Regression and Classification Applied to Precision Agriculture

Conference on Applied Statistics in Agriculture and Natural Resources

UtahStateUniversity

May 16-19, 2022

Regression and Classification Applied to Precision Agriculture

Contents:

Conference on Applied Statistics
in Agriculture and Natural Resources
May 16-19, 2022

UtahStateUniversity

Regression and Classification Applied to Precision Agriculture

Conference on Applied Statistics in Agriculture and Natural Resources

UtahStateUniversity

May 16-19, 2022

Guilherme J. M. Rosa (Gee-Ler-Mee)
Department of Animal and Dairy Sciences
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Lab website: www.gjmrosa.org

2

Wisdom of Crowds

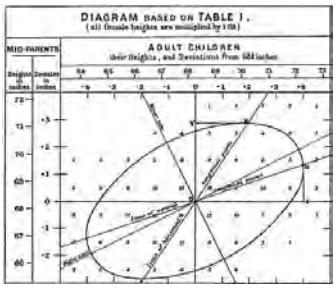
- Francis Galton (1822-1911)
- Ox weight guessing context
- Wisdom of Crowds
  Democratic principle:
  "One Vote One Value" (Vox Populi)

3

Francis Galton

(England, 1822-1911)

- Behavioral genetics, "Nature vs Nurture"
- Weather map, Isochrone map, Anticyclone
- Regression toward the mean, Standard deviation, Galton board, Galton distribution (log-normal), Galton–Watson process, Galton's problem (autocorrelation)



4

Wisdom of Crowds

- In 1906 Galton attended a farmers' fair in Plymouth where he was intrigued by an ox weight guessing contest. Around 800 people entered the contest and wrote their guesses on tickets. The person who guessed closest to the butchered weight of the ox won a prize.
- After the contest Galton took the tickets and ran a statistical analysis on them. He discovered that the average guess of all the entrants was remarkably close (under by only 1 lb !) to the actual weight of the butchered ox (1,198 lbs).
- The collective guess was not only better than the actual winner of the contest but also better than guesses made by cattle experts.

5

Results

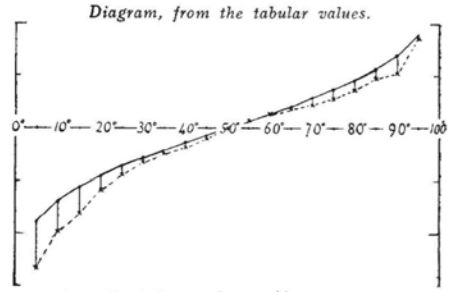
Galton, F. (1907) Vox Populi. Nature 75: 450-451.

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e. =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q_1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
m 50	1207	0	0	0
55	1214	+ 7	− 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q_3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	−10
85	1254	+ 47	+57	−10
90	1267	+ 52	+70	−18
95	1293	+ 86	+90	− 4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Diagram, from the tabular values.



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

Actual weight: 1,198 lbs
Guesses average: 1,197 lbs
Guesses median: 1,207 lbs

6

Remarks

- Democratic principle: "one vote one value"
- *Vox Populi* : middlemost estimate
- Model averaging
- Ensemble (Boosting) methods, combination of weak predictors
- Very useful in regression and classification
- Resulting combined model is better than any of the models alone

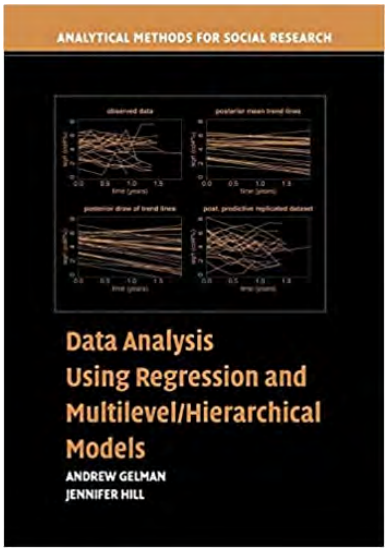
7

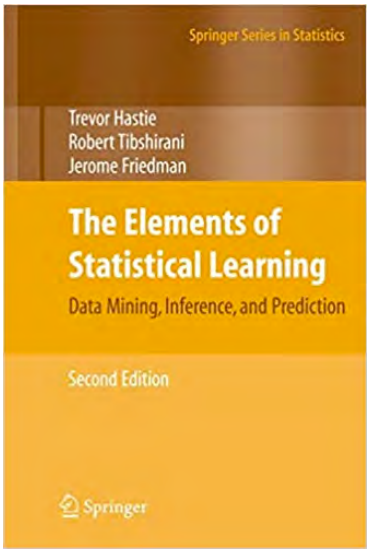
Regression/Classification Topics

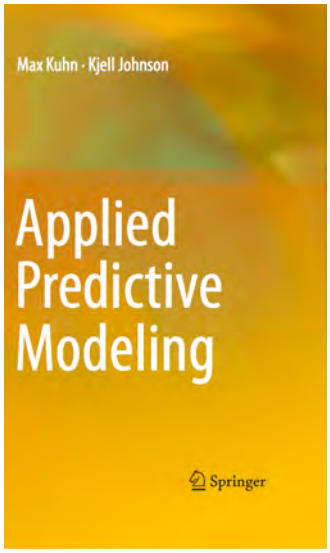
- Least squares and beyond
- Checking model assumptions
- Non-Gaussian models
- Heteroscedasticity
- Variable transformation
- Model (variable) selection
- Linear and non-linear models
- Multi-collinearity
- Dimension reduction techniques
- Shrinkage estimation
- Parametric and non-parametric
- Measurement error
- Measurement error
- Missing data imputation
- Multivariate models
- Mixed effects (multilevel)
- Power and sample size calculation
- Bayesian methods
- Monte Carlo methods
- Prediction, interpretation, causality
- Robust regression
- Kernel regression
- Machine learning approaches
- Software

etc., etc., etc.

8







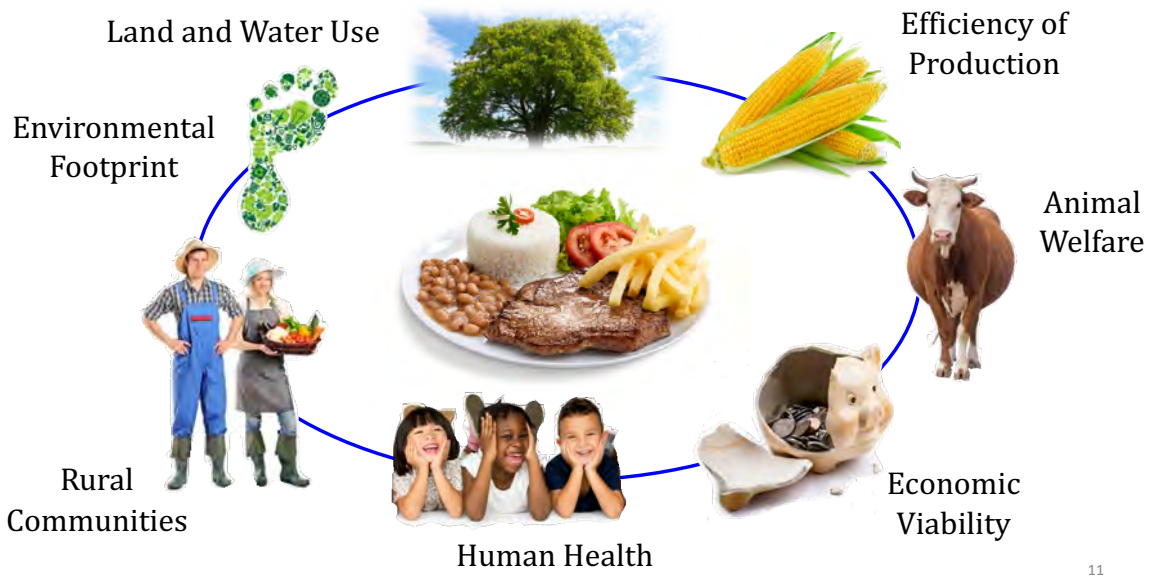
9

Outline

- Overall Introduction; Regression and Classification, Digital Agriculture
- Multiple Regression
- Multilevel and Hierarchical Models
- Regularization Approaches
- Machine Learning
- Kernel Regression
- Causality

10

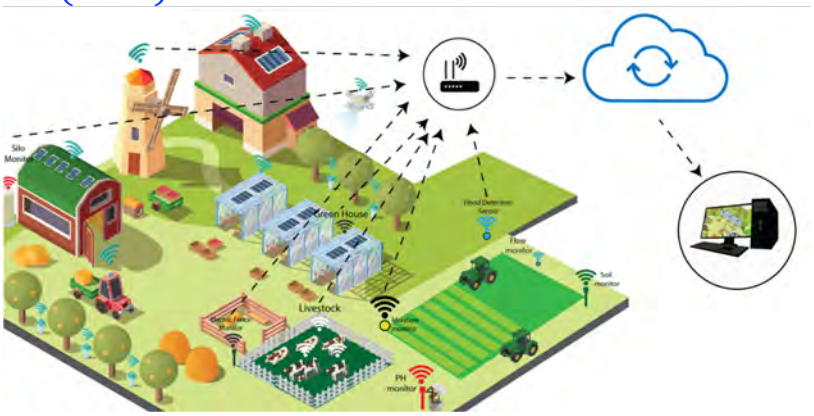
Sustainability on Food Production



11

Digital Agriculture

- Sensors
- Communication Networks
- Unmanned Aerial Vehicles (UAVs)
- Robotic Machinery
- Data Analytics
- Data Visualization
- Artificial Intelligence
- Other Technologies



12

Sensor Technology

- Automated data recording systems
- Robotics and artificial intelligence
- Real time measurements; sensors
  - Image
  - Motion
  - Sound
  - Chemical composition
  - Spectroscopy
  - etc.

13

Data Collection

- Spatial and temporal dimensions
- Multilevel: animal or plant, pen or plot, farm, geographical region
- Historical data and data streaming (real time)
- Myriad of data formats (structured and unstructured)

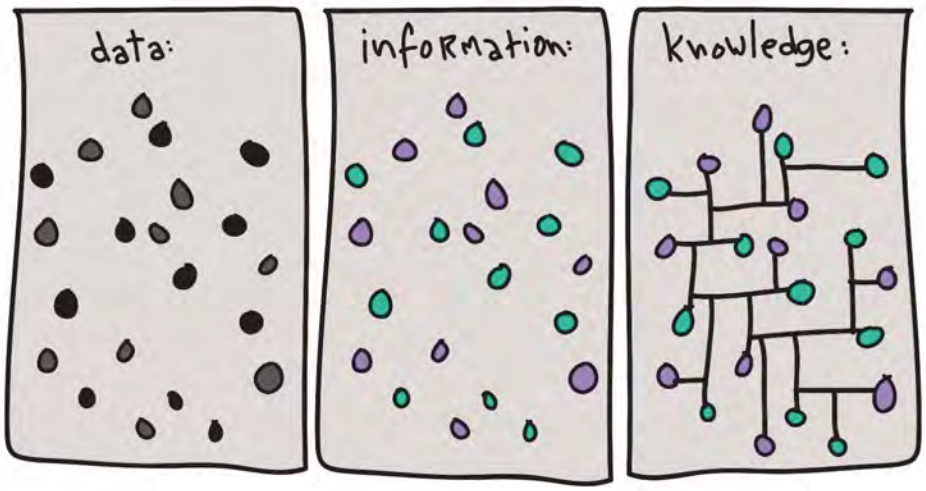
14

Data Integration and Data Processing

- Database strategy and architecture (unstructured and structured data; temporal scale, etc.)
- Centralized or distributed, local or cloud storage (security, privacy)

16

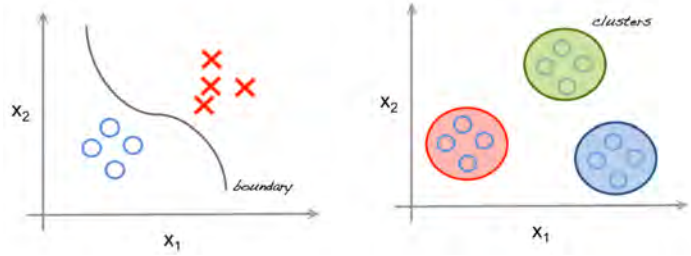
Data is the Fuel for AI



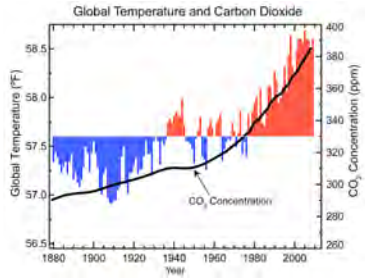
17

Data Analytics Tools and Goals

- Supervised and Unsupervised



- Prediction, Interpretation and Causal Inference



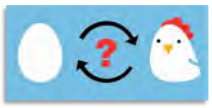
18

From Data to Decisions

Data

Descriptive

Predictive

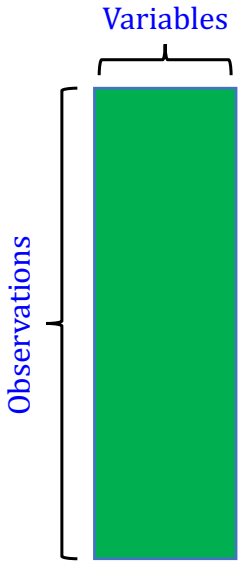


Prescriptive

Optimization

19

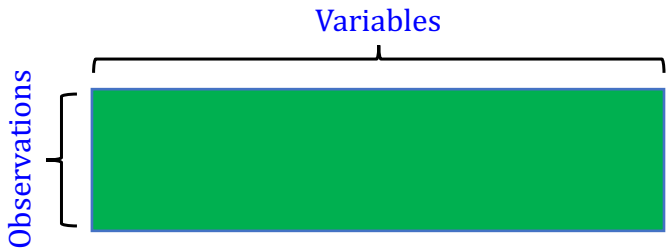
Tall Data



- High power
  (statistical vs. practical significance)
- Asymptotic properties
- Plenty of d.f. ☺

20

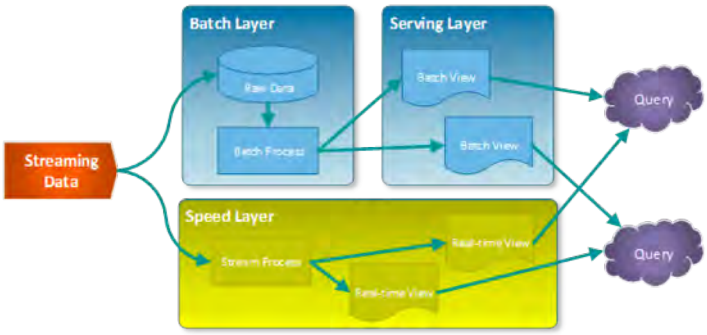
Wide Data



- "big p, small n" paradigm
- Collinearity
- Multiple testing
- Penalized/regularized regression
- Dimension reduction techniques

21

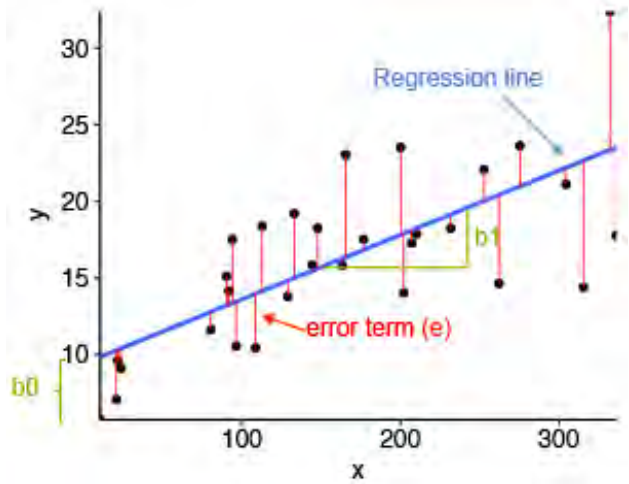
Data Streaming and Batch Processing



- Real time monitoring:
  Animal- and Farm (or pen)-level
- Management optimization; Genetic improvement
  Product quality, production efficiency, animal wellbeing, etc.

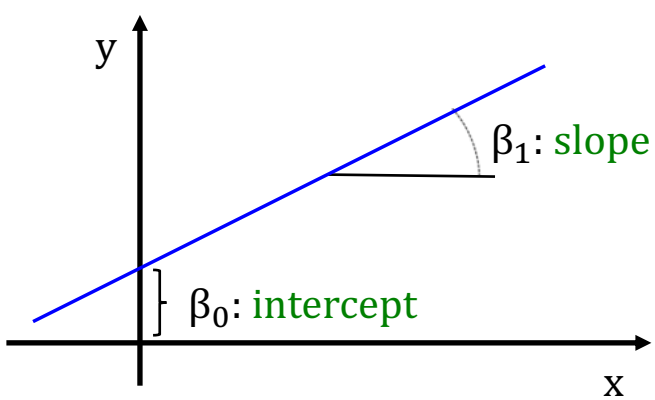
22

Intro to Regression and Classification



23

Simple Linear Regression



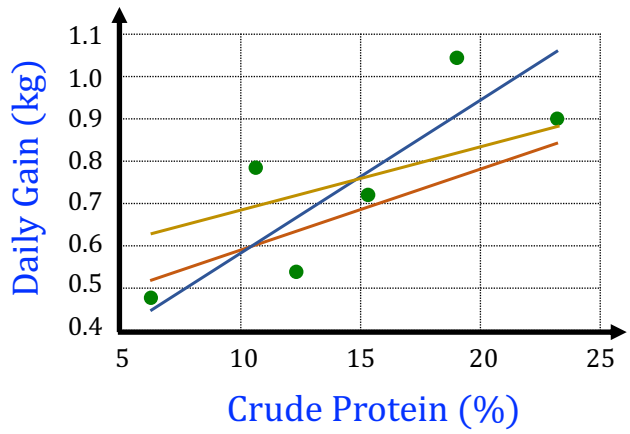
$$E[y] = \beta_0 + \beta_1 x$$

- If $x = 0$, then $y = \beta_0$ (regression intercept)
- Each additional unit in x is associated with β_1 units of change in y
- Note: regression parameters (β_0 and β_1) should be interpreted only within the range of x values in the dataset.

