Sparse Modelling

Model Selection

Sparse Modelling

Model Selection

Lots of this is based on work with Josh Jahner, Alex Buerkle and the Modelscape Consortium

Learning objectives

- Understand inference vs in-sample vs out of sample prediction
- Understand reducible vs irreducible error
- Simulate data for analysis
- Run 1 type of sparse analysis on simulations, extract inference, IS, OS predictions
- Contrast results with other techniques

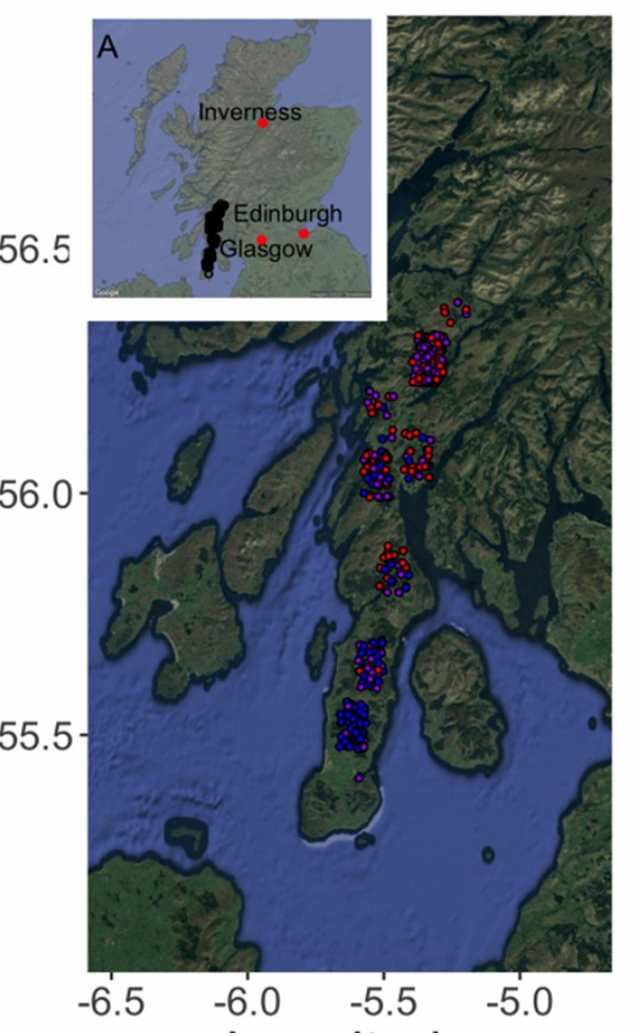
How do we decide which parameters to include?

- Population density~environment
- Phenotype~genotypes (or expression)
- Primary productivity ~ remote sensed river parameters

Purpose of Sparse Modelling

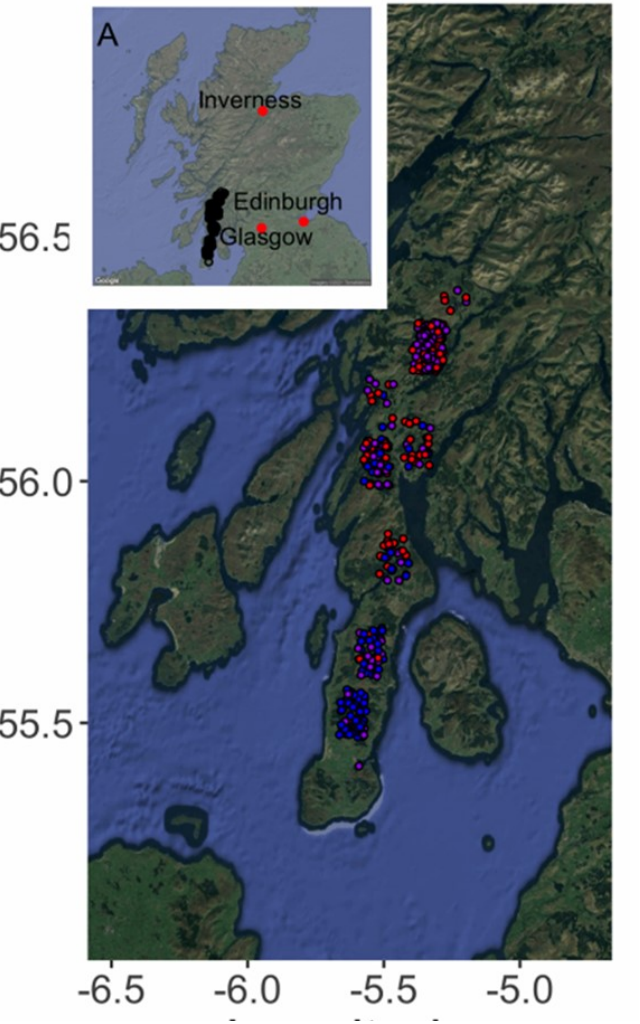
- Choose among predictor values when n<<p
- Increases interpretability of models
- Prevents overfitting

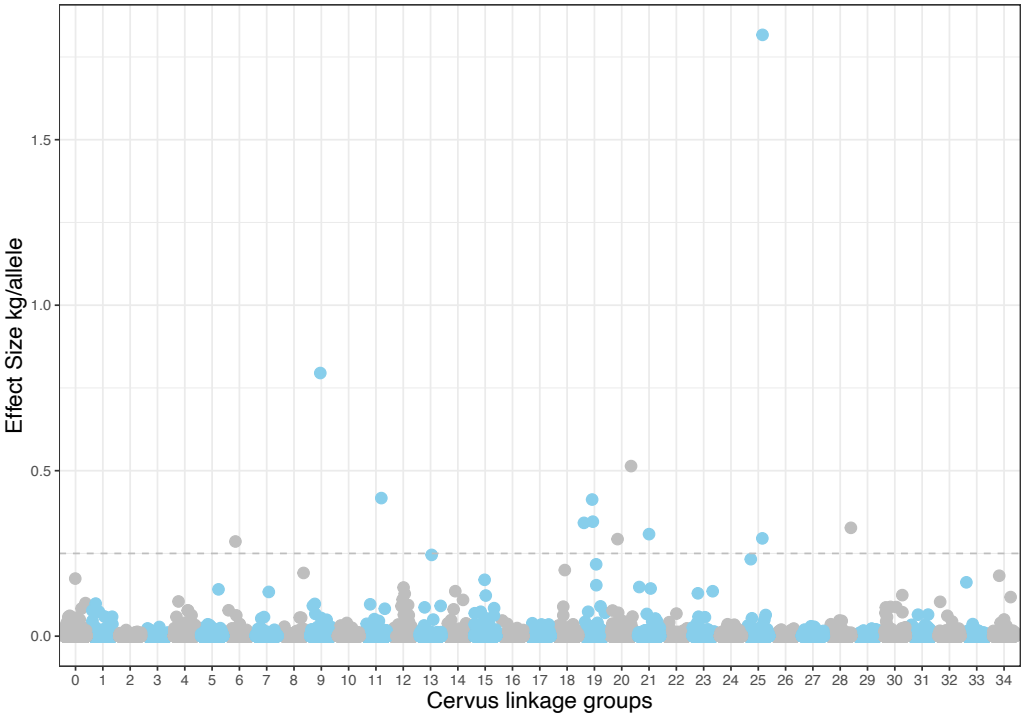
Some example data:



