficult as the number of possible two-way interactions is quadratic in the number of loci
  For example, with 5,000 loci there are over 10 million possible two-way interactions
- Could look at using genetic algorithms
  - Each generation of models compete to produce the next generation of models
  - The result is a population of models including a set plausible models

Summary

- Adding genomics to genetic evaluation has presented a number of challenges
  - missing covariates
  - systems of equations which are no longer sparse
- Introduction of single step methods for genetic evaluation
- We can expect that both the variety and the amount of data available for use in genetic evaluation will only increase
  - Traits that are not well represented by a linear mixed model and its variants
- Which in turn will necessitate a new generation of methods for genetic evaluation