24

Example: Forage crude protein (% of dry matter) and beef cattle average daily weight gain (kg)

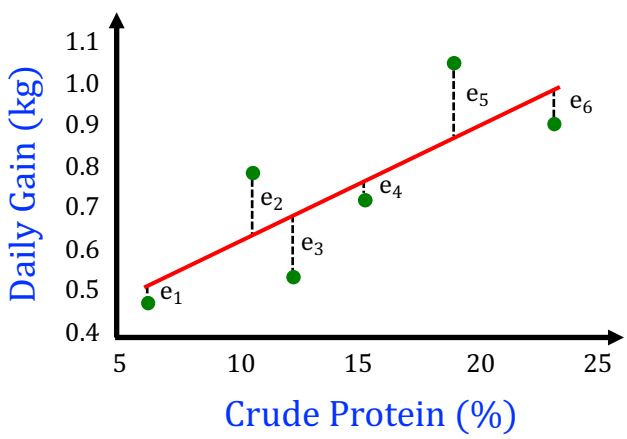
CP (%)	DG (kg)
6.3	0.48
10.7	0.79
12.4	0.55
15.4	0.72
19.1	1.03
23.3	0.89



- How should we choose the line (which values of $\hat{\beta}_0$ and $\hat{\beta}_1$) that best describes the data?

25

Least Squares



Errors: $e_i = y_i - (\beta_0 + \beta_1 x_i)$

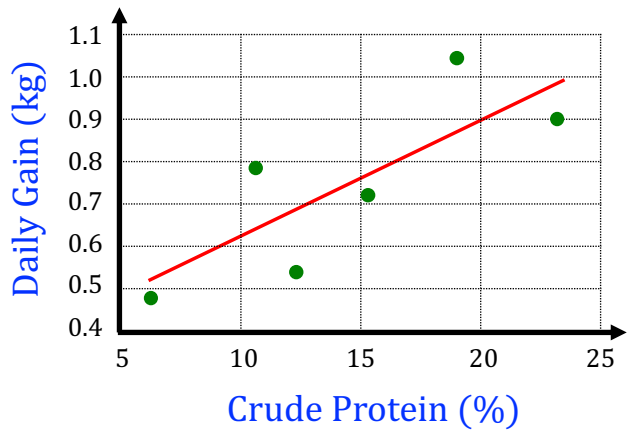
Sum of Squares: $\sum_{i=1}^{n} e_i^2$

- An alternative is to minimize the sum of the squares of the errors.

26

Fitted Regression

CP (%)	DG (kg)
6.3	0.48
10.7	0.79
12.4	0.55
15.4	0.72
19.1	1.03
23.3	0.89



- Estimated regression: $DG = 0.3534 + 0.0268 \text{ x } CP$
- What is the interpretation of the regression coefficient (slope)?

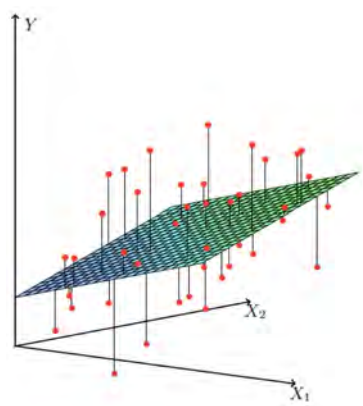
27

Multiple Linear Regression

- Response variable described as a linear function of multiple predictors:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

- Multiple linear regression includes also models with interaction between predictor variables, and polynomial regression:



Interaction: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) + e$

Polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + e$

28

Example: Carcass Quality of Cull Dairy Cows

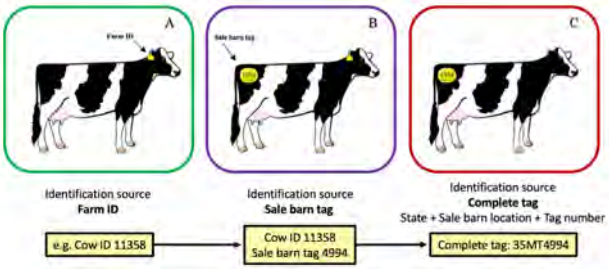
- Investigating the relationship between life history factors, live animal auction price, and carcass quality of cull dairy cows.

Dairy Farms | Sales Barn | Meat packing plant

Moreira, L. C., Passafaro, T. L., Schaefer, D. M. and Rosa, G. J. M. (2021) The effect of life history events on carcass merit and price of cull dairy cows. *Journal of Animal Science* 99(1): skaa401.

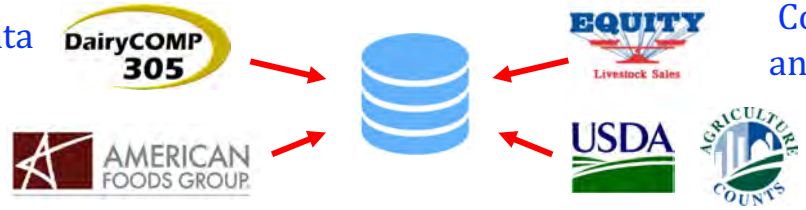
29

Data Integration



Farm data

Carcass data



Cow price and weight

Average national price

30

Available Variables

Variables	Description	Unit	Levels
DIM	Days in milk	d	—
DOPN	Days open	d	—
DDRY	Days dry	d	—
CINT	Calving interval	d	—
TOTM	Total milk production for current lactation	kg	—
TOTP	Total protein production for current lactation	kg	—
TOTF	Total fat production for current lactation	kg	—
PTOTM	Total milk production for previous lactation	kg	—
PTOTP	Total protein production for previous lactation	kg	—
PTOTF	Total fat production for previous lactation	kg	—
305M	Dairy Comp internal projected 305-d milk production	kg	—
305ME	Dairy Comp projected 305ME	kg	—
ME305	305ME	kg	—
PDIM	Days in milk for previous lactation	d	—
PDOPN	Days open for previous lactation	d	—
FCM1	Fat corrected milk for lactation 1	kg	—

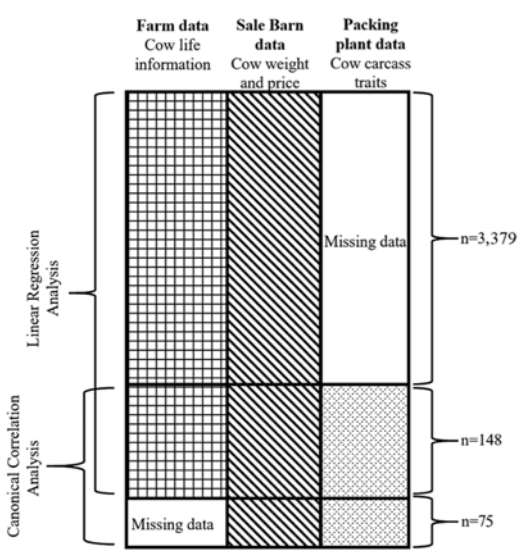
Variables	Description	Unit	Levels
305ME1	305 d mature herd equivalent for lactation 1	kg	—
FCM2	Fat corrected milk for lactation 2	kg	—
305ME2	305 d mature herd equivalent for lactation 2	kg	—
LACT	Lactation number	—	7
EASE	Calving ease score	—	2
MAST	Lifetime Mastitis	—	3
Culling	Culling reason	—	7
Month	Month of sale	—	12
Year	Year of sale	—	4
Farm	Farm of origin	—	4
Weight	Live weight	kg	—
BP	Price paid per 100 lb (45.34 kg) of live weight	$	—
price ratio	BP divided by the national average cow price ($/cwt) for the respective month and year of sale	$	—
Carcasswt	Carcass weight	kg	—
Dressing	Dressing percentage	%	—
Grade	Paid grade for carcass	—	9
Maturity	Animal maturity	—	2
Trim	Score of carcass trimming loss weight	—	2

Farm management software

Sales barn, USDA, Meatpacking plant

31

Canonical Correlation



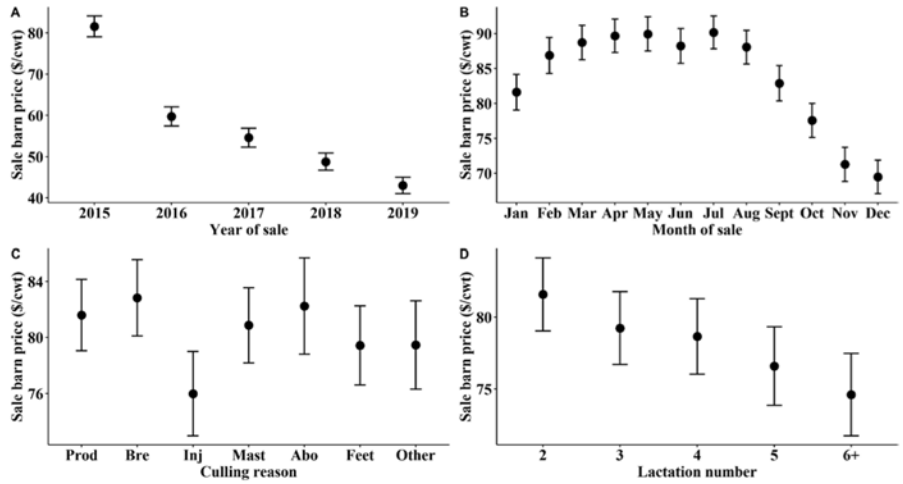
Overall structure of the dataset, including information from farm, sale barn and packing plant.

$$\mathrm{Var}\begin{bmatrix}\theta\\ \eta\end{bmatrix} = \begin{bmatrix}V_{11} V_{12}\\ V_{21} V_{22}\end{bmatrix}$$

$$\rho = \max_{a,b}\left\{\frac{a^{T}V_{12}b}{\sqrt{a^{T}V_{11}ab^{T}V_{22}b}}\right\}$$

32

Results



Least squares means and confidence intervals of sale BP for (A) year of sale, (B) month of sale, (C) culling reason (Prod: low production, Bre: breeding problem, Inj: injury, Mast: Mastitis and udder problem, Abo: abort, Feet: Leg and feet problem, Other: Other reasons), and (D) lactation number.

33

Results

Table 4. Pearson correlation coefficients between sale BP, price ratio, and carcass characteristics variables[1]

	Price ratio	Weight	Carcasswt	Dressing	Maturity	Grade	Trim
BP	0.91***	0.377***	0.536***	0.445***	−0.313***	0.608***	−0.258***
Price ratio		0.407***	0.571***	0.472***	−0.397***	0.662***	−0.275***
Weight			0.876***	0.068	0.224***	0.186**	−0.121*
Carcasswt				0.534***	0.069	0.465***	−0.254***
Dressing					−0.277***	0.661***	−0.306***
Maturity						−0.453***	0.141*
Grade							−0.24***

Table 5. Canonical coefficients, loadings, correlation, and aggregate redundancy coefficient for the first canonical variates of the price ratio and carcass merit characteristics of 223 cull dairy cows[1]

Variables[2]	Loadings	Cross-loadings	Correlation	ARC
η				
Price ratio			0.761	0.216
θ				
Livewt	0.393	0.299		0.579
Carcasswt	0.680	0.517		
Dressing	0.662	0.503		
Grade	0.851	0.647		
Maturity	0.580	0.441		
Trim	−0.355	−0.270		

34

Results

Predictive performance for cull cow price

Barn Price
(nominal value)

$r^2 = 0.75$, RSME = \$7.6/cwt,
and MAE = \$5.8/cwt

$$\begin{aligned} BP = \mu &+ \text{lactation} + \text{culling} + \text{farm} + \text{month} + \text{year} + b_1 \\ &\times \text{DDRY} + b_2 \times \text{TOTM} + b_3 \times \text{PTOTP} \\ &+ b_4 \times \text{305M} + b_5 \times \text{FCM1} + b_6 \times \text{FCM2} \\ &+ b_7 \times \text{305ME1} + b_8 \times \text{305M2} + b_9 \times \text{weight} + e \end{aligned}$$

Adjusted Price
(corrected for seasonality)

$r^2 = 0.47$, RSME = 0.1045,
and MAE = 0.076

$$\begin{aligned} \text{Price ratio} = \mu &+ \text{lactation} + \text{culling} + \text{farm} \\ &+ \text{month} + \text{year} + b_1 \times \text{DOPN} \\ &+ b_2 \times \text{ME305} + b_3 \times \text{weight} + e \end{aligned}$$

35

Variable Transformation

- Centering and scaling: $y^* = \frac{y-\bar{y}}{s_y}$ and $x^* = \frac{x-\bar{x}}{s_x}$

  $y^* = \beta_1^* x^* + \varepsilon^* \longrightarrow \hat{\beta}_1^* = \text{Corr}(x, y)$

- Polynomial: $x^* = x^\lambda$, where $\lambda = 2, 3, \ldots$

  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

- Log: $y = \exp\{\beta_0 + \beta_1 x + \varepsilon\} = B_0 \times B_1^x \times \epsilon$,

  where $B_0 = \exp\{\beta_0\}$, $B_1 = \exp\{\beta_1\}$, and $\epsilon = \exp\{\varepsilon\}$

  $\log(y) = \beta_0 + \beta_1 x + \varepsilon$, with $y > 0$

36

Variable Transformation (cont'ed)

- Log-log: $y = \beta_0 \times x^{\beta_1} \times \varepsilon$

  $\log(y) = \log(\beta_0) + \beta_1 \log(x) + \log(\varepsilon)$, with $y > 0$ and $x > 0$

  $y^* = \beta_0^* + \beta_1 x^* + \varepsilon^*$

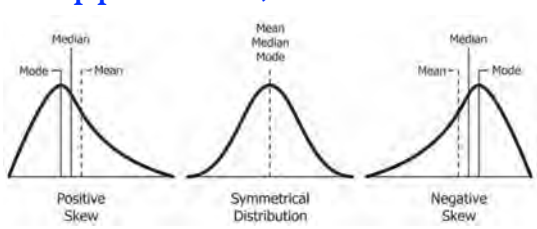
- Others: square-root, inverse, etc.

- Box-Cox $y = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$

37

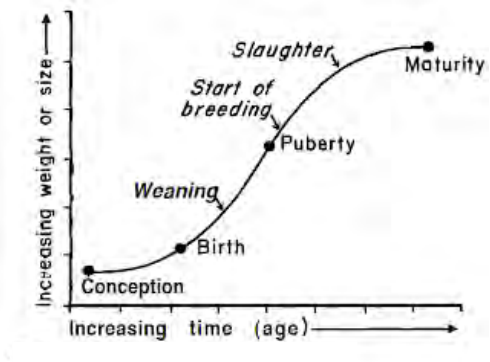
Non-Normal Data

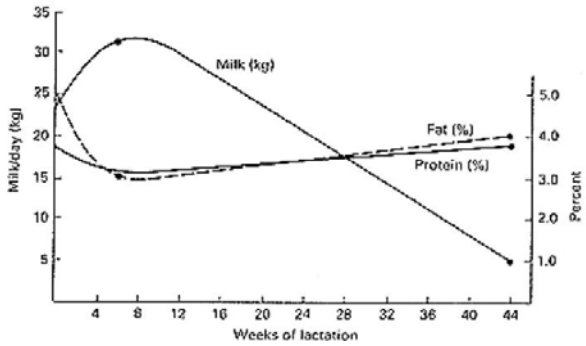
- Least squares coupled with resampling techniques (e.g. bootstrap and permutation)
- Data transformation (e.g. Box-Cox)
- Generalized linear model (exponential family)
- More general models (e.g. mixtures) using Bayesian MCMC
- Nonparametric approaches, Machine Learning



38

Non-Linear Models





- Growth curves: Brody, Gompertz, and Von Bertalanffy models
- Lactation curves: Wood and Wilmink models

39

Example: Lactation Modeling of Milk Protein Profile

- The protein profile of milk includes several caseins, whey proteins, and nonprotein nitrogen compounds, all important for human nutrition and cheesemaking properties

- Objective was to model the pattern of each N compound expressed qualitatively (as % of total milk N), quantitatively (in g/L milk), and as daily yield (in g/d)

Amalfitano, N., Rosa, G. J. M., Cecchinato, A. and Bittante, G. (2021) Nonlinear modeling to describe the pattern of 15 milk protein and nonprotein compounds over lactation in dairy cows. *Journal of Dairy Science* 104: 10950-10969.

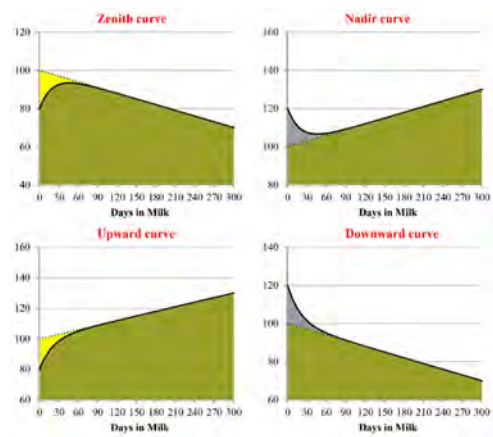
40

Wilmink's Model

- Four-parameter model (Wilmink 1987):

$$y_t = a + b \times \exp(-k \times t) + c \times t + e$$

where y_t is the milk production in time t, and e is the error term. The four parameters represent the persistency coefficient (parameter c) that explains the variation in the long-term milk component (parameter a), the short-term milk component (parameter b), and the speed of adaptation (parameter k).



Example of shapes of lactation curves.

41

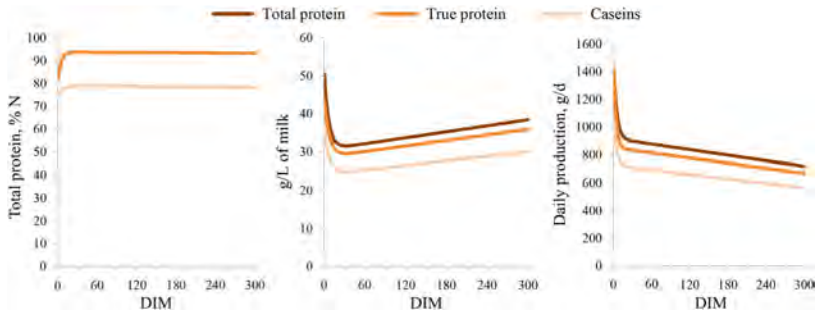
M&M and Results

- Data on detailed milk nitrogenous compound profiles (15 fractions for each expression mode: 45 traits) obtained from 1,342 cows
- Data from each milk trait analyzed with the NLMIXED procedure of SAS using the final model:

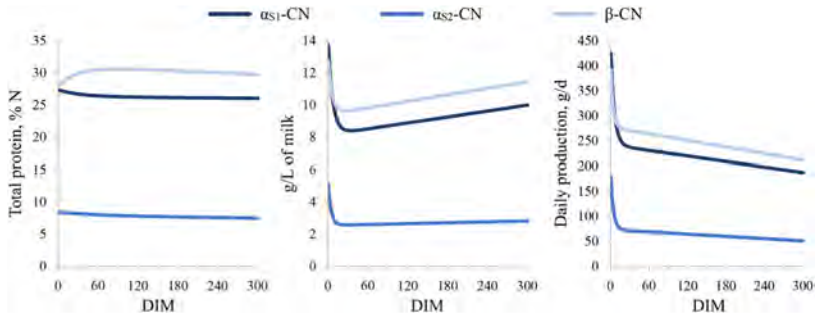
$$y_{ijkl} = a + b \times \exp(-k \times t) + c \times t + breed_i + parityj + herd\_datek + e_{ijkl}$$

with fixed effect of the breed of the cow (4 classes: Holstein-Friesian, Brown Swiss, Simmental, local breeds) and parity (3 classes: 1, 2, and ≥3), and random effect of herd-date (n = 41)

42



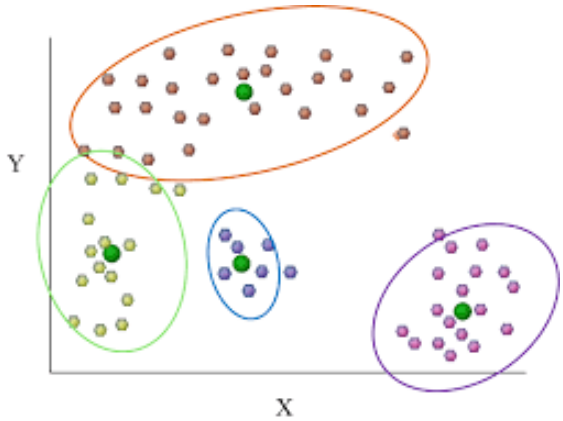
Pattern of total protein, true protein, and total CN content during lactation expressed in percentage of total protein (% N), grams per liter of milk (g/L), and daily production (g/d).