386 deer with phenotypes, ~45,000 SNPs

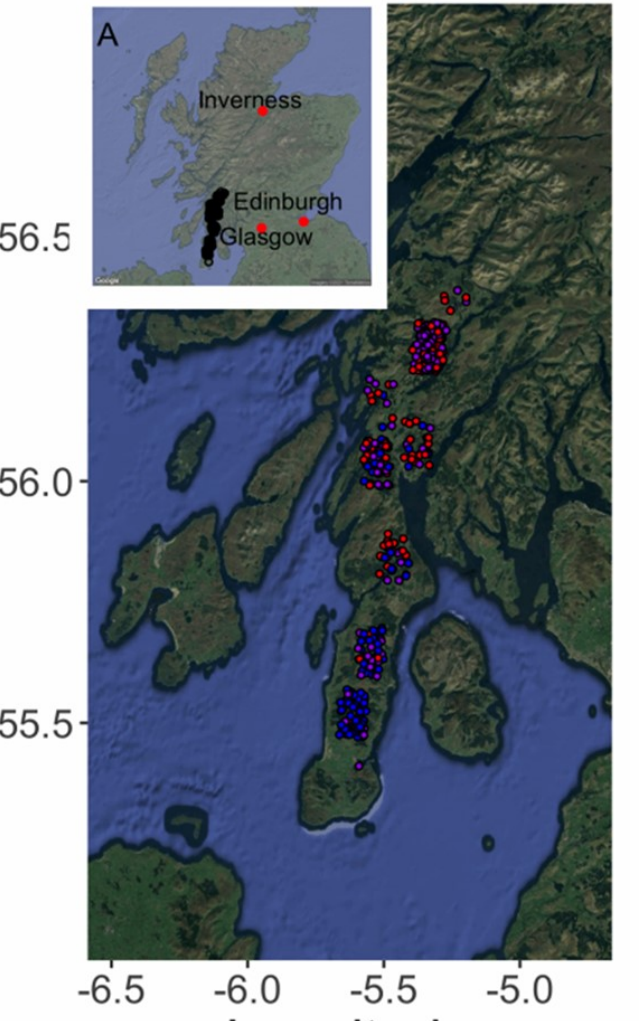
Some example data:

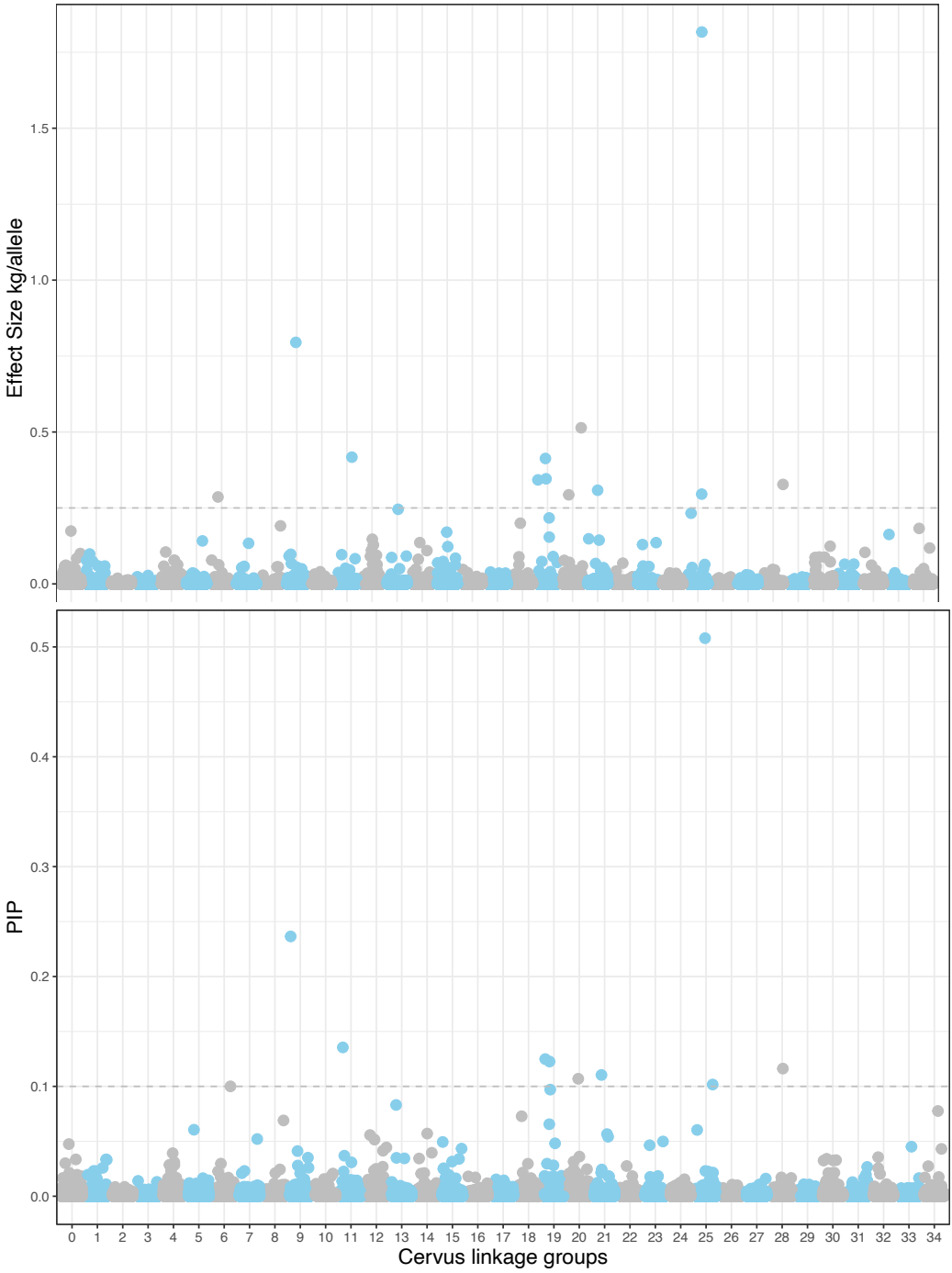




386 deer with phenotypes, ~45,000 SNPs

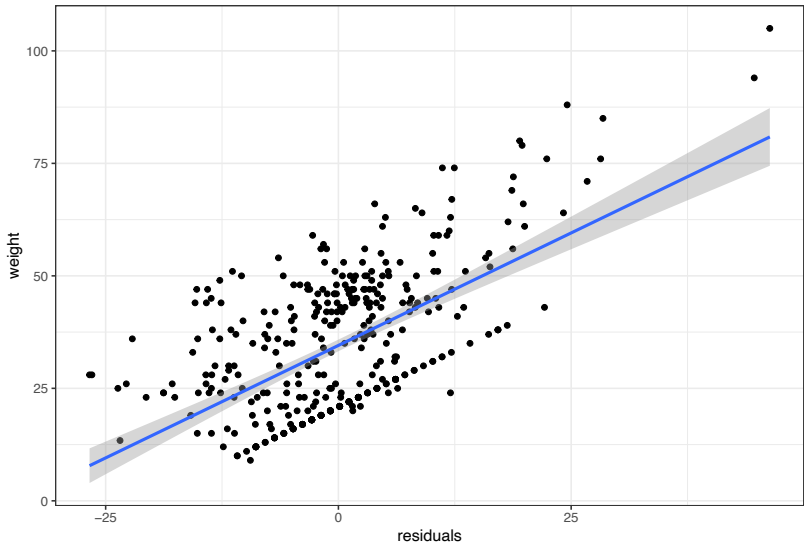
Some example data:

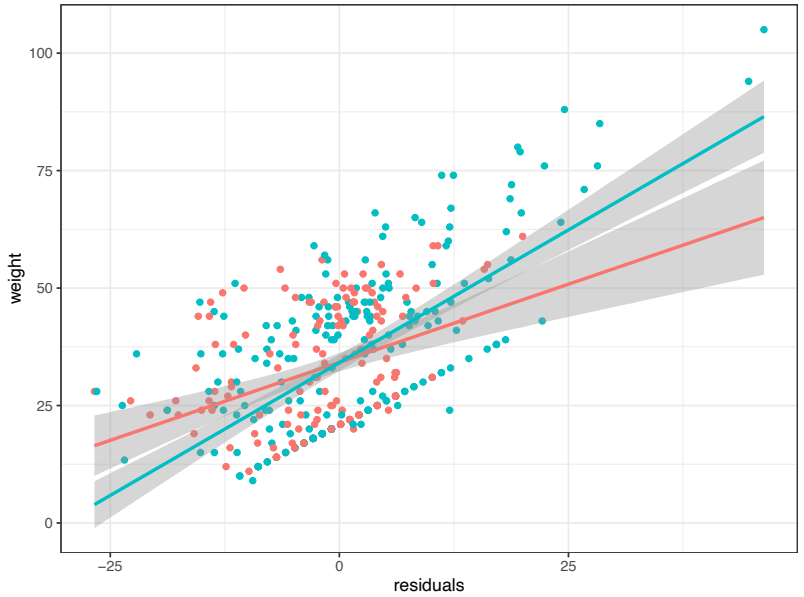




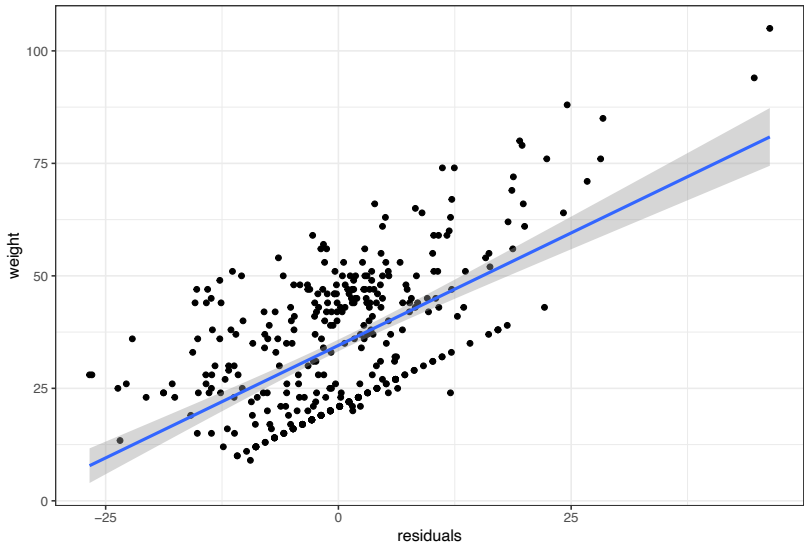
386 deer with phenotypes, ~45,000 SNPs

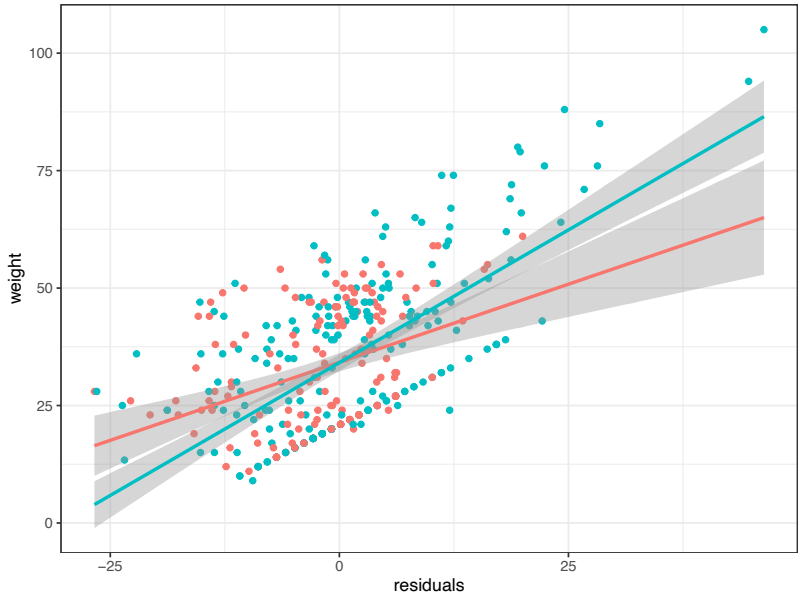
Carcass mass ~sex + age +SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10