Pattern of α_{S1}-CN, α_{S2}-CN, and β-CN content during lactation expressed in percentage of total protein (% N), grams per liter ofmilk (g/L), and daily production (g/d).

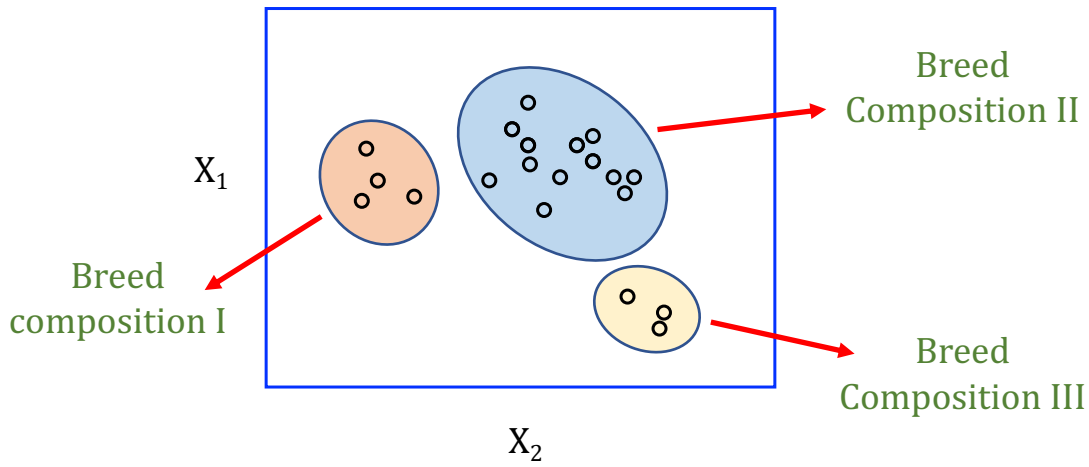
43

Classification and Clustering



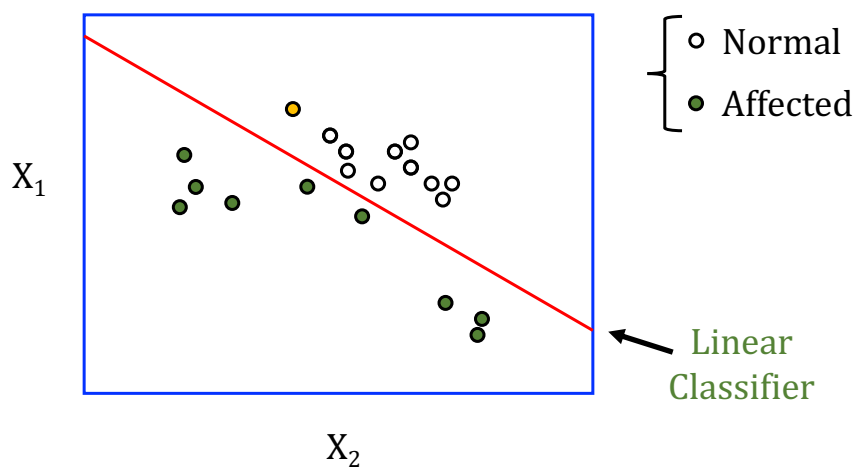
44

Clustering vs. Classification



45

Clustering vs. Classification



46

Cluster Analysis

- Cluster analysis (or clustering) is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar (in some sense) to each other than to those in other groups (clusters)
- It is an unsupervised exploratory data mining technique used in many fields, including pattern recognition, image analysis, etc.
- Many algorithms available, such as K-means, mixture models, hierarchical clustering

47

K-means algorithm

1) Define the number K of clusters
2) Randomly selected the K centroids, which are used as the beginning points for every cluster
3) Performs iterative (repetitive) calculations to optimize the positions of the centroids
4) The algorithm halts when the centroids have stabilized, or a pre-defined number of iterations has been achieved

48

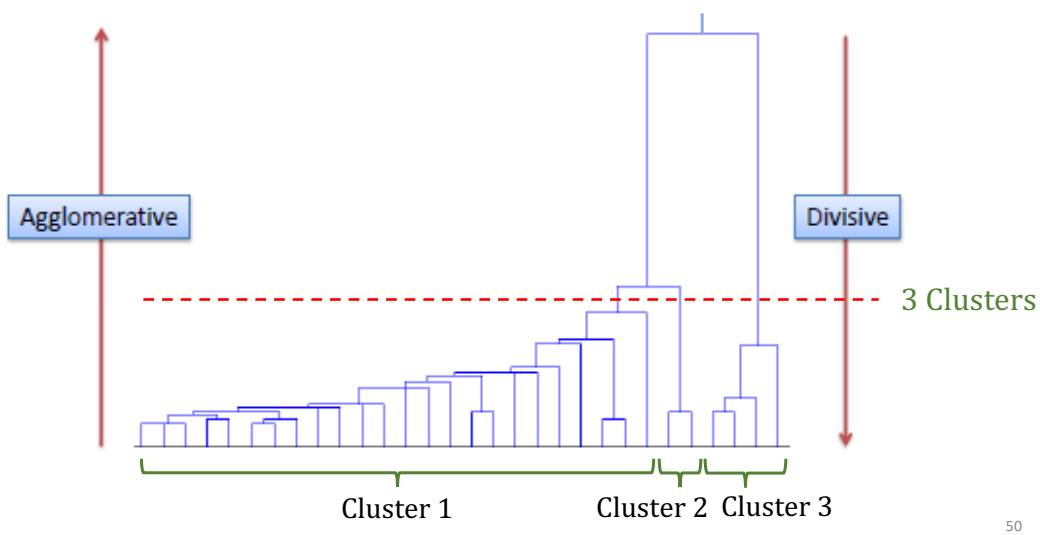
Hierarchical Clustering

- Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. There are two types of hierarchical clustering: Divisive and Agglomerative.
- Divisive or top-down method: all of the observations are initially assigned to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation.
- Agglomerative or bottom-up method: each observation is assigned to its own cluster. Based on a similarity measure (e.g., distance) the two most similar clusters are merged. The process is repeated until there is only a single cluster left.

49

Hierarchical Clustering

Dendrogram



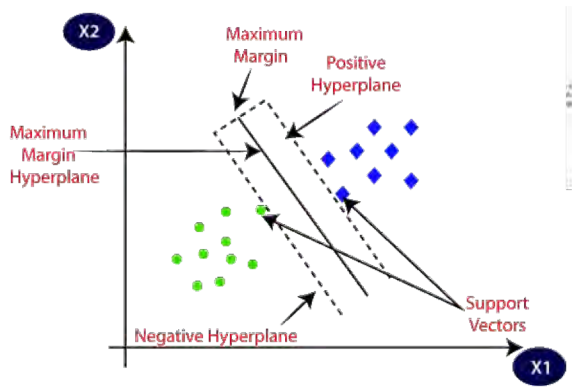
50

Classification

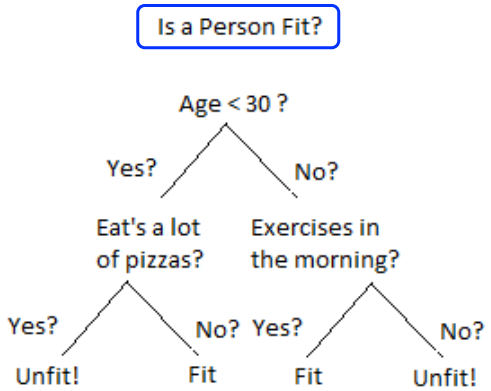
- Discriminant or Classification techniques seek to categorize samples into groups based on the predictor characteristics
- Examples are: assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.).
- Classification is a supervised approach of pattern recognition.
- Linear models: logistic regression, linear discriminant analysis. Non-Linear models: Neural networks, support vector machines, K-nearest neighbors, Naïve Bayes, Classification trees, etc.

51

Classification



Support Vector Machine



Decision Trees

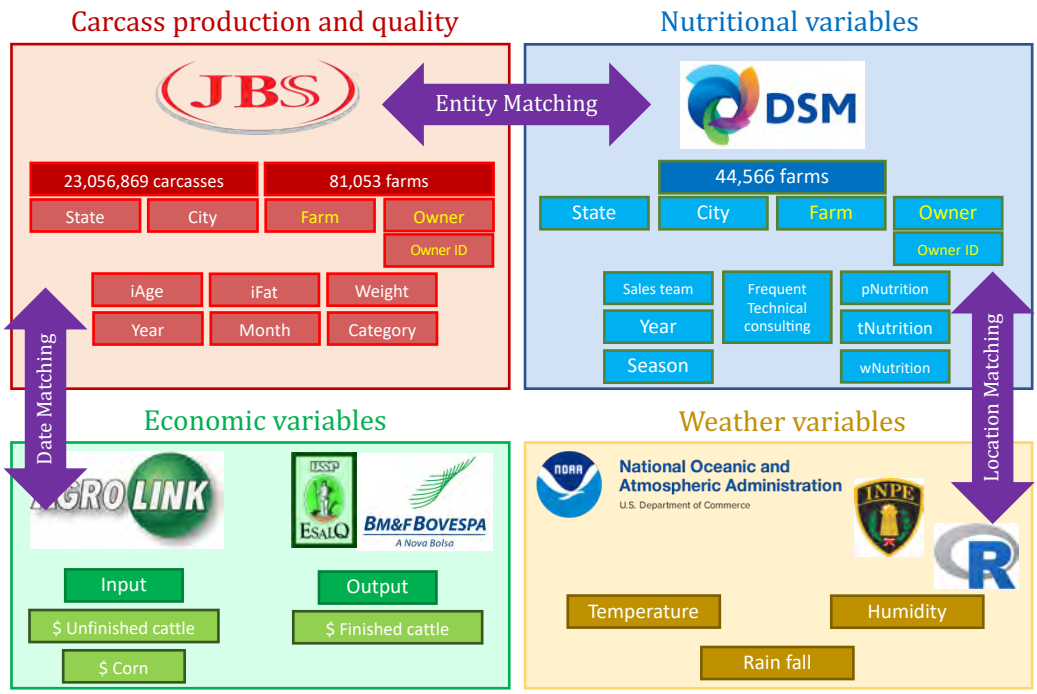
52

Example: Investigating Factors that Affect Beef Production and Quality in Brazil

Objectives

1. Forecast beef cattle production and quality, using a large scale data set integrated from different sectors of industry in Brazil
2. Compare prediction quality of alternative methods: Generalized Linear Model, Random Forest, and Neural Network

53

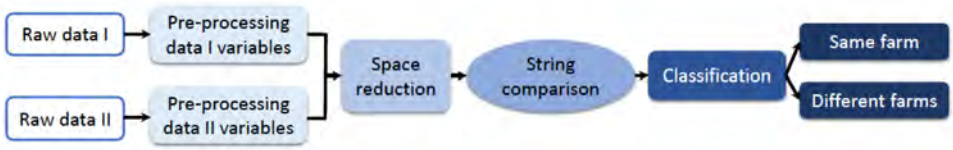


54

Data Integration

Dataset 1	Dataset 2
Saint Johns	Glenview
Green view	St. John's
Laredo	_platte_
Platte	Larredo

Analysis Pipeline



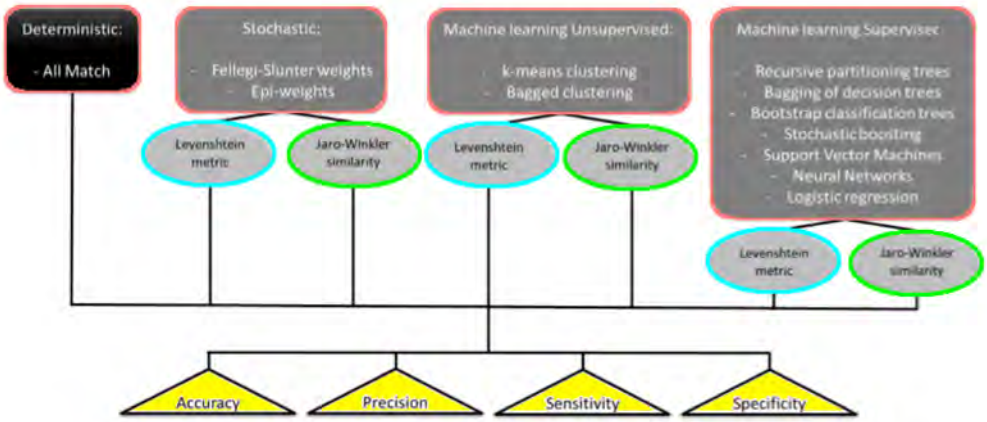
81K farms JBS x 44.5K farms DSM ~ 3.6G comparisons

55

Data Integration

Comparison of Approaches for Farm Data Linkage

(Entity Matching)



56

Data Integration

- Best classification methods:

  Support Vector Machine (SVM)
  (acc = 99.9%, prec = 91.1%, sens = 97.3%, spec = 99.9%)

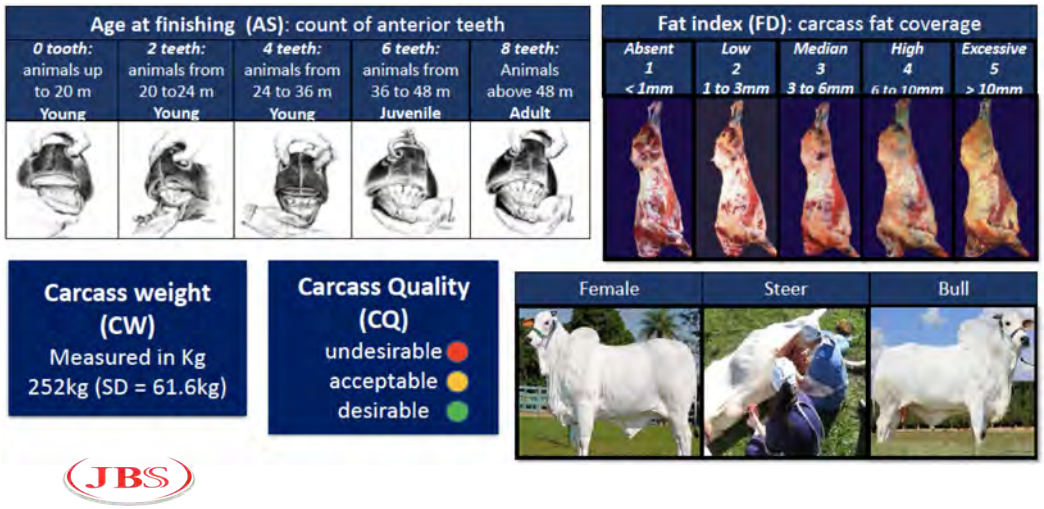
  Bagged Clustering (BC)
  (acc = 99.9%, prec = 90.8%, sens = 93.2%, spec = 99.9%)

- Results indicate that both SVM and BC are suitable for farm matching in scenarios where training labels are available, or not, respectively.

Aiken VCF, Dorea JRR, Acedo JS, Dias F and Rosa GJM (2019) Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods. *Computers and Electronics in Agriculture* 163: 104857.