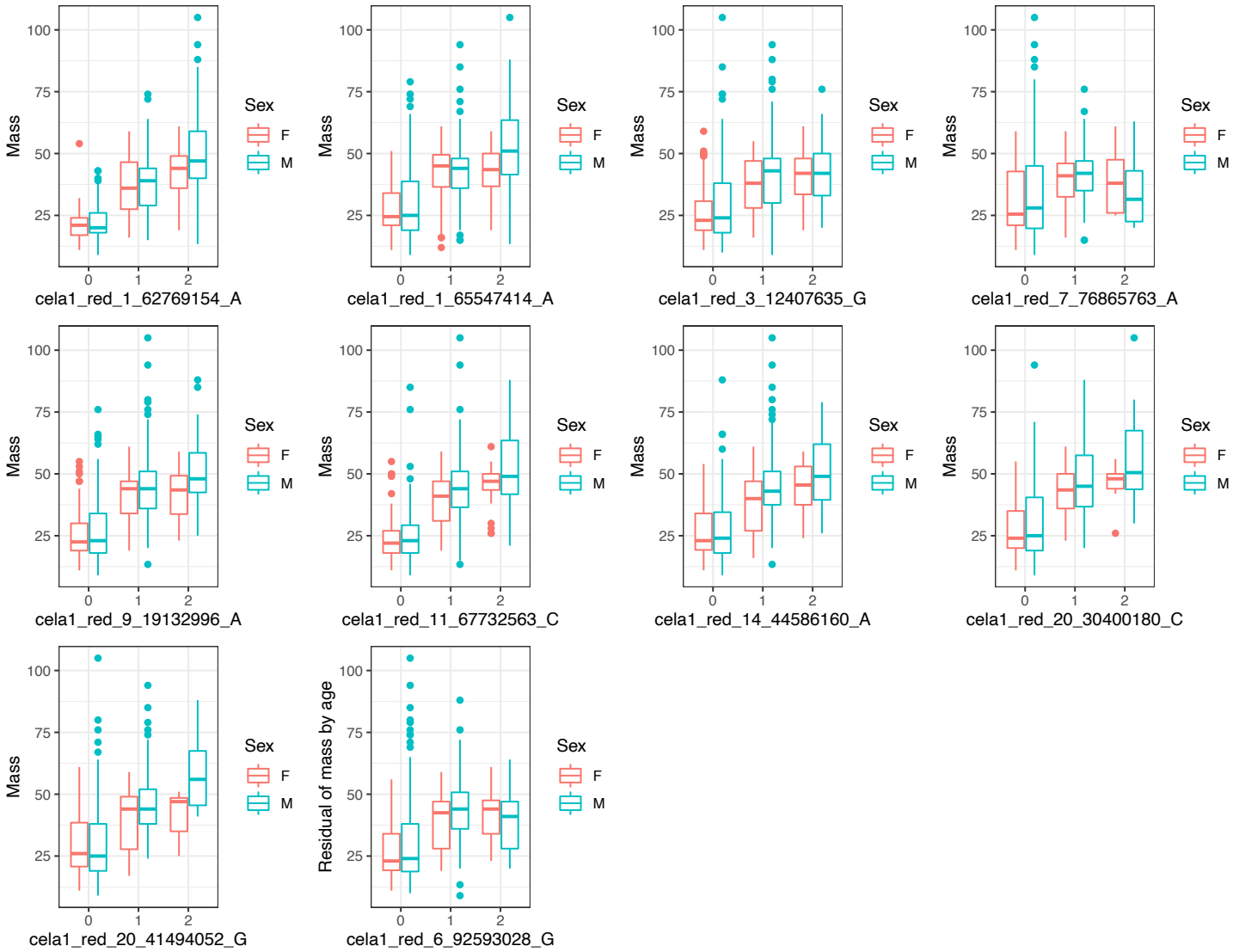




Carcass mass ~sex + age +SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10