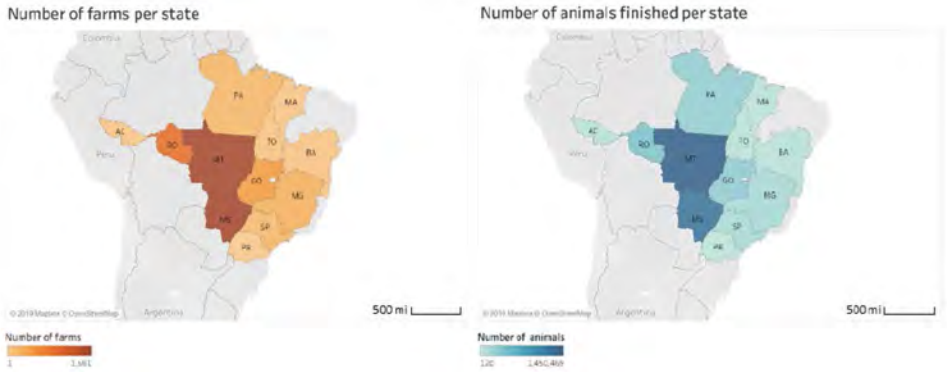
57

Response Variables



58

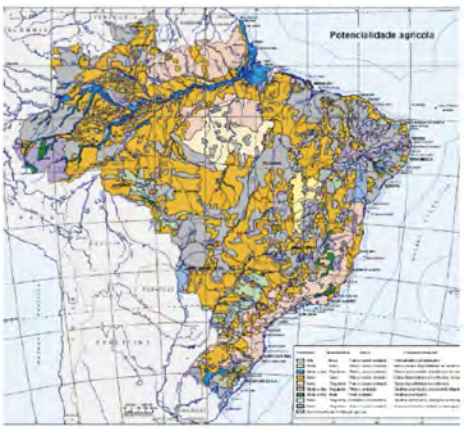
Distribution of Farms and Market Cattle



Distribution of farms and of finished animals in the data set per state in Brazil

59

Soil and Climate Distribution



IBGE
Instituto Brasileiro de Geografia e Estatística

60

Methods

Models: Linear Regression (LR)
Generalized Linear Regression (GLR)
Random Forest (RF)
Multilayer Perceptron Neural Networks (NN)

Predictors: Animal Category (female, steer, bull), Technician Consulting, Nutrition Product, Corn Price, Sales price, Soil, Climate, Month, and Age at Slather (only for CW and FD)

Predictive ability: 10-fold Cross-Validation; training with 542,935 (2014/2015) and testing with 285,357 observations (2016)

Continuous: Root Mean Square Error (RMSEp), Coefficient of Determination (R^2), and Mean Absolute Error (MAE)

Categorical: Accuracy and the Cohen's kappa coefficient (Kappa)

Software: R package "caret" (Kuhn, 2019)
Center for High Throughput Computing (CHTC)

61

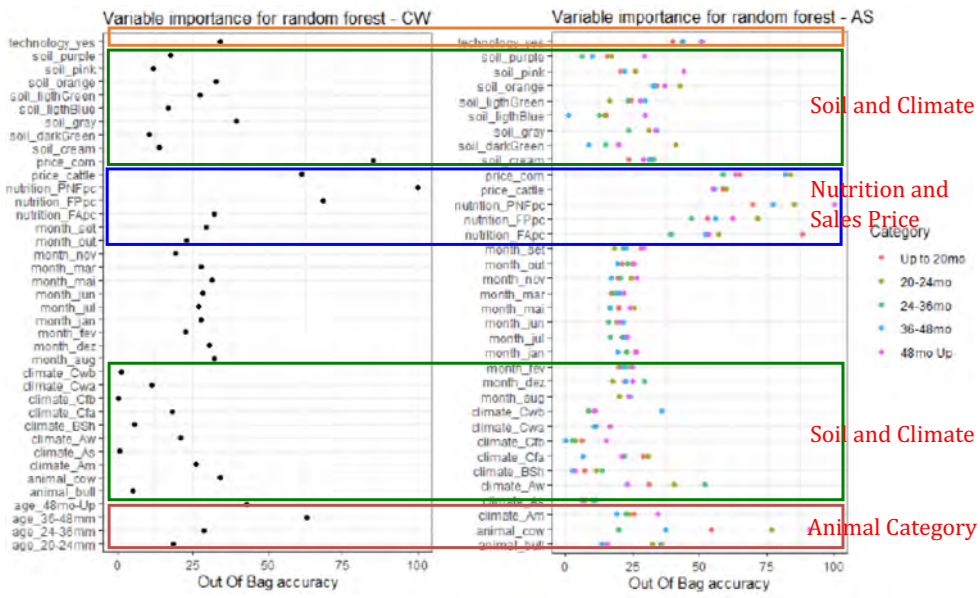
Models Predictive Ability

		Outcome variable		
		Categorical		
Model	**Measure**	**AS**	**FD**	**CQ**
Generalized linear regression	Accuracy	0.2867 (±0.0011)	**0.4576 (±0.0022)**	**0.5867 (±0.0019)**
	Kappa	0.0666 (±0.0015)	**0.0476 (±0.0037)**	**0.0862 (±0.0037)**
Random Forest	Accuracy	**0.2871 (±0.0019)**	0.4494 (±0.0020)	0.5390 (±0.0016)
	Kappa	**0.0759 (±0.0026)**	0.0523 (±0.0032)	0.0930 (±0.0032)
Multilayer perceptron neural networks	Accuracy	0.2536 (±0.0028)	0.3742 (±0.0019)	0.4640 (±0.1999)
	Kappa	0.0237 (±0.0034)	0.0501 (±0.0160)	0.0670 (±0.0017)

		Continuous	
		CW (centered and scaled)	**CW (original scale)**
Linear regression	RMSEp	0.6765 (±0.0027)	41.2697 kg
	R^2	0.6017 (±0.0017)	0.6017
	MAE	0.5097 (±0.0017)	31.0941 kg
Random Forest	RMSEp	**0.6626 (±0.0025)**	**40.4217 kg**
	R^2	**0.5920 (±0.0024)**	**0.5920**
	MAE	**0.5018 (±0.0013)**	**30.6122 kg**
Multilayer perceptron neural networks	RMSEp	0.8073 (±0.0030)	49.2491 kg
	R^2	0.4657 (±0.0037)	0.4657
	MAE	0.5905 (±0.0045)	36.0233 kg

62

Importance of Predictor Variables



63

Computational Requirements

For All Response Variables Combined:

Regression: 6 h with 4 CPUs and 40 GB memory

Random Forest: 2,370 h with 109 CPUs and 8 TB memory

Neural Network: 15,482 h with 5,580 CPUs and 223 GB memory

Aiken VCF, Fernandes AFA, Passafaro TL, Acedo JS, Dias F, Dorea JRR and Rosa GJM (2020) Forecasting beef production and quality using large scale integrated data from Brazil. *Journal of Animal Science* 98(4): skaa089.

64

Regression Modeling Goals and Applications

- Prediction: no specific interest on interpretation of regression coefficients (black-box and nonparametric models are useful as well), contribution of each variable on prediction accuracy, explores association (not causal relationship) between target variables and predictors

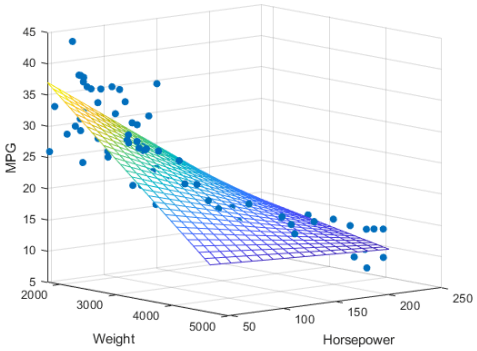
- Interpretation of model parameter estimates: parametric model backed-up by theory related to domain of application, e.g. infinitesimal model in quantitative genetics, non-linear curves (digestibility, fluid dynamics, growth, lactation, etc.)

- Causal inference: hypothesis testing in the context of controlled randomized trials and also observational data (issues of confounding and selection bias)

65

Multiple Regression

- Least-Squares
- Maximum Likelihood
- Logistic Regression
- Generalized Linear Models



66

Multiple Linear Regression

Response variable (Y)	Predictor (explanatory) variables			
	X_1	X_2	$\cdots$	X_p
y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

- Response variable described as a linear function of multiple predictors: $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$

67

Multiple Linear Regression

- Model: $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$
- Predictors (explanatory variables) can be continuous or categorical (regression and ANOVA)
- Error terms (ε_i) assumed independent from each other, with mean 0 and variance σ_ε^2, i.e. $\varepsilon_i \sim^{iid} (0, \sigma_\varepsilon^2)$
- Some additional assumptions related to the distribution of ε_i will be considered later, such as normality

68

Multiple Linear Regression

- General linear model: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

where $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ is the vector of observations on the response variable, $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]^T$ is the vector of location parameters (regression coefficients), $\mathbf{X}$ is a known incidence /design $(n \times k)$ matrix linking each observation y_j to the vector $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ is a vector of error terms, assumed $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$

- Notice: $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$, $k = p + 1$ **and** $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X\beta}$

69

Least Squares

- Seek estimate $\widehat{\boldsymbol{\beta}}$ that minimizes the residual sum of squares (RSS): $\mathrm{RSS} = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$, where $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$
- Matrix notation: $\mathrm{RSS} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$
$$= \mathbf{y}^T\mathbf{y} - 2\widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}}$$
- Partial derivatives: $\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}}$
- Equating to zero: $\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \rightarrow \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (LS estimate)
- Proof of minimum: $\frac{\partial^2 \mathrm{RSS}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2\mathbf{X}^T\mathbf{X}$
(Hessian matrix; positive definite if $\mathrm{rank}(\mathbf{X}) = \mathbf{k}$)

70

Least Squares

- The errors $\boldsymbol{\varepsilon}$ come for a distribution with mean 0 and variance σ_ε^2, which can be estimated from the residuals as:

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Coefficient of determination: fraction of the variation in the response variable that is predictable from the explanatory variable(s):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- Adjusted R^2: $R_{adj}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k)}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2/(n-1)} = 1 - \frac{(n-1)}{(n-k)}\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

71

Testing Regression Coefficients

- Model: $y_i = E[y_i|\mathbf{x}_i] + \varepsilon_i = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} + \varepsilon_i$
- If normality is assumed for the error terms, i.e. $\varepsilon_i \sim^{iid}(0, \sigma_\varepsilon^2)$, then:

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma_\varepsilon^2) \quad \text{and} \quad (n-k)\, s^2 \sim \sigma^2\chi_{(n-k)}^2$$

- For any regression coefficient:

$$H_0: \beta_j = 0 \rightarrow z_j = \frac{\hat{\beta}_j}{s\sqrt{\nu_j}} \sim t_{n-k}$$

where $\nu_j = j^{th}$ diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$

72

"In Least Squares We Trust"

- Unbiased estimator:

$$E[\widehat{\boldsymbol{\beta}}] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E[\mathbf{y}] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

- Variance: $\mathrm{Var}[\widehat{\boldsymbol{\beta}}] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathrm{Var}[\mathbf{y}]\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$

$$= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{I}\sigma_\varepsilon^2\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

$$= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma_\varepsilon^2$$

- Distribution of $\widehat{\boldsymbol{\beta}}$ however depends on the distribution of $\mathbf{y}$
- Inference about $\boldsymbol{\beta}$ can be performed using for example Monte Carlo methods such as Bootstrap

73

Gauss-Markov Theorem

- Linear combination of the parameters: $\boldsymbol{\theta} = \mathbf{k}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]^T$
- LS estimate: $\widehat{\boldsymbol{\theta}} = \mathbf{k}^T\widehat{\boldsymbol{\beta}} = \mathbf{k}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- $E[\widehat{\boldsymbol{\theta}}] = \mathbf{k}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{k}^T\boldsymbol{\beta}$
- Consider any linear combination $\widetilde{\boldsymbol{\theta}} = \mathbf{c}^T\mathbf{y}$ such that $E[\widetilde{\boldsymbol{\theta}}] = \mathbf{k}^T\boldsymbol{\beta}$, i.e. unbiased
- It can be shown that $\mathrm{Var}\left[\mathbf{k}^T\widehat{\boldsymbol{\beta}}\right] \leq \mathrm{Var}[\mathbf{c}^T\mathbf{y}]$
- Mean squared error: $\mathrm{MSE} = E\left[\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right] = \mathrm{Var}\left[\widetilde{\boldsymbol{\theta}}\right] + \left(E[\widetilde{\boldsymbol{\theta}}] - \boldsymbol{\theta}\right)^2$

74

Biased or Unbiased…

- LS Estimator of σ^2: $s^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Unbiased: $s^2 \sim \frac{\sigma^2}{(n-k)}\chi^2_{(n-k)} \rightarrow E[s^2] = \frac{\sigma^2}{(n-k)}E\left[\chi^2_{(n-k)}\right] = \sigma^2$
- What about the estimator of σ?

$$\mathrm{Var}[s^2] > 0$$

$$\mathrm{Var}[s^2] = E[s^2] - (E[s])^2$$

$$= \sigma^2 - (E[s])^2$$

- So that $(E[s])^2 < \sigma^2 \rightarrow E[s] < \sigma$

75

More on the LS Methodology

- The estimator $\widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is called ordinary least squares (OLS) estimator, and it is indicated only in situations with homoscedastic and uncorrelated residuals.
- If the residual variance is heterogeneous (i.e., $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 = w_i\sigma^2$), the residual variance matrix can be expressed as $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}\sigma^2$, where $\mathbf{W}$ is a diagonal matrix with the elements w_j, a better estimator of $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}$, which is generally referred to as weighted least squares (WLS) estimator.
- Furthermore, in situations with a general residual variance-covariance matrix $\mathbf{V}$, including correlated residuals, a generalized least squares (GLS) estimator $\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ is obtained by minimizing the generalized sum of squares, given by $\mathrm{GSS} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$.

76

Maximum Likelihood

- Likelihood Function: any function of the model parameters that is proportional to the density function of the data.
- Hence, to use a likelihood-based approach for estimating model parameters, some extra assumptions must be made regarding the distribution of the data.
- In the case of the linear model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, if the residuals are assumed normally distributed with mean vector zero and variance-covariance matrix $\mathbf{V}$, i.e. $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{V})$, the response vector $\mathbf{y}$ is also normally distributed, with expectation $\mathrm{E}[\mathbf{y}] = \mathbf{X\beta}$ and variance $\mathrm{Var}[\mathbf{y}] = \mathbf{V}$.

77

Maximum Likelihood Estimation

- The distribution of $\mathbf{y}$ has a density function given by:

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{V}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})\right\}$$

so that the likelihood and the log-likelihood functions can be expressed respectively as:

$$\mathrm{L}(\boldsymbol{\beta}, \mathbf{V}) \propto |\mathbf{V}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})\right\}$$

and

$$\mathrm{l}(\boldsymbol{\beta}, \mathbf{V}) = \log[\mathrm{L}(\boldsymbol{\beta}, \mathbf{V})] \propto -\frac{1}{2}|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})$$

78

Maximum Likelihood Estimation

- Assuming $\mathbf{V}$ known, the likelihood equations for $\boldsymbol{\beta}$ are given by taking the first derivatives of $\mathrm{l}(\boldsymbol{\beta}, \mathbf{V})$ with respect to $\boldsymbol{\beta}$ and equating it to zero:

$$\frac{\partial \mathrm{l}(\boldsymbol{\beta}, \mathbf{V})}{\partial \boldsymbol{\beta}} \equiv \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

from which the following system of equations is obtained:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- The maximum likelihood estimator (MLE) for $\boldsymbol{\beta}$ is given then by: $\mathrm{MLE}(\boldsymbol{\beta}) = \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$

79

Maximum Likelihood Estimation

- If the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}$ does not exist, a generalized inverse $(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-}$ can be used to obtain a solution for the system of likelihood equations:

$$\boldsymbol{\beta}^{0} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- Note: Under normality the MLE coincides with the GLS estimator discussed previously. Similarly, in situations in which the matrix $\mathbf{V}$ is diagonal, or when $\mathbf{V}$ can be represented as $\mathbf{V} = \mathbf{I}\sigma^2$, the MLE coincides with the WLS and the OLS estimators, respectively.

80

Maximum Likelihood Estimation

- The expectation and the variance-covariance matrix of the MLE are given by:

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathrm{E}[\mathbf{y}] = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$
$$\mathrm{Var}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Var}[\mathbf{y}]\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$$

- As $\widehat{\boldsymbol{\beta}}$ is a linear combination of the response vector $\mathbf{y}$, we have that $\widehat{\boldsymbol{\beta}} \sim \mathrm{MVN}\left(\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\right)$, from which confidence intervals (regions) and hypothesis testing regarding any (set of) element(s) of $\boldsymbol{\beta}$ can be easily obtained.

81

Maximum Likelihood Estimation

- Note: In the case of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}\sigma^2)$, it can be shown that:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \rightarrow \widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2)$$
$$\widehat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|^2$$

$$\widehat{\sigma}^2 \sim \sigma^2 \frac{\chi^2_{(n-k)}}{n} \quad \rightarrow \quad \mathrm{E}[\widehat{\sigma}^2] = \frac{n-k}{n}\sigma^2$$

$$\widetilde{\sigma}^2 = \frac{n}{n-k}\widehat{\sigma}^2 = \frac{1}{n-k}\left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|^2 = s^2 \quad \rightarrow \quad \widetilde{\sigma}^2 \sim \sigma^2 \frac{\chi^2_{(n-k)}}{n-k}$$

82

Properties of Maximum Likelihood Estimators

- Consistency: $E[\hat{\theta}] \xrightarrow{n\to\infty} \theta$
- Invariance: $\hat{\theta} = \mathrm{MLE}(\theta) \rightarrow g(\hat{\theta}) = \mathrm{MLE}[g(\theta)]$
- Asymptotic normality and efficiency:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$$

  where $I(\theta)$ is the Fisher information matrix (Crámer-Rao lower bound: $\mathrm{Var}(\tilde{\theta}) \geq I(\theta)^{-1}$)
- Relation to Bayesian inference: A maximum likelihood estimator coincides with the posterior mode given a uniform prior distribution on the parameters

83

Multicollinearity

- Multicollinearity (also collinearity) is a linear association between predictors variables, i.e. the predictor variables are correlated.
- Consequence: regression coefficient estimates may change erratically in response to small changes in the model or the data.
- Multicollinearity however does not reduce the predictive power or reliability of the model.
- Under extreme multicollinearity, parameters may be not estimable.
- Detection: Large changes in the estimates when a predictor variable is added or deleted; variance inflation factor (VIF)
- Modelling alternatives: Variable Selection, Dimension Reduction, Shrinkage Estimation

84

Building a Regression Model for Prediction

- Descriptive analysis; one-variable-at-a-time models, pairwise relationships (scatter plots and correlations)
- Prior knowledge (application domain expertise) to get a starting point, i.e. variables to include in the model
- Try adding more variables, for example using results from descriptive analysis
- Pruning of variables based on results (coefficients sign and p-values)
- Try interactions, especially between inputs with large effects
- Some trial & error, there is not a universal recipe

85

High-Dimensional Model for Prediction

- Exhaustive search generally impractical
- Search algorithms (simulated annealing, genetic algorithms)
- Alternative model comparison criteria (AIC, BIC, etc.)
- Model building strategies will depend on sample size, number of input variables, and other models characteristics (random effects, covariance structure search, non-linear terms, etc.)
- Some dimension reduction techniques, variable selection, and shrinkage estimation will be discussed later

86

Logistic Regression

- Linear (simple or multiple) regression is used to model continuous outcomes while logistic regression deals with binary (yes or no) outcomes
- $y_i = 0$ or $y_i = 1$ → $p_i = \text{Prob}(y_i = 1)$
- In the logistic model, the log-odds (the logarithm of the odds) is a linear combination of the predictor variables:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

87

Logistic Regression

- $\text{Prob}(y_i = 1) = \text{logit}^{-1}(\eta_i)$, where $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ is the linear predictor
- The function $\text{logit}^{-1}(w) = \frac{e^w}{1+e^w}$ transforms continuous values to the range (0,1)
- $\text{Prob}(y_i = 1) = p_i$, $\text{logit}(p_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$
- Odds ratio: odds: $\frac{p}{1+p}$, ratio of two odds: $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

$$\log\left(\frac{\text{Prob}(y_i=1|x_i)}{\text{Prob}(y_i=0|x_i)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

88

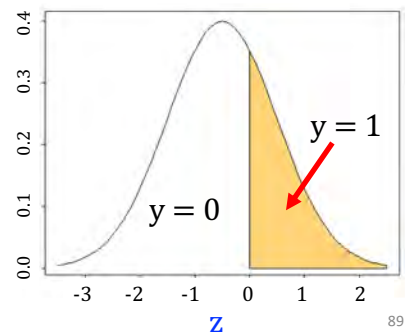
Logistic Regression

- Latent formulation: $y_i = \begin{cases} 1, & \text{if } z_i > 0 \\ 0, & \text{if } z_i < 0 \end{cases}$

  where $z_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$, with ε_i independent logistic probability distribution, i.e. $\text{Prob}(\varepsilon_i < w) = \text{logit}^{-1}(w)$

- Hence:

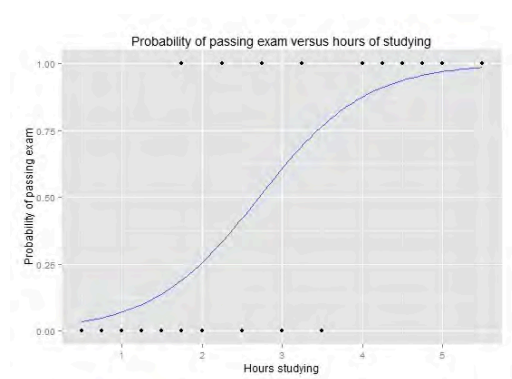
$$\begin{aligned} \text{Prob}(y_i = 1) &= \text{Prob}(z_i > 0) \\ &= \text{Prob}\left(\varepsilon_i > -\mathbf{x}_i^T \boldsymbol{\beta}\right) \\ &= \text{logit}^{-1}\left(\mathbf{x}_i^T \boldsymbol{\beta}\right) \end{aligned}$$



89

Example: Probability of passing an exam versus hours of study

Hours	Pass
0.50	0
0.75	0
1.00	0
1.25	0
1.50	0
1.75	0
1.75	1
2.00	0
2.25	1
2.50	0
2.75	1
3.00	0
3.25	1
3.50	0
4.00	1
4.25	1
4.50	1
4.75	1
5.00	1
5.50	1



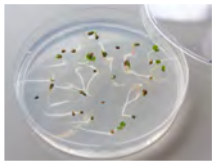
$$\log\left(\frac{p}{1-p}\right) = -4.0777 + 1.5046 \times \text{Hours}$$

$$p = \frac{1}{1 + \exp(4.0777 - 1.5046 \times \text{Hours})}$$

90

Generalized Linear Models

- The models discussed so far assumed a Gaussian (normal) distribution of the response variables
- Often however such variables are expressed as a binary (e.g., pregnancy in dairy cattle, or germination in seeds) or count variable (e.g., litter size in swine, or fruits in trees)
- In such cases the linear (Gaussian) model is not appropriate, and a generalized linear model (GLM) approach is necessary




91

Generalized Linear Models

- GLM can actually model outcomes (response variables) generated from any distribution from the exponential family, which includes the normal, binomial, Poisson and gamma distributions, among others
- The GLM consists of three elements:
  1. Probability distribution from the exponential family
  2. Linear predictor $\eta = X\beta$
  3. Link function g such that $E(Y) = \mu = g^{-1}(\eta)$

92

The Exponential Family of Distributions

- Exponential family: set of probability distributions whose probability density (or mass) function can be expressed as:

$$p(y|\theta) = h(y)\exp[\eta(\theta) \cdot T(y) - A(\theta)]$$

where $h(y)$, $\eta(\theta)$, $T(y)$ and $A(\theta)$ are known functions.

- Exponential families include: Bernoulli, beta, binomial (with fixed number of trials), categorical, chi-squared, Dirichlet, exponential, gamma, geometric, inverse Wishart, multinomial (with fixed number of trials), negative binomial (with fixed number of failures), normal, Poisson, Wishart, among others.

93

The Exponential Family of Distributions

- Example with Gaussian Distribution:

$$\begin{aligned} p(y|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\ &= \frac{1}{\sqrt{2\pi}}\exp\left[-\log(\sigma) - \frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \\ &= h(y)\exp[\eta(\theta) \cdot T(y) - A(\theta)] \end{aligned}$$

where $h(y) = \frac{1}{\sqrt{2\pi}}$, $\eta(\theta) = [\mu/\sigma^2 \quad -1/(2\sigma^2)]^T$,

$T(y) = [y \quad y^2]$ and $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma)$.

94

Generalized Linear Models

- Common distributions and canonical link functions:

Distribution	Link name	Link function, $X\beta = g(\mu)$	Mean function, $\mu = X\beta$
Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Exponential	Negative inverse	$X\beta = -\mu^{-1}$	$\mu = -(X\beta)^{-1}$
Gamma			
Inverse Gamma	Inverse squared	$X\beta = \mu^{-2}$	$\mu = (X\beta)^{-1/2}$
Poisson	Log	$X\beta = \log(\mu)$	$\mu = \exp(X\beta)$
Bernoulli, Binomial	Logit	$X\beta = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$
Categorical, Multinomial			

95

Overdispersion

- Example with Poisson: $y_i = \text{Poisson}(\mu_i)$, where $\mu_i = \exp(X_i\beta)$
- $E[y_i] = \text{Var}[y_i] = \mu_i = \exp(X_i\beta)$
- Exposure input: $y_i = \text{Poisson}(h_i\mu_i)$, where $\mu_i = \exp(X_i\beta)$ and $\log(h_i)$ is called offset; $E[y_i] = \text{Var}[y_i] = h_i\mu_i$
- Z-score: $z_i = \frac{y_i - \hat{y}_i}{\text{sd}(\hat{y}_i)} = \frac{y_i - h_i\hat{\mu}_i}{\sqrt{h_i\hat{\mu}_i}} \approx N(0,1)$, where $\hat{\mu}_i = \exp(X_i\hat{\beta})$
- Estimated overdispersion: $\frac{1}{n-k}\sum_{i=1}^{n} z_i^2$, as $\sum_{i=1}^{n} z_i^2 \sim \chi^2_{n-k}$

96

Overdispersion

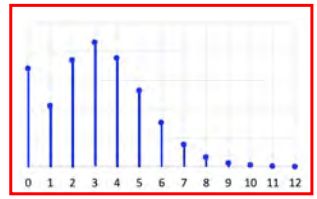
- Overdispersed-Poisson (or negative-binomial model)

  y_i = overdispersed Poisson$(h_i \exp(X_i\beta), w)$

  where w is the overdispersion parameter: $\mathrm{Var}[y_i] = wE[y_i]$

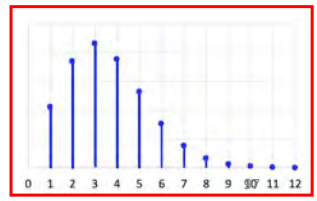
- Zero-inflated Poisson (ZIP)

$$\begin{cases} \Pr(Y=0) = \pi + (1-\pi)e^{-\lambda} \\ \Pr(Y=k) = (1-\pi)\dfrac{\lambda^k e^{-\lambda}}{k!} \quad , k = 1,2,3,\ldots \end{cases}$$

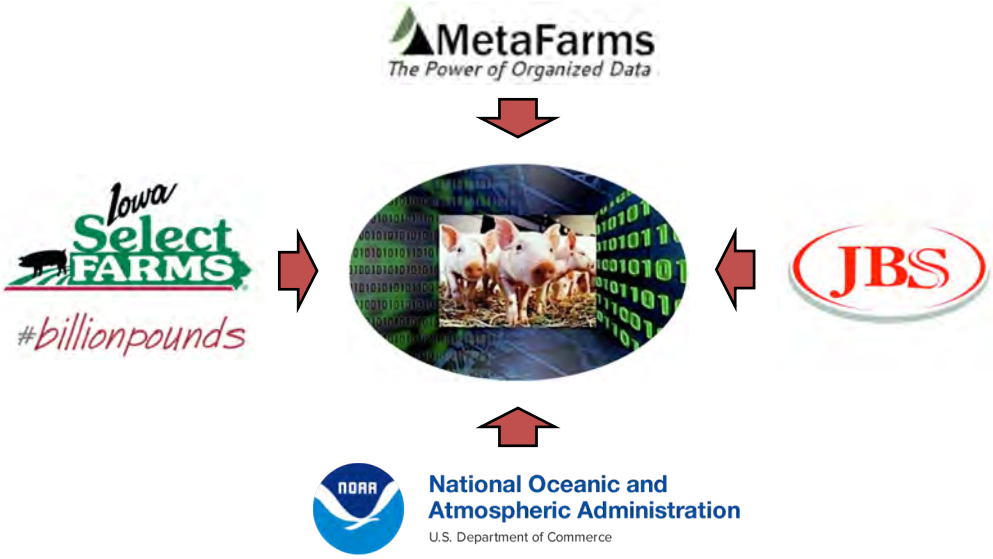


- Zero-truncated Poisson (ZTP)

$$\Pr(Y=k) = \frac{\lambda^k}{(e^{\lambda}-1)k!}, \quad k = 1,2,3,\ldots$$



Example: Pig Production Data Analytics



98

Generalized Additive Mixed Models

General Linear Model: $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

Non-Gaussian distribution (exponential family)

$$g(E[y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Generalized Linear Model

Random effects

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} b_k x_{ik} + \varepsilon_i$$

$$b_k \sim N(0, \sigma_k^2)$$

Linear Mixed Model

Non-linear relationships (smooth functions)

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i$$

Additive Model

Hastie T and Tibshirani R (1986) Generalized additive models. Stat. Sci. 1: 297-318.

GAM:

$$g(\mu_i) = \boldsymbol{a}_i \boldsymbol{\theta} + \sum_{j=1}^{p} f_j(x_{ij}) \quad \begin{cases} \mu_i = E(y_i) \\ y_i \sim \mathrm{EF}(\mu_i, \varphi) \quad (i = 1, \ldots, n) \end{cases}$$

row i of incidence matrix

Vector of parameters

Unknown smooth functions of covariates x_{ij}

- Smooth functions commonly depicted by reduced rank smoothing splines, including different kind of polynomials such as the P-spline, adaptive variants, tensor products, thin plate, and cubic splines
- Any reduced rank smoothing spline can be represented as $f_j = \mathbf{X}_j \beta_j$, in which $\mathbf{X}_j$ is an $n \times p_j$ incidence matrix containing the smooth spline basis functions evaluated at vector $\mathbf{x}_j$, and β_j is the corresponding regression coefficient vector
- Type and size of the basis functions must be defined to prevent model overfitting, for example using a penalization term in the model likelihood
- Fitting a GAM can be performed by penalized iteratively re-weighted least squares, given the smoothing parameters
- Smoothing parameters can be estimated by generalized cross-validation or by restricted maximum likelihood estimation
- GAM can be extended to accommodate random effects using empirical Bayesian approach

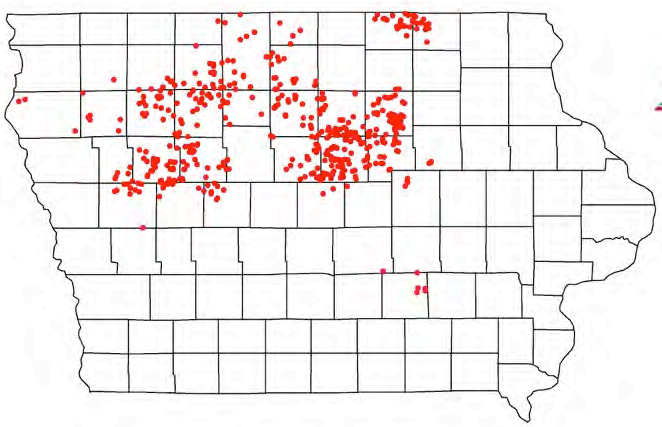
Wood, S. N. 2017. Generalized Additive Models: An Introduction with R. 2th ed. CRC Press.

Pig Production Data Analytics

- Data from 2013 to 2016
- More than 100 variables:
  - Performance: Average daily gain, feed conversion, mortality, final weight, initial weight, days on feed, etc.
  - Economics: Profit, income, expenses, feed cost, genetic sales, etc.
  - Management: Number of empty days, vaccinations, etc.
  - Facilities: Type of feeder, type of drinker, construction age, supervisor, manager, etc.

101

Location of ISF Finishing Farms



100 km

102

Factors Associated with Total Transport Losses

- Dead on arrival (DOA)
- Downer or slower hogs
  - Direct economic losses for producers
  - Animal welfare and well-being concern

Passafaro TL, Van de Stroet D, Bello NM, Williams NH and Rosa GJM (2019) Generalized additive mixed model on the analysis of total transport losses of market-weight pigs. *Journal of Animal Science* 97: 2025-2034.

103

Material and Methods

- Integration of movement and weather data
  - Market-weight pigs
  - July of 2014 to December of 2015
- Data editing
  - Missing information
  - Truck companies with less than 20 shipments
  - Shipments with <100 or >210 pigs
  - Farm - quarter of year combination with <5 records
- Final data
  - 26,819 shipments
  - 420 farms
  - 2 processing plants
  - 4,567,514 market-weight hogs

104

Description of the variables recorded per shipment during 1.5 years at ISF.

Variables	Description	Number of levels	Variable type	Comments
Total losses, %	DOA plus slower pigs	-	Response	-
Truck company	The transportation company	78	Explanatory/random	-
Site and year	Concatenation of farm and year	797	Explanatory/random	-
Group type	Finish or wean to finish	2	Explanatory/fixed	-
Abattoir	Meat plant of destination	2	Explanatory/fixed	
Season	Fall, spring, summer, and winter	4	Explanatory/fixed	
Driver	Owner or employee	2	Explanatory/fixed	-
Number of pigs	Number of pigs in the shipment	-	-	
Average body weight	Measured in lbs	-	Explanatory/fixed	
Travel distance	Measured in km	-	Explanatory/fixed	*gmapdistance* (Melo & Zaruk, 2010)
Wind speed	Measured in mps	-	Explanatory/fixed	-
Precipitation	Measured in mm	-	Explanatory/fixed	-
THI	-	-	Explanatory/fixed	(NOAA, 1978) I = T −[0.55−(0.0055xRH)]x[T−14.5]

The weather conditions were estimated with a WKNN using the R package *kknn* (Hechenbichler & Schliep, 2004), with 22 weather stations

105

Descriptive statistics for transport losses and continuous explanatory variables.

Variable	Mean	SD	Minimum	Maximum
DOA, %	0.19	0.45	0.00	7.69
DOWN, %	0.57	0.85	0.00	12.80
Total losses, %	0.76	1.05	0.00	14.02
Number of pigs per shipment	170.3	8.4	100.00	201.00
Average body weight, lbs	276.6	12.7	230.6	319.8
Travel distance, km	136.6	63.4	35.6	396.5
Wind speed, mps	4.2	1.8	0.5	11.0
Precipitation, mm	2.3	5.9	0.00	58.1
THI	9.7	9.6	-16.5	26.3

DOA = Dead on arrival; DOWN = Losses due to downer hogs; THI = Temperature humidity index

106

Materials and Methods

- Statistical model
  - Generalized Additive Mixed Models (GAMM): linear predictor specified in terms of smooth functions of covariates (Lin and Zhang, 1999)
- Base generalized linear mixed model
  - Random effects: combination of farm - quarter of the year, and truck company
  - Fixed effects: abattoir, type of driver, management group, distance traveled, average weight, wind speed, precipitation, and THI
- Forward stepwise procedure
  - Model deviance, Biological meaning, Significance
  - Pairwise interactions

107

Final Model

- Base model without management group plus two interactions:
  - Abattoir x average market-weight
  - Wind speed x precipitation

$$(y_i|\alpha_{k[i]}, \gamma_{j[i]}) \sim \text{Overdispersed Binomial}(n_i, p_i, \varphi)$$

$$\begin{aligned}\text{logit}(p_i) = {} & \alpha_{k[i]} + \gamma_{j[i]} + a_{b[i]} + d_{o[i]} + \sum_{e=1}^{9} b_{t_e}(t_i)\beta_{t_e} \\ & + \sum_{g=1}^{9} b_{s_g}(s_i)\beta_{s_g} + \sum_{h=1}^{9} b_{r_h}(r_i)\beta_{r_h} \\ & + \sum_{p=1}^{9} b_{u_p}(u_i)\beta_{u_p} + \sum_{c=1}^{9} b_{w_c}(w_i)_{[a_{b[i]}]}\beta_{w_c[a_{b[i]}]} \\ & + \sum_{l=1}^{5}\sum_{m=1}^{5} \beta_{s_l,r_m} b_{s_l}(s_i) b_{r_m}(r_i)\end{aligned}$$

- Analysis implemented with the R package mgcv (Wood, 2017)

108

Results

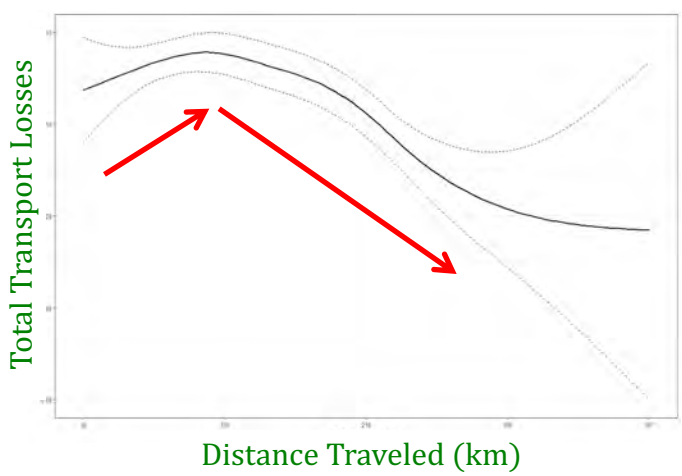
Table 3. Parameter estimates and approximate significance level of smoothing functions on total transport losses of market-weight pigs using a generalized additive mixed model

			Confidence interval (odds ratios)		
Parameters	**Estimates**	**Odds ratios**	**Lower limit**	**Upper limit**	***P*-value**
Intercept	−4.945	0.007	0.006	0.008	<0.0001
Abattoir: B	−0.323	0.720	0.719	0.722	<0.0001
Driver: Owner	−0.142	0.868	0.866	0.870	<0.0001
Smoothing functions			**EDF[1]**	**Ref. DF[2]**	***P*-value**
Distance travelled			3.2862	9	0.0034
THI			5.0547	9	<0.0001
Wind speed			0.0004	9	0.9998
Precipitation			1.0021	9	0.8837
Average market weight × abattoir A			4.9742	9	<0.0001
Average market weight × abattoir B			5.0547	9	<0.0001
Wind speed × precipitation			1.5295	16	0.0209

EDF = effective degree of freedom; Ref. DF = reference number of degrees of freedom.