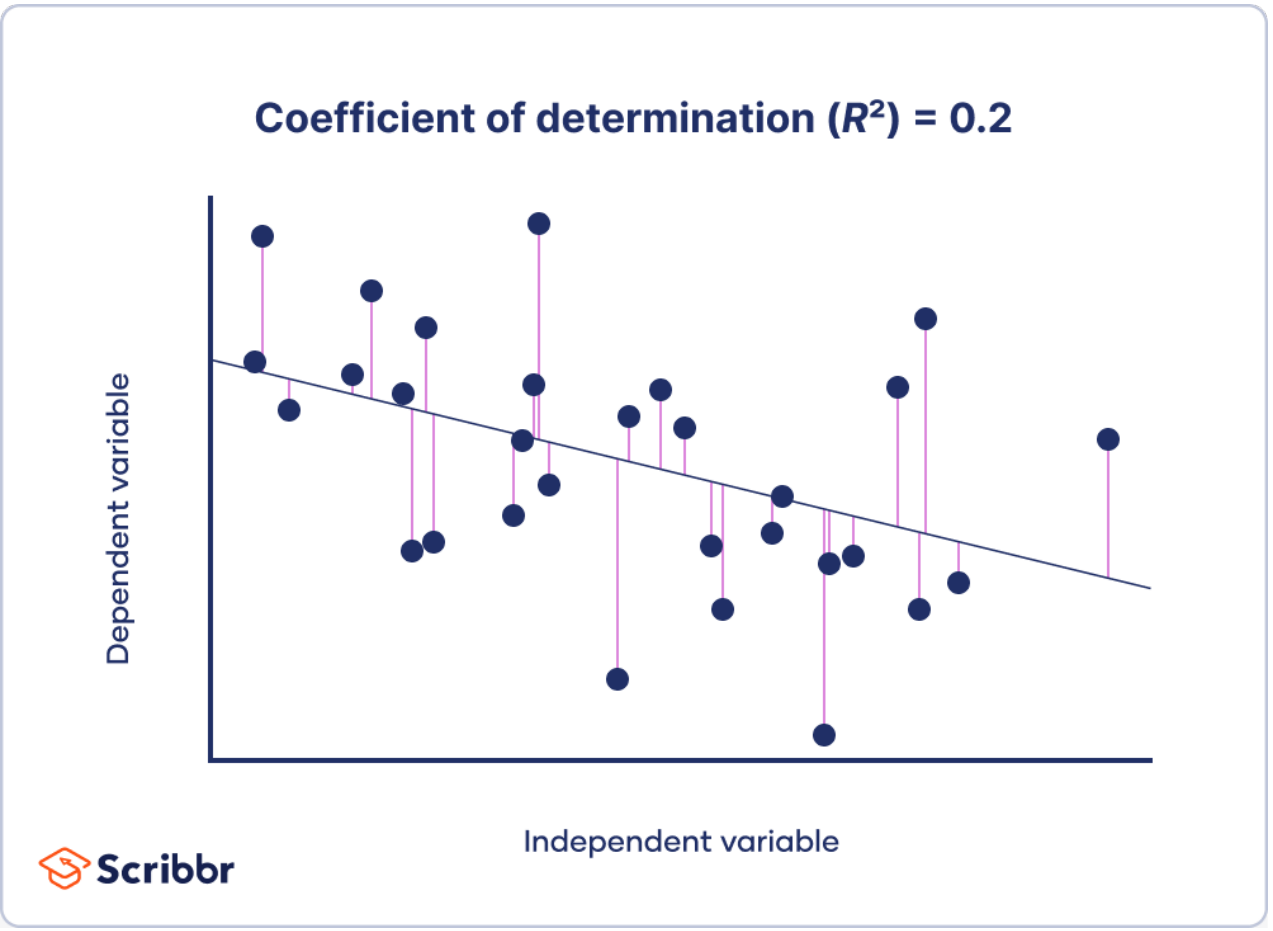
What’s missing in the deer model?

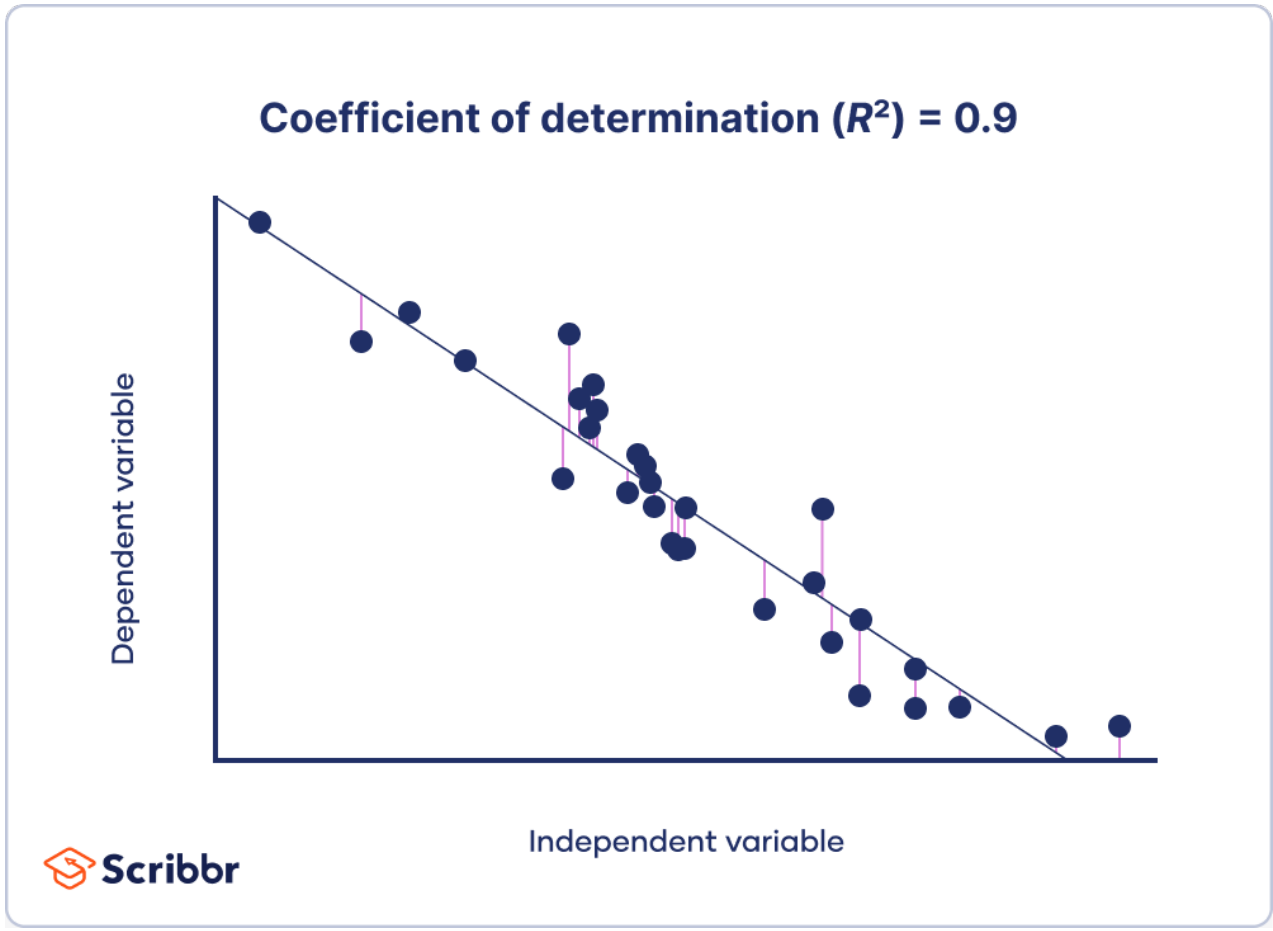
- Carcass mass ~sex + age +SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10

What’s missing in the deer model?

- Carcass mass ~sex + age +SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10

+ error?

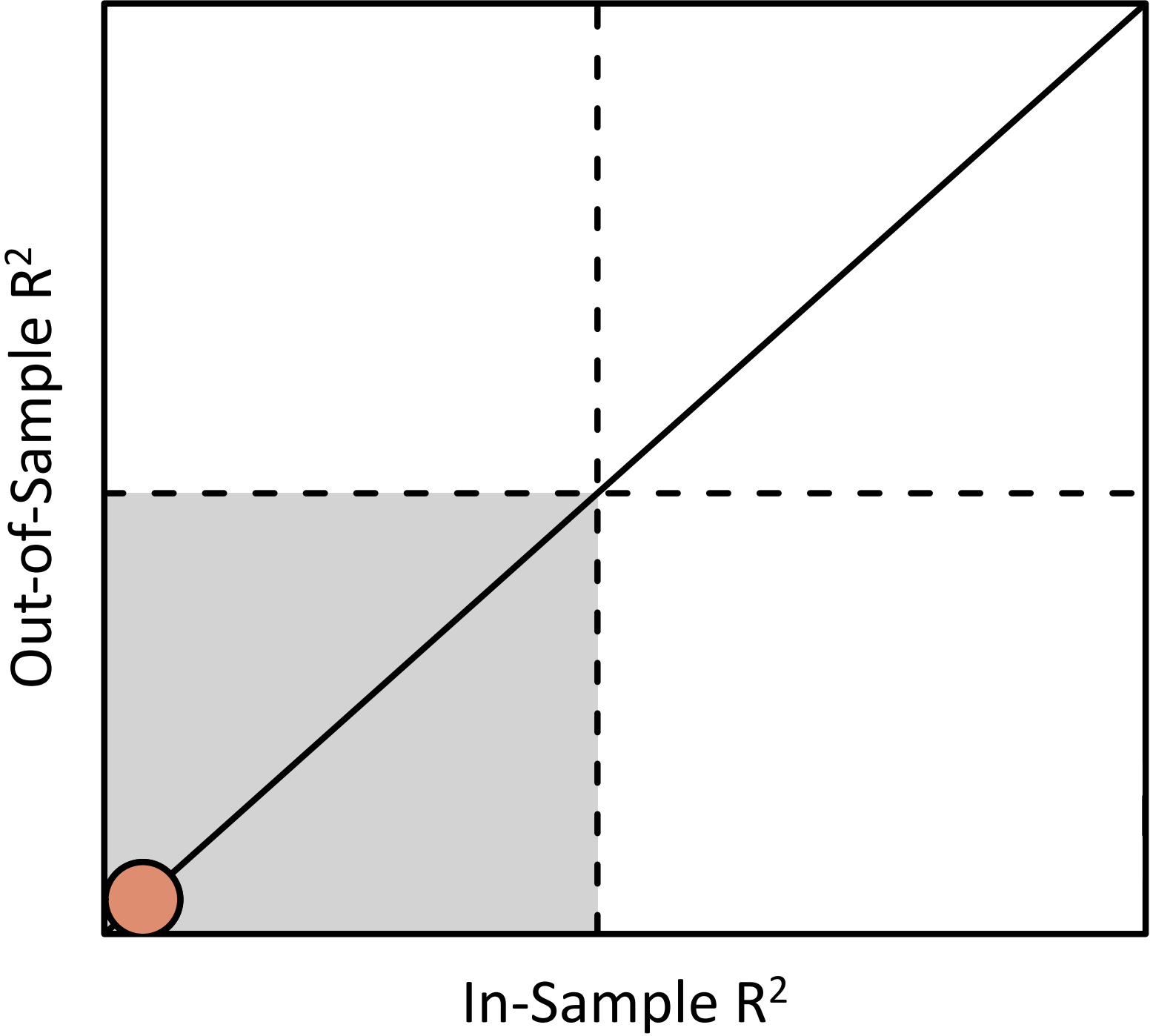


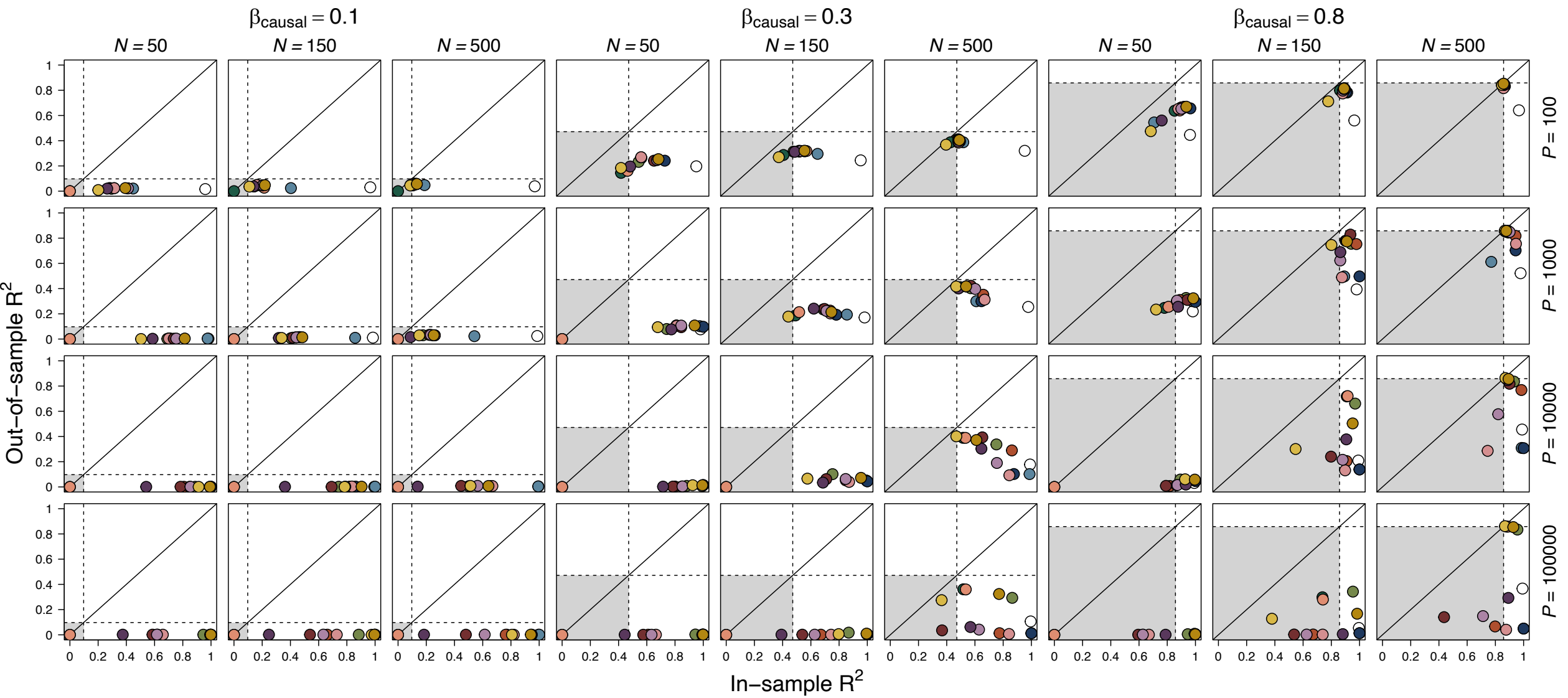


https://www.scribbr.com/statistics/coefficient-of-determination/

- Inference?
- In sample prediction?
- Out of sample prediction?

- Inference – which parameters are associated with a response
- In sample prediction – accurate predictions about the sampled pop
- Out of sample prediction – generalizations about pops for which we have no prior information


Out-of-Sample R^2
In-Sample R^2


βcausal = 0.1
βcausal = 0.3
βcausal = 0.8
N = 50
N = 150
N = 500
P = 100
P = 1000
P = 10000
P = 100000
Out−of−sample R2
In−sample R2
0
0.2
0.4
0.6
0.8
1

Some sparse modelling techniques

- Ridge Regression
- LASSO
- Elasticnet
- Susie
- BSLMM
- Random Forest

Some sparse modelling techniques

- Ridge Regression