109

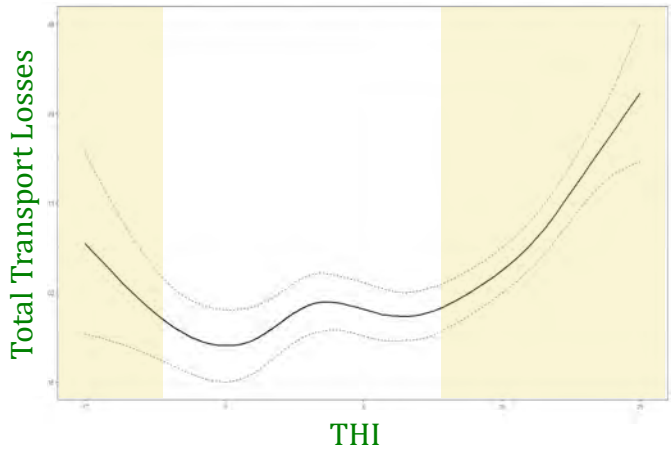
Results



Predicted total transport losses of market weight pigs on the odds ratio scale

110

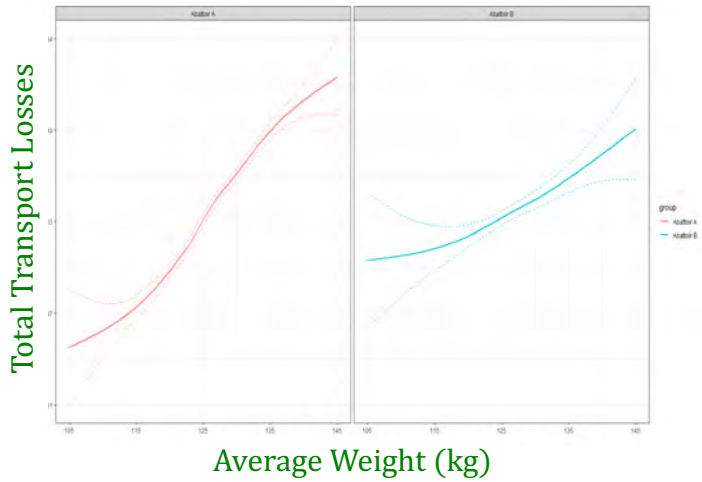
Results



Predicted total transport losses of market weight pigs on the odds ratio scale

111

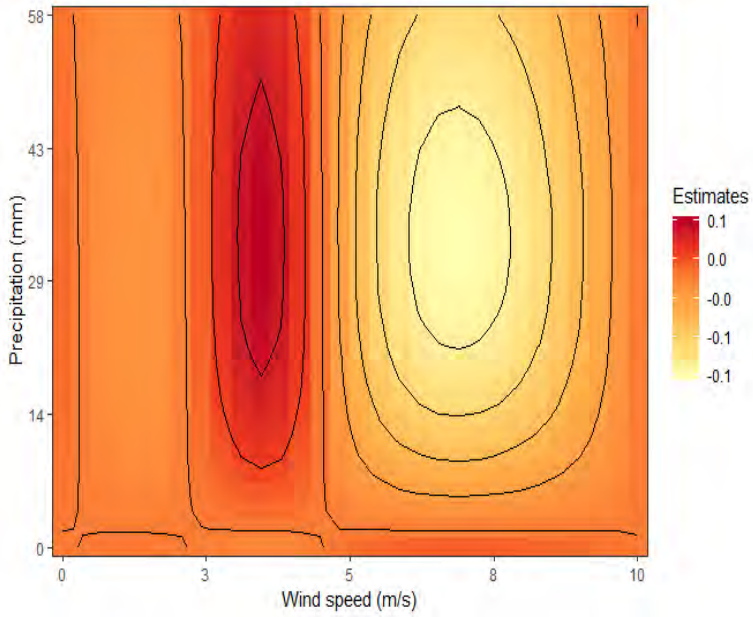
Results



Predicted total transport losses of market weight pigs on the odds ratio scale

112

Results



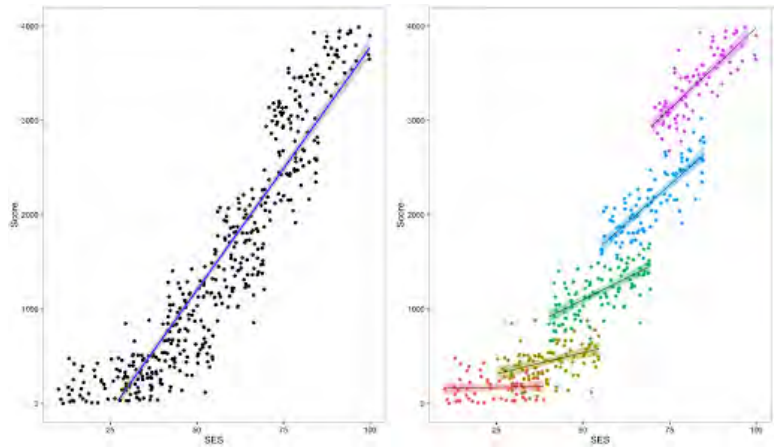
113

Conclusion

- Total transport losses caused by a complex system involving multiple interacting factors, and non-linear relationships
- Understanding factors associated with total transport losses might assist farmers to improve management, profit, and animal welfare
- GAMM is a flexible approach to model total transport losses, accommodating both random and fixed effects, and non-linear relationships

114

Multilevel and Hierarchical Models



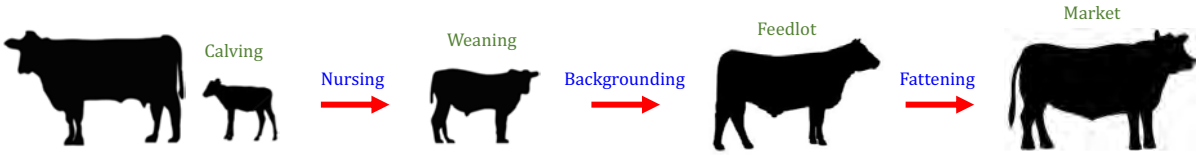
115

Outline

- Prediction with Multilevel Data
- Mixed Model Methodology
- Overview and Derivation
- BLUE and BLUP
- Example

116

Beef Feedlots

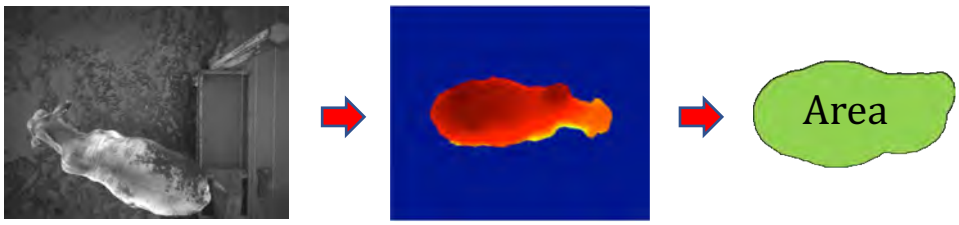


Once young calves reach a weight of 300-700 pounds (140 to 320 kg) they are rounded up and transferred to a feedlot, where they gain an additional 400-600 pounds (220 kg) on about 6-8 months.

117

Prediction of Final Weight

- Data on image feature (e.g. top-view body area) of cattle (explanatory variable x) at beginning of finishing phase and final carcass weight (response variable y):



- Predictive model: $y = \mu + b\,x + e$

Intercept slope error term

118

Prediction of Final Weight

- Data:

Animal	Area (x)	Weight (y)
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

- Animal j: (x_j, y_j), $(j = 1,2, \dots, n)$
- Model: $y_j = \mu + \beta\, x_j + e_j$, with $e \sim (0, \sigma_e^2)$
- Predictions: $\tilde{y} = \hat{\mu} + \hat{\beta}\, x_0$, $\mathrm{Var}(\tilde{y}) = \hat{\sigma}_e^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$

119

Data from Multiple Feedlots

- Suppose data on top-view body area (x) and carcass weight (y) are obtained on cattle $(j = 1,2, \dots, n_i)$ from multiple feedlots $(i = 1,2, \dots, k)$
- Proposed model: $y_{ij} = \mu + \beta\, x_{ij} + e_{ij}$
- Notice: a model that ignores group effects (feedlot effect in this case) will tend to understate the prediction error because of group-to-group variability

120

Multilevel (Hierarchical) Regression

- Model with farm effects (i.e. farm-specific intercepts) as well as the interaction between farm and the covariable x (i.e. farm-specific slopes): $y = \text{farm} + \text{farm} \times x + \text{error}$
- Equivalently: $y_{ij} = f_i + \beta_i x_{ij} + e_{ij}$
- Assuming farm effect as fixed and $e_{ij} \sim N(0, \sigma_e^2)$:

$$\boxed{E[y_{ij}] = f_i + \beta_i x_{ij} \quad \text{and} \quad Var[y_{ij}] = \sigma_e^2}$$

- Predictions:

Future animal on surveyed feedlot: $\tilde{y}_{i0} = \hat{f}_i + \hat{\beta}_i x_{i0}$, $Var(\tilde{y}_{i0}) = \hat{\sigma}_{\tilde{y}}^2$

Future animal on future feedlot: $\tilde{y}_{00} = \text{wild guess}$, $Var(\tilde{y}_{00}) = \infty$

121

Multilevel (Hierarchical) Regression

- Model: $y_{ij} = (\alpha + a_i) + (\beta + b_i) x_{ij} + e_{ij}$

  with $a_i \sim N(0, \sigma_a^2)$, $b_i \sim N(0, \sigma_b^2)$ and $Cov(a_i, b_i) = \sigma_{ab}$
- Marginally: $E[y_{ij}] = \alpha + \beta x_{ij}$ and $Var[y_{ij}] = \sigma_a^2 + x_{ij}^2 \sigma_b^2 + 2\sigma_{ab} + \sigma_e^2$
- Conditionally: $E[y_{ij} | a_i, b_i] = (\alpha + a_i) + (\beta + b_i) x_{ij}$ and $Var[y_{ij} | a_i, b_i] = \sigma_e^2$
- Predictions:

Future animal on surveyed feedlot:

$\tilde{y}_{i0} = (\hat{\alpha} + \hat{a}_i) + (\hat{\beta} + \hat{b}_i) x_{i0}$, $Var(\tilde{y}_{i0}) = \hat{\sigma}_{\tilde{y}}^2$

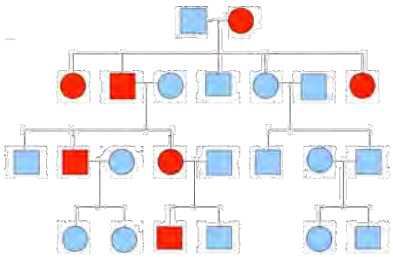
Future animal on future feedlot:

$\tilde{y}_{00} = \hat{\alpha} + \hat{\beta} x_{ij}$, $Var(\tilde{y}_{00})$ includes (co)variance components

122

Mixed Models

Overview and Derivation of the Mixed Model



Charles Roy Henderson
(1911-1989)

General Linear Model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

responses

residuals

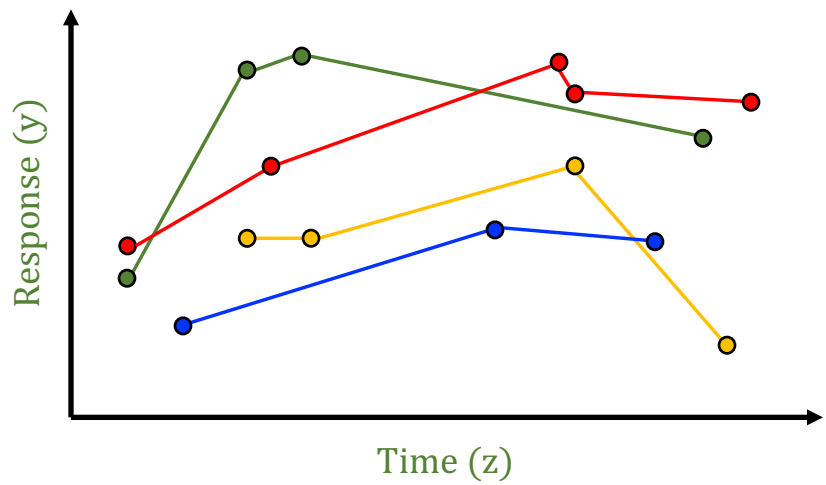
design/incidence matrix (known)

overall mean + fixed effects parameters

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, I_n\sigma^2) \;\rightarrow\; \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Analysis of Longitudinal Data

- Suppose a series of longitudinal data (e.g., repeated measurements on time) on n individuals.



125

Two-stage Analysis of Longitudinal Data

Step 1

- Let y_{ij} represent the observation j ($j = 1,2,\ldots,n_i$) on individual i ($i = 1,2,\ldots,n$), and the following quadratic regression of measurements on time (z_{ij}) for each individual:

$$y_{ij} = \beta_{0i} + \beta_{1i} z_{ij} + \beta_{2i} z_{ij}^2 + \varepsilon_{ij}$$

where β_{0i}, β_{1i} and β_{2i} are subject-specific regression parameters, and ε_{ij} are residual terms, assumed normally distributed with mean zero and variance σ_ε^2

126

- In matrix notation such subject-specific regressions can be expressed as:

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad (1)$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})^T$,

$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})^T \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$ and

$$\mathbf{Z}_i = \begin{bmatrix} 1 & z_{i1} & z_{i1}^2 \\ 1 & z_{i2} & z_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & z_{in_i} & z_{in_i}^2 \end{bmatrix}$$

127

- Under these specifications, the least-squares estimate of β_i is:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{y}_i$$

- Note that this is also the maximum likelihood estimate of β_i
- Such estimates can be viewed as summary statistics for the longitudinal data, the same way one could use area under the curve (AUC), or peak (maximum value of y_{ij}), or mean response.

128

Two-stage Analysis of Longitudinal Data

Step 2

- Suppose now we are interested on the effect of some other variables (such as gender, treatment, year, etc.) on the values of β_i
- Such effects could be studied using a model as:

$$\hat{\boldsymbol{\beta}}_i = \mathbf{W}_i\boldsymbol{\beta} + \mathbf{u}_i$$

where $\mathbf{u}_i \sim N(\mathbf{0},\mathbf{D})$, which is an approximation for the model:

$$\boldsymbol{\beta}_i = \mathbf{W}_i\boldsymbol{\beta} + \mathbf{u}_i \qquad (2)$$

129

Single-stage Analysis of Longitudinal Data

- The two step-analysis described here can be merged into a single stage approach by substituting (2) in (1):

$$\mathbf{y}_i = \mathbf{Z}_i[\mathbf{W}_i\boldsymbol{\beta} + \mathbf{u}_i] + \boldsymbol{\varepsilon}_i$$

which can be expressed as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

where $\mathbf{X}_i = \mathbf{Z}_i\mathbf{W}_i$. By concatenating observations from multiple individuals, we have the following mixed model:

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}}$$

130

Mixed Effects Models

- Frequently, linear models contain factors whose levels represent a random sample of a population of all possible factor levels
- Models containing both fixed and random effects are called mixed effects models
- Linear mixed effects models have been widely used in analysis of data where responses are clustered around some random effects, such that there is a natural dependence between observations in the same cluster
- For example, consider repeated measurements taken on each subject in longitudinal data, or observations taken on members of the same family in a genetic study

131

Linear Mixed Effects Model

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}}$$

where:

- $\mathbf{y}$: response vector; observations
- $\boldsymbol{\beta}$: vector of fixed effects
- $\mathbf{u}$: vector of random effects; $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{G})$
- $\mathbf{X}$ and $\mathbf{Z}$: (known) incidence matrices
- $\mathbf{e}$: residual vector; $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$

132

Linear Mixed Effects Model

- Generally, it is assumed that $\mathbf{u}$ and $\mathbf{e}$ are independent from each other, such that:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}\end{bmatrix}\right)$$

- Inferences regarding mixed effects models refer to the estimation of fixed effects, the prediction of random effects, and the estimation of variance and covariance components, which are briefly discussed next

133

Estimation of Fixed Effects

- Let $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = \mathbf{Zu} + \mathbf{e}$

$$\begin{cases} \mathrm{E}[\boldsymbol{\varepsilon}] = \mathrm{E}[\mathbf{Zu} + \mathbf{e}] = \mathbf{Z}\mathrm{E}[\mathbf{u}] + \mathrm{E}[\mathbf{e}] = \mathbf{0} \\ \mathrm{Var}[\boldsymbol{\varepsilon}] = \mathrm{Var}[\mathbf{Zu} + \mathbf{e}] = \mathbf{Z}\mathrm{Var}[\mathbf{u}]\mathbf{Z}^{\mathrm{T}} + \mathrm{Var}[\mathbf{e}] = \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{\Sigma} \end{cases}$$

such that $\mathbf{y} \sim \mathrm{MVN}(\mathbf{X\beta}, \mathbf{V})$, where $\mathbf{V} = \mathbf{ZG}\boldsymbol{Z}^{T} + \Sigma$

- Under these circumstances, the MLE for $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} \sim \mathrm{MVN}(\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1})$$

134

Estimation of Fixed Effects

- As $\mathbf{G}$ and $\boldsymbol{\Sigma}$ are generally unknown, an estimate of $\mathbf{V}$ is used instead such that the estimator becomes:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{y}$$

- The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is now approximated by $(\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$

- Note: $(\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$ is biased downwards as a consequence of ignoring the variability introduced by working with estimates of (co)variance components instead of their true (unknown) parameter values

135

Estimation of Fixed Effects

- Approximated confidence regions and test statistics for estimable functions of the type $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta}$ can be obtained by using the result:

$$\frac{(\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta}^{0})^{\mathrm{T}}(\mathbf{K}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{K})^{-1}(\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta}^{0})}{\mathrm{rank}(\mathbf{K})} \approx \mathrm{F}_{[\varphi_{\mathrm{N}},\varphi_{\mathrm{D}}]}$$

where $\mathrm{F}_{[\varphi_{\mathrm{N}},\varphi_{\mathrm{D}}]}$ refers to an F-distribution with $\varphi_{\mathrm{N}} = \mathrm{rank}(\mathbf{K})$ degrees of freedom for the numerator, and φ_{D} degrees of freedom for the denominator, which is generally calculated from the data using, for example, the Satterthwaite's approach

136

Estimation (Prediction) of Random Effects

- In addition to the estimation of fixed effects, very often in genetics interest is also on prediction of random effects.
- In linear (Gaussian) models such predictions are given by the conditional expectation of **u** given the data, i.e. $\mathrm{E}[\mathbf{u}|\mathbf{y}]$.
- Given the model specifications, the joint distribution of **y** and **u** is:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix}\mathbf{X\beta}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{V} & \mathbf{ZG}\\ \mathbf{GZ}^{\mathrm{T}} & \mathbf{G}\end{bmatrix}\right)$$

137

Estimation (Prediction) of Random Effects

- From the properties of multivariate normal distribution, we have that:

$$\begin{aligned}\mathrm{E}[\mathbf{u}\,|\,\mathbf{y}] &= \mathrm{E}[\mathbf{u}] + \mathrm{Cov}[\mathbf{u},\mathbf{y}^{\mathrm{T}}]\mathrm{Var}^{-1}[\mathbf{y}](\mathbf{y}-\mathrm{E}[\mathbf{y}])\\ &= \mathbf{GZ}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X\beta}) = \mathbf{GZ}^{\mathrm{T}}(\mathbf{ZGZ}^{\mathrm{T}}+\mathbf{\Sigma})^{-1}(\mathbf{y}-\mathbf{X\beta})\end{aligned}$$

- The fixed effects **β** are typically replaced by their estimates, so that predictions are made based on the following expression:

$$\hat{\mathbf{u}} = \mathbf{GZ}^{\mathrm{T}}(\mathbf{ZGZ}^{\mathrm{T}}+\mathbf{\Sigma})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

138

Mixed Model Equations

- Henderson (1950) presented the mixed model equations (MME) to estimate $\boldsymbol{\beta}$ and $\mathbf{u}$ simultaneously, without the need for computing $\mathbf{V}^{-1}$
- The MME were derived by maximizing (for $\boldsymbol{\beta}$ and $\mathbf{u}$) the joint density of $\mathbf{y}$ and $\mathbf{u}$, and can be expressed as:

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} \end{bmatrix}$$

139

BLUE and BLUP

- Using the second part of the MME, we have that:

$$\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{G}^{-1})\hat{\mathbf{u}} = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

so that: $\hat{\mathbf{u}} = (\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$

- It can be shown that this expression is equivalent to:

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^{\mathrm{T}}(\mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}}+\boldsymbol{\Sigma})^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

and, more importantly, that $\hat{\mathbf{u}}$ is the best linear unbiased predictor (BLUP) of $\mathbf{u}$

140

BLUE and BLUP

- Using this result into the first part of the MME, we have that:

$$\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

$$\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

$$\hat{\boldsymbol{\beta}} = \{\mathbf{X}^T[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}]\mathbf{X}\}^{-1}\mathbf{X}^T[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}]\mathbf{y}$$

- Similarly, it can be shown that this expression is equivalent to $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$, which is the best linear unbiased estimator (BLUE) of β.