- LASSO
- Elasticnet
- Susie (used in genomics, happy to chat)
- BSLMM (used in genomics, happy to chat)
- Random Forest

Some sparse modelling techniques

- Ridge Regression
- LASSO
- Elasticnet
- Susie (used in genomics, happy to chat)
- BSLMM (used in genomics, happy to chat)
- Random Forest (?) won’t talk about again today

Some sparse modelling techniques

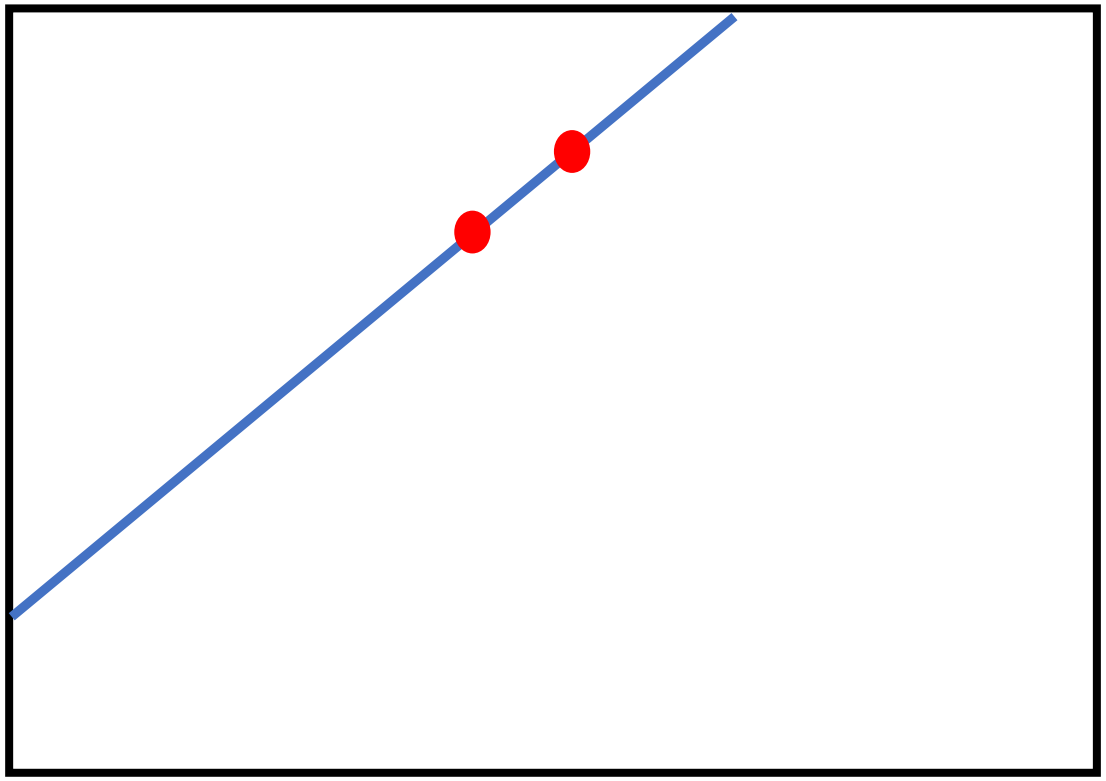
- Ridge Regression
- LASSO
- Elasticnet
- Susie (used in genomics, happy to chat)
- BSLMM (used in genomics, happy to chat)
- Random Forest (?) won’t talk about again today

-> a lot of these are considered ‘machine learning'

Ridge Regression

- Constraint on the coefficients, using a penalty term λ
  - When λ equals 0, ridge regression is basically a linear regression