141

Variance Components

- Notice that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ require knowledge of $\mathbf{G}$ and $\boldsymbol{\Sigma}$. These matrices, however, are rarely known. This is a problem without an exact solution using classical methods.
- The practical approach is to replace $\mathbf{G}$ and $\boldsymbol{\Sigma}$ by their estimates ($\hat{\mathbf{G}}$ and $\hat{\boldsymbol{\Sigma}}$) into the MME:

$$\begin{bmatrix} \mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}+\hat{\mathbf{G}}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y} \end{bmatrix}$$

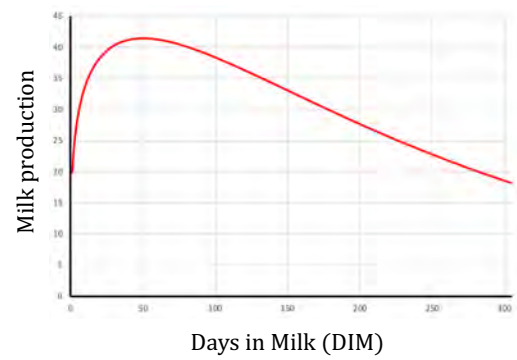
- Variance and covariance components estimation:
  - Analysis of Variance (ANOVA)
  - Maximum likelihood
  - Restricted maximum likelihood (REML)
  - Bayesian approach

142

Example: Lactation Curves

- Wood's model: $y = at^b e^{-ct} + \varepsilon$

  where y is the milk production in day t, and parameters interpreted as scale (parameter a), rate of increase (parameter b), and rate of decay (parameter c)



Li, M., Rosa, G. J. M., Reed, K. F and Cabrera, V. E. (2022) Investigating the impact of temporal, geographic, and management factors on US Holstein lactation curve parameters. *Journal of Dairy Science* 00(0): 0–0. (submitted)

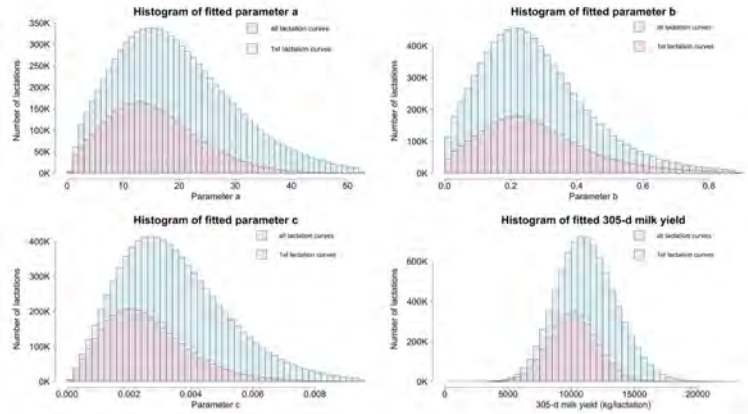
143

Material and Methods

- Test-day milk records on 10M+ lactations of US Holstein cows
- Effects of spatial (farm region), temporal (calving year, and calving month), and management (milking frequency, age at calving for 1st lactation, and parity) factors on lactation curve parameters
- Two-step approach:
  1) Individual animal-parity parameter estimation using the Wood's model (non-linear least-squares optimization)
  2) Mixed-effects model analysis of parameter estimates (a, b, and c) from individual lactation curves
- Fixed effects of spatial, temporal, and management factors, plus the random effects of animals and herds.

144

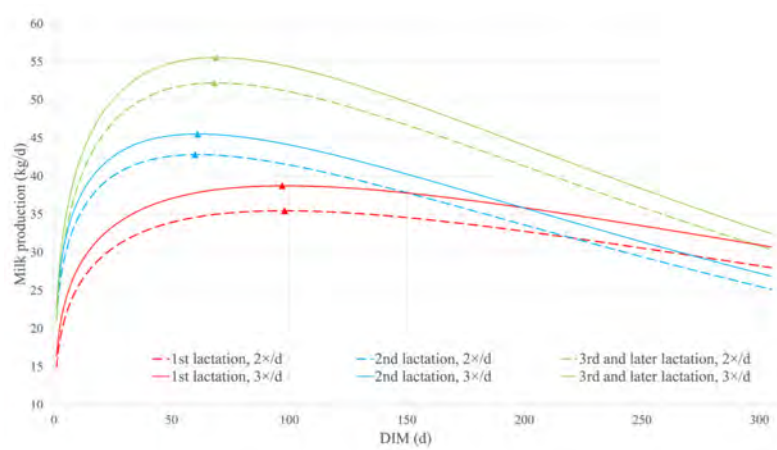
Results



Histograms of fitted individual lactation curve ($y = at^b e^{-ct} + \varepsilon$, Wood 1967) parameters a, b, c, and 305-d milk yield ($a \int_1^{305} t^b e^{-ct} dt$) for all-lactations models and for 1st lactation models.

145

Results

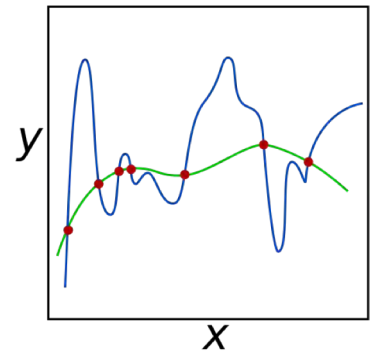


Lactation curves plotted according to the estimated mean of the lactation curve parameters for each lactation group and milking frequency. Triangles indicate the peak.

146

Regularized Regression

- Pros and Cons of Least-Squares
- Regularization Techniques
  - Variable Selection
  - Dimension Reduction
  - Shrinkage Estimation
- Model Selection
- Cross-Validation Techniques
- Predictive Quality Metrics



147

Least Squares Regression

- Model assumptions: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, with $\mathbf{e} \sim (\mathbf{0}, \mathbf{I}\sigma^2)$, i.e. $e_i \overset{\text{i.i.d.}}{\sim} (0, \sigma^2)$
- The least squares estimate of $\mathbf{b}$ minimizes the residual sum of squares, which is given by: $\text{RSS} = \hat{\mathbf{e}}^T\hat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$
- Taking the derivatives and equating them to zero…

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

- Hat matrix (projection matrix): $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\mathbf{H}}\mathbf{y}$

148

Least Squares

- Expectation: $E[\hat{\mathbf{b}}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{y}]$
  $= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{b}$ (unbiased estimator)
- Variance: $\mathrm{Var}[\hat{\mathbf{b}}] = \mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}[\mathbf{y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$
  $= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{I}\sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$
  (Gauss-Markov: smallest variance among unbiased estimators)
- Estimator of residual variance:

  Model with (p+1) parameters:
  intercept + p regression coefficients

  $$E[\mathrm{RSS}] = E[(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})] = (n - p - 1)\,\sigma^2$$

  so that an unbiased estimator of σ^2 is: $s^2 = \frac{1}{(n-p-1)}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$

149

Bias vs. Variance

- Let $\hat{\theta}_1$ be an unbiased estimator of θ with variance equal to V,
  i.e., $E[\hat{\theta}_1] = \theta$ and $\mathrm{Var}[\hat{\theta}_1] = V$
- Suppose now an estimator given by $\hat{\theta}_2 = c \times \hat{\theta}_1$, where $0 < c < 1$,
  so that $E[\hat{\theta}_2] = c \times \theta$ (biased estimator) and $\mathrm{Var}[\hat{\theta}_2] = c^2 \times V < V$
- Which estimator, $\hat{\theta}_1$ or $\hat{\theta}_2$, is better? ($\hat{\theta}_1$ is unbiased, $\hat{\theta}_2$ has smaller variance…)
- Mean squared error (MSE): $\mathrm{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$
  $= \mathrm{Var}(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2$
  $=$ variance + squared bias

150

Problems with Least Squares

- Multicollinearity: regression coefficient estimates may change erratically in response to small changes in the model or the data
- Under extreme multicollinearity, parameters may be not estimable
- Prediction accuracy: unbiased but large variance
- Modelling alternatives: Some sort of regularization technique

151

Regularization Techniques

- Variable Selection: Best subset regression, Stepwise regression (forward, backward, hybrid)
- Dimension Reduction: Principal Component Regression, Partial Least Squares
- Shrinkage Estimation: Ridge Regression, LASSO (variable selection and shrinkage simultaneously), Elastic Net

152

Dimension Reduction

- Stepwise Regression:

Intercept only: $y = b_0 + e$

Forward: start with an intercept model and add predictors based on some model selection criteria

Backward: start with a full model and remove predictors based on some model selection criteria

Full model: $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e$

- Common model selection/comparison criteria: AIC, BIC, LRT, etc.

153

Dimension Reduction

- Least Squares: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \rightarrow \hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \rightarrow \hat{\mathbf{y}} = \mathbf{X}_{new}\hat{\mathbf{b}}$

- Principal Component (PC) Regression:

1. Use singular value decomposition (SVD) to form new latent vectors (PCs) associated with a low-rank approximation of $\mathbf{X}$

$$\overset{(n\times p)}{\mathbf{X}} = \underbrace{\overset{(n\times p)(p\times p)}{\mathbf{U}\mathbf{D}}}_{\mathbf{T}}\underbrace{\mathbf{V}^T}_{\mathbf{P}^T}$$

$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$

$\mathbf{D}$: diagonal matrix of singular values in descending order ($d_1 \geq d_2 \geq \cdots \geq d_p$)

Columns of $\mathbf{T}$: "principal components" (factor scores, latent variables)

Columns of $\mathbf{V}$: "loadings"

154

- PC Regression (Cont'ed):

2. Form a low-rank approximation of $\mathbf{X}$ by keeping just the first $k < p$ PCs (the ones associated with the k largest singular values): $\mathbf{X} \approx \mathbf{T}_k\mathbf{P}_k^T$

3. Regress $\mathbf{y}$ on this lower-dimensional feature space using the PCs as the new features: $\mathbf{y} = \mathbf{T}_k\mathbf{c} + \boldsymbol{\varepsilon} \rightarrow \hat{\mathbf{c}} = (\mathbf{T}_k^T\mathbf{T}_k)^{-1}\mathbf{T}_k^T\mathbf{y}$

   Notice: The columns of $\mathbf{T}$ are orthogonal to each other ($\mathbf{T} = \mathbf{UD}$), so $\mathbf{T}_k^T\mathbf{T}_k$ is a diagonal matrix

4. Prediction of future $\mathbf{y}$: $\mathbf{X}_{new} \rightarrow \hat{\mathbf{y}} = \mathbf{X}_{new}\mathbf{P}_k\hat{\mathbf{c}}$

   Notice: As $\mathbf{X} = \mathbf{T}_k\mathbf{P}_k^T \rightarrow \mathbf{XP}_k = \mathbf{T}_k\mathbf{P}_k^T\mathbf{P} = \mathbf{T}_k$

155

- Partial Least Squares:

- PC Regression: $\mathbf{X} = \mathbf{UDV}^T = \mathbf{TP}^T$; $\mathbf{T} = \mathbf{XP}$ (columns of T are the PCs)

  Note that vectors in $\mathbf{P}$ are eigenvectors of $\mathbf{X}^T\mathbf{X}$; $\mathbf{X}^T\mathbf{X} = \mathbf{VDU}^T\mathbf{UDV}^T = \mathbf{VD}^2\mathbf{V}^T$

- If columns of $\mathbf{X}$ are centered on zero, then $\mathbf{X}^T\mathbf{X}$ is proportional to the sample covariance matrix
- Thus, the first k PCs maximize the ability to describe the covariance or spread of the data in $\mathbf{X}$
- Problem: Rotation and data reduction to explain variation in $\mathbf{X}$ does not guarantee to yield latent features that are good for predicting $\mathbf{y}$
- Solution: Projection of latent variables to maximize the covariance between $\mathbf{X}$ and $\mathbf{y}$. For example, for the first latent vector, search for a vector $\mathbf{t} = \mathbf{Xw}$ that maximizes $\mathrm{Cov}(\mathbf{Xw}, \mathbf{y})$ subject to $\mathbf{w}^T\mathbf{w} = 1$

156

Shrinkage Estimation

Complexity parameter ($\lambda > 0$)

- Ridge Regression: $\hat{\mathbf{b}}^{\text{ridge}} = \arg\min\left\{(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb}) + \lambda\sum_{j=1}^{p} b_j^2\right\}$

Or, equivalently: $\hat{\mathbf{b}}^{\text{ridge}} = \arg\min\{(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb})\}$, subject to $\sum_{j=1}^{p} b_j^2 \le s$

→ "*squared magnitude*" of coefficients added as a penalty term

$$\begin{cases} \hat{b}_0 = \bar{y} = \frac{1}{n}\sum y_i \\ \text{After centering } y_i \text{ and } x_i's \text{ (i.e. } y_i - \bar{y} \text{ and } x_i - \bar{x}) \end{cases}$$

$$\mathrm{RSS}(\lambda) = (\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb}) + \lambda\mathbf{b}^{\mathrm{T}}\mathbf{b} \;\rightarrow\; \hat{\mathbf{b}}^{\text{ridge}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

157

Shrinkage Estimation

- LASSO: least absolute shrinkage and selection operator

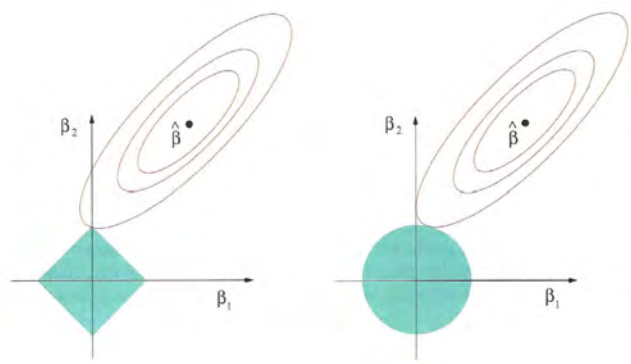
$$\hat{\mathbf{b}}^{\text{lasso}} = \arg\min\{(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb})\}, \text{ subject to } \sum_{j=1}^{p} |b_j| \le t$$

→ "*absolute value of magnitude*" of coefficients added as a penalty term

- Advantages: Lasso shrinks the less important features' coefficient to zero (i.e. feature selection)
- Disadvantages: In "large p, small n" situations (i.e. high-dimensional data with few examples), LASSO selects at most n variables before it saturates. If there is a group of highly correlated variables, then the LASSO tends to select one variable from the group and ignore the others

158

Representation of lasso (left) and ridge regression (right) estimation

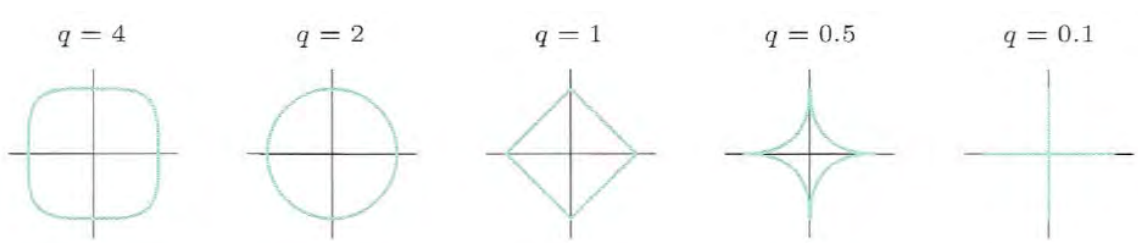


The solid green areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ (lasso) and $\beta_1^2 + \beta_2^2 \leq t^2$ (ridge regression), while the red ellipses are the contours of the least squares error function.

159

Shrinkage Estimators: Generalization

$$\hat{\mathbf{b}} = \arg\min \left\{ (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) + \lambda \sum_{j=1}^{p} |b_j|^q \right\}, q \geq 0$$



Contours of constant value of $\sum_{j=1}^{p} |b_j|^q$ for given values of q.

160

Shrinkage Estimation

- Elastic Net Regression:

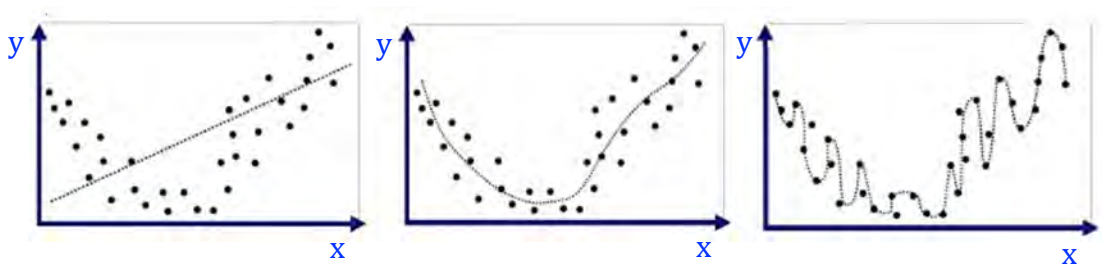
$$\hat{\mathbf{b}}^{\text{elast}} = \arg\min \left\{ (\mathbf{y} - \mathbf{Xb})^{\mathrm{T}}(\mathbf{y} - \mathbf{Xb}) + \lambda_1 \sum_{j=1}^{p} |b_j| + \lambda_2 \sum_{j=1}^{p} b_j^2 \right\}$$

$$\begin{cases} \lambda_1 = 0 \rightarrow \text{ridge (i.e. } L_2 \text{ regularization only)} \\ \lambda_2 = 0 \rightarrow \text{lasso (i.e. } L_1 \text{ regularization only)} \\ \lambda_1 > 0 \text{ and } \lambda_2 > 0 \rightarrow \text{both } L_1 \text{ and } L_2 \text{ regularization} \end{cases}$$

161

Model Selection

- Goodness-of-Fit vs. Model Complexity



Over-reduction Good Fit Overfitting

☞ Bias-variance tradeoff

162

Example: Prediction of Cattle Grazing Activities

- Wearable sensors have been explored as an alternative for real-time monitoring of cattle feeding behavior in grazing systems.
- Goal to evaluate the the effect of different cross-validation strategies on the prediction of grazing activities in cattle using wearable sensor (accelerometer) data and ML algorithms.

Ribeiro, L. A. C., Bresolin, T., Rosa, G. J. M., Casagrande, D. R., Danes, M. A. C. and Dórea, J. R. R. (2021) Nonlinear modeling to describe the pattern of 15 milk protein and nonprotein compounds over lactation in dairy cows. *Journal of Animal Science* 99(9): 1–8.

163

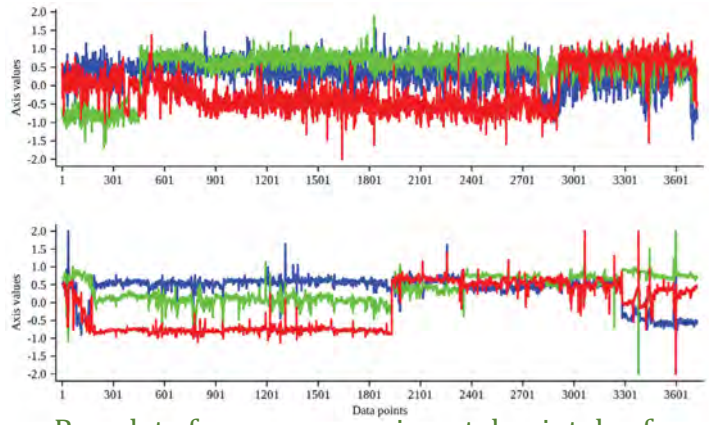
Material and Methods

- Six steers (average live weight of 345 $\pm$ 21 kg) had their behavior visually classified as grazing or not-grazing for a period of 15 d.
- Elastic Net Generalized Linear Model (GLM), Random Forest (RF), and Artificial Neural Network (ANN) were employed to predict grazing activity (grazing or not-grazing) using 3-axis accelerometer data.
- Three CV strategies were evaluated: holdout, leave-one-animal-out (LOAO), and leave-one-day-out (LODO), all with similar dataset sizes (n ~ 57,000).

164

Accelerometer

- 3-axes (X, Y, and Z) wireless accelerometer sensor was attached to the halter on the back of each animal's neck.
- The X, Y, and Z axes indicate longitudinal (front-to-back), horizontal (side-to-side), and vertical (up-to-down) head movements, respectively.



Raw data from one experimental point day for grazing (top) or not-grazing (bottom) behavior categories. The X, Y, and Z accelerometer axis values (g-force) are represented in blue, green, and red colors, respectively.

165

Results

Table 4. Validation of the machine learning approaches to predict grazing or not-grazing behavior categories visually observed in Nellore cattle using different validations strategies

Method[1]	Accuracy	Error Rate	Sensitive	Specificity	PPV	NPV
Leave-one-animal-out						
GLM	52.01	47.99	54.64	49.77	48.08	56.31
RF	56.61	43.39	59.98	53.74	52.47	61.19
ANN	57.06	42.94	53.63	59.98	53.29	60.31
Leave-one-day-out						
GLM	48.57	51.43	26.70	67.66	41.86	51.41
RF	61.20	38.80	71.98	51.79	56.57	67.94
ANN	63.50	36.50	69.12	58.59	59.29	68.51
Holdout (20%)						
GLM	58.88	41.12	66.68	51.98	55.14	63.79
RF	76.48	23.52	78.30	74.86	73.39	79.58
ANN	74.18	25.82	75.35	73.14	71.29	77.02
Holdout (20%) – Replicates[2]						
GLM	59.20 (0.29)	40.80	67.33	52.03	55.30	64.38
RF	75.86 (0.30)	24.14	77.23	74.49	75.03	76.73
ANN	72.21 (1.64)	27.79	70.92	73.48	72.69	71.83

[1]GLM, generalized linear model; RF, random forest; ANN, artificial neural network; PPV, positive predictive values; NPV, negative predictive values.
[2]Average accuracy using 20 holdout replicates and standard deviation within parenthesis.

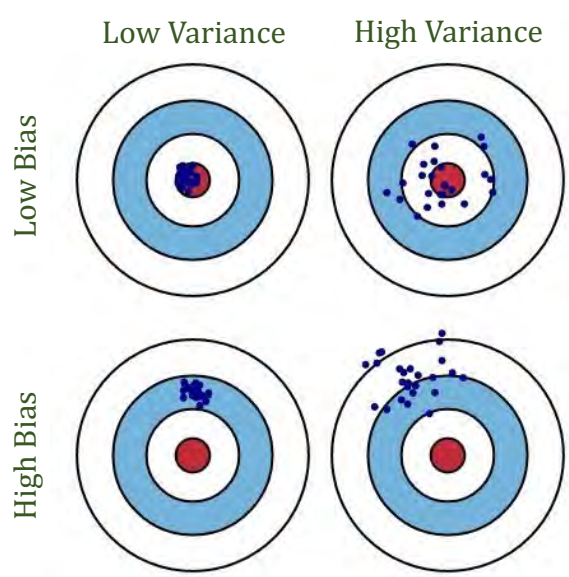
166

Results

- Overall, GLM delivered the worst prediction accuracy compared with the ML techniques, and ANN performed slightly better than RF for LOAO and LODO across CV strategies.
- The holdout yielded the highest nominal accuracy values for all three ML approaches, followed by LODO and LOAO.
- Nonetheless, the greater prediction accuracy of holdout CV may simply indicate a lack of data independence and the presence of carry-over effects from animals and grazing management.

167

Model Selection



168

Model Selection

⇨ Goodness-of-fit

- likelihood ratio approach (LRT; nested models)

$$\mathrm{LRT} = -2\times\ln\left(\frac{L_1}{L_2}\right) \sim \chi^2_{(p_1-p_2)}$$

⇨ Model complexity

- number of free parameters, p (effective number)

Linear (regularized) fitting: $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \rightarrow p = \mathrm{trace}(\mathbf{H})$

169

Effective Number of Parameters

- Example with a simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + e_i \;\rightarrow\; \mathbf{y} = \underbrace{[\mathbf{1}\quad \mathbf{x}]}_{X}\boldsymbol{\beta} + \mathbf{e} \qquad \mathbf{X'X} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} = \frac{1}{\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right]}\begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \qquad \mathbf{X}(\mathbf{X'X})^{-1} = k^{-1}\begin{bmatrix} \Sigma x_i^2 - x_1\Sigma x_i & nx_1 - \Sigma x_i \\ \Sigma x_i^2 - x_2\Sigma x_i & nx_2 - \Sigma x_i \\ \vdots & \vdots \\ \Sigma x_i^2 - x_n\Sigma x_i & nx_n - \Sigma x_i \end{bmatrix}$$

$$k^{-1} = 1/\det(\mathbf{X'X})$$

$$\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'} = k^{-1}\begin{bmatrix} \boxed{\Sigma x_i^2 - 2x_1\Sigma x_i + nx_1^2} & \Sigma x_i^2 - (x_1+x_2)\Sigma x_i + nx_1x_2 & \cdots & \Sigma x_i^2 - (x_1+x_n)\Sigma x_i + nx_1x_n \\ \Sigma x_i^2 - (x_1+x_2)\Sigma x_i + nx_1x_2 & \boxed{\Sigma x_i^2 - 2x_2\Sigma x_i + nx_2^2} & \cdots & \Sigma x_i^2 - (x_2+x_n)\Sigma x_i + nx_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma x_i^2 - (x_1+x_n)\Sigma x_i + nx_1x_n & \Sigma x_i^2 - (x_2+x_n)\Sigma x_i + nx_2x_n & \cdots & \boxed{\Sigma x_i^2 - 2x_n\Sigma x_i + nx_n^2} \end{bmatrix}$$

$$\mathrm{trace}\left[\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right] = k^{-1}\times\Sigma\left[\Sigma x_i^2 - 2x_i\Sigma x_i + nx_i^2\right] = k^{-1}\times\left[2n\Sigma x_i^2 - 2(\Sigma x_i)^2\right] = k^{-1}\times[2k] = 2$$

170

Model Selection

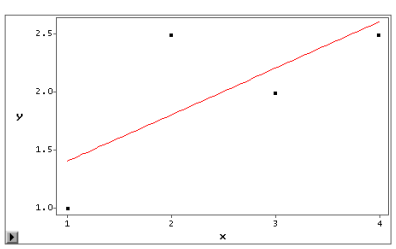
- Balancing goodness-of-fit and complexity
  - Akaike information criterion (AIC): $\mathrm{AIC} = 2p - \ln(L)$
  - Bayesian information criterion (BIC): $\mathrm{BIC} = p \times \ln(n) - 2\ln(L)$ (or Schwarz Criterion)
- If $e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ then:

$$\mathrm{AIC} = 2p + n \times \ln\left(\frac{\mathrm{RSS}}{n}\right) \text{ and } \mathrm{BIC} = \frac{1}{\sigma^2}\mathrm{RSS} + p \times \ln(L)$$

Model Selection

- Example: linear vs. quadratic regression

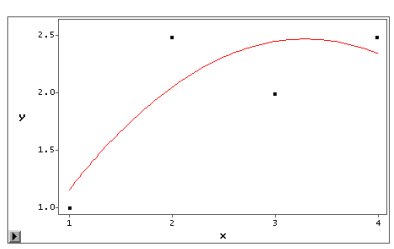
$$y_i = \beta_0 + \beta_1 x_i + e_i$$



$$\hat{y}_i = 1.0 + 0.4x_i$$

$$\begin{cases} R^2 = 0.53 \\ R^2_{adj} = 0.30 \\ \hat{\sigma}^2_e = 0.35 \end{cases}$$

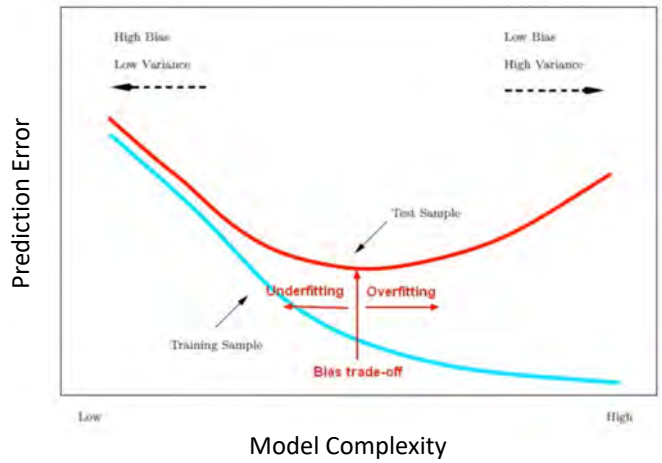
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$



$$\hat{y}_i = -0.25 + 1.65x_i - 0.25x_i^2$$

$$\begin{cases} R^2 = 0.70 \\ R^2_{adj} = 0.10 \\ \hat{\sigma}^2_e = 0.45 \end{cases}$$

Predictive Ability

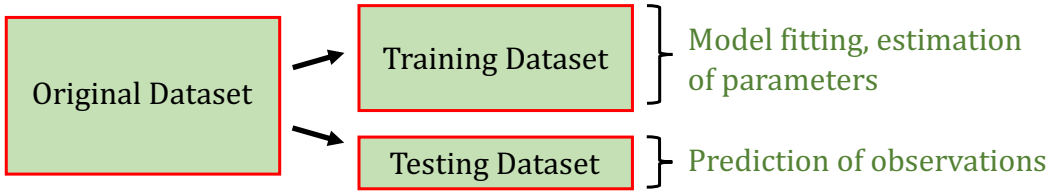


Behavior of test sample and training sample error as the model complexity is varied.

173

Cross-Validation

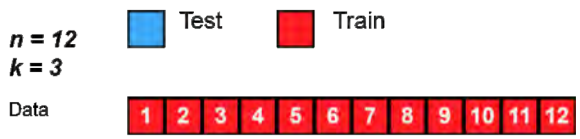
- Holdout method: In the holdout method, the data points are randomly assigned to two sets d_0 and d_1, usually called the training set and the test set, respectively.
- The size of each of the sets is arbitrary although typically the test set is smaller than the training set. The model is then trained on d_0 and tested (i.e. evaluate its performance) on d_1.



174

Cross-Validation

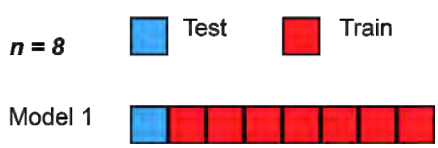
- K-Fold Cross-Validation: The original sample is randomly partitioned into K equal sized subsamples. One of the subsamples is retained as the validation data, and the remaining K - 1 subsamples are used as training data.
- The CV process is repeated K times, one for each subsample. The K results are then averaged to produce a single estimation.
- Illustration of K-fold CV when n = 12 observations and K = 3.



175

Cross-Validation

- Leave-one-out cross-validation (LOOCV): One observation is removed from the original dataset, to be used as validation. The model is trained on the remaining n-1 observations, and tested on the observation left out. The process is repeated n times, so that each observation is used once as validation. Results are averaged across the n validation data points.
- Illustration of a LOOCV when n = 8 observations. A total of 8 models will be trained and tested.

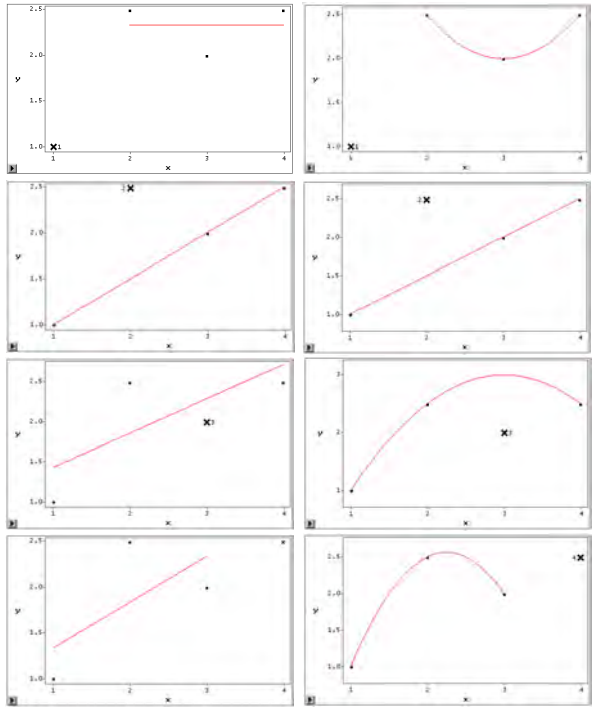


176

Leave-One-Out Cross-Validation (LOOCV)

Obs	Linear	Quadr
1	-1.333	-3.0
2	1.000	1.0
3	-0.286	-1.0
4	-0.333	3.0
PRESS	2.971	20.0

Linear



Quadratic

Predictive Quality Metrics

- Prediction of Binary Outcomes

  Metrics usually assess the frequency of two types of error: false positive (a.k.a. nuisance alarm) and false negative (a.k.a. missing alarm) errors via tables of errors, or confusion matrix:

Prediction	True Category (Ground Truth)	
	$y = 0$	$y = 1$
$\hat{y} = 0$	True Negative (TN)	False Negative (FN)
$\hat{y} = 1$	False Positive (FP)	True Positive (TP)

Example:

Y = 0 for healthy
Y = 1 for disease

Predictive Quality Metrics

Prediction	True Category (Ground Truth)	
	$y = 0$	$y = 1$
$\hat{y} = 0$	True Negative (TN)	False Negative (FN)
$\hat{y} = 1$	False Positive (FP)	True Positive (TP)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

Accuracy can be a misleading metric for imbalanced data sets. Consider a sample with 95 negative and 5 positive values. Classifying all values as negative in this case gives 0.95 accuracy score.

179

- Some other commonly used metrics:

$$\text{Precision} = \frac{TP}{TP + FP}$$

(Positive Predictive Value)

$$\text{Recall} = \frac{TP}{TP + FN}$$

(Sensitivity)

$$\text{Specificity} = \frac{TN}{TN + FP}$$

(Selectivity)

$$F_1 \text{ score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

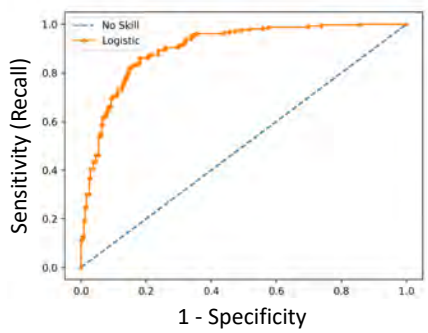
(Harmonic Mean of Recall and Precision)

$$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2}$$

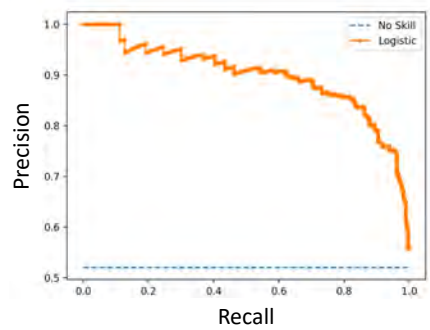
180

- Two useful plots:

Receiver Operating Characteristic Curve (ROC)



Precision-Recall Curve



Generally, ROC curves should be used when there are roughly equal numbers of observations in each class. Precision-Recall curves should be used when there is a moderate to large class imbalance.

181

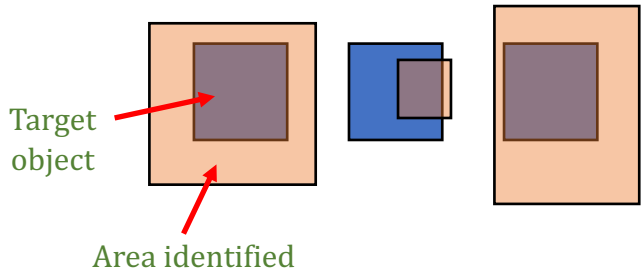
Predictive Quality Metrics

- Prediction of Continuous Outcomes
  - Predictive Correlation: $r = \mathrm{Corr}(y, \hat{y})$, or its square r^2
  - Mean Squared Error: $\mathrm{MSE} = \mathrm{mean}([y - \hat{y}]^2)$
  - Root Mean Squared Error: $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$
  - Mean Absolute Error (MAE): $\mathrm{MAE} = \mathrm{mean}(|y - \hat{y}|)$
  - Mean Absolute Scaled Error (MASE): $\mathrm{MASE} = \mathrm{mean}\left(\left|\frac{y-\hat{y}}{\mathrm{mean}(y)}\right|\right)$

182

Predictive Quality Metrics