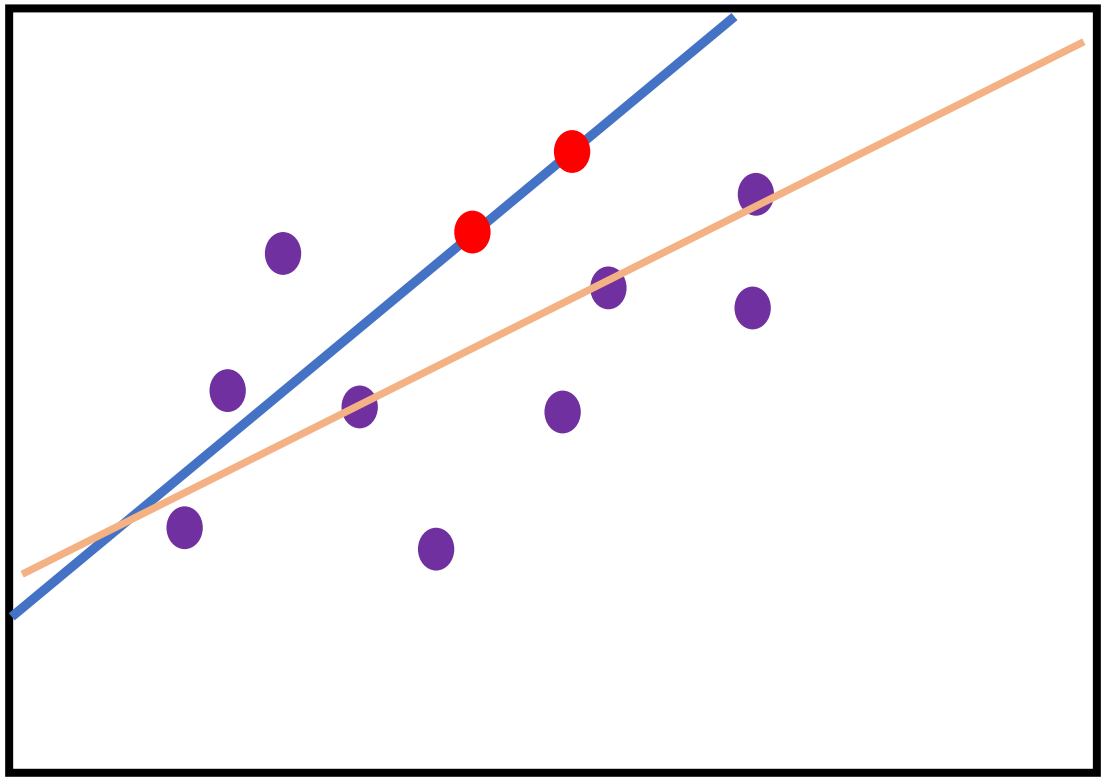
Ridge Regression



Sum of squares for training data is small

Awesome for sparse modelling because: while least squares needs at least n=p to estimate sum of squares, ridge regression can find a solution using cross validation

Ridge Regression



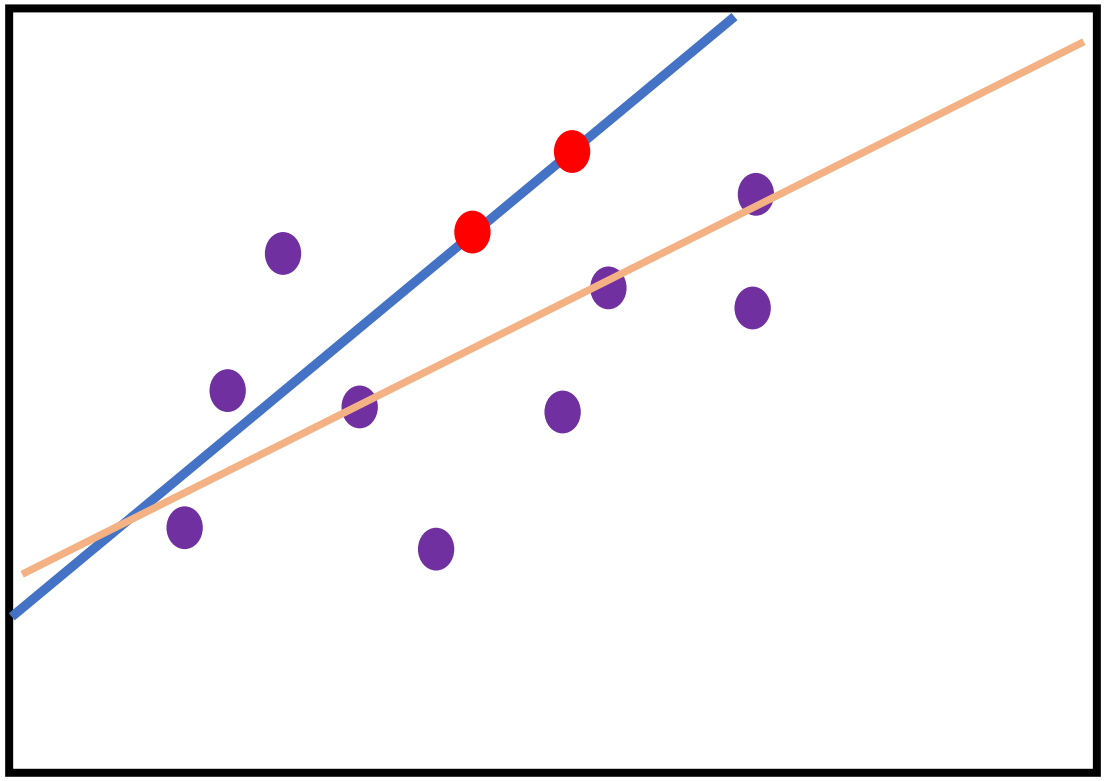
Sum of squares for
training data is small
Sum of squares for
testing data is large

Ridge regression purposely doesn't fit the training data as well, so that it fits the testing data better

This is known as 'shrinkage' or 'regularization'

Awesome for sparse modelling because: while least squares needs at least n=p to estimate sum of squares, ridge regression can find a solution using cross validation

Ridge Regression



Sum of squares for
training data is small
Sum of squares for
testing data is large

Ridge regression purposely doesn't fit the training data as well, so that it fits the testing data better

This is known as 'shrinkage' or 'regularization'

Ridge minimizes sum of squares + $\lambda * slope^2$

This makes predictions of y less sensitive to changes in x

Choose λ using cross validation

Awesome for sparse modelling because: while least squares needs at least n=p to estimate sum of squares, ridge regression can find a solution using cross validation

Lasso and Ridge are really similar

- Ridge: sum of squared residuals + $\lambda * slope^2$
- Lasso: sum of squared residuals + $\lambda * |slope|$

Lasso and Ridge are really similar

- Ridge: sum of squared residuals + $\lambda * slope^2$
- Lasso: sum of squared residuals + $\lambda * |slope|$
- Both make predictions of y less sensitive to x from a small training dataset
- Ridge can only shrink the slope close to 0, while LASSO can shrink it all the way to 0 – makes it easier to get rid of useless variables
- ElasticNet uses both Ridge and Lasso depending on the variable.

Today:

- We’re going to simulate some data, and run some sparse models.
- I’m going to give you the code
- 3 levels
  - Get the code to work for each of the methods (as given to you)
  - Implement the code for a new data set (code not given)
  - Understand inference, in sample and out of sample for this new data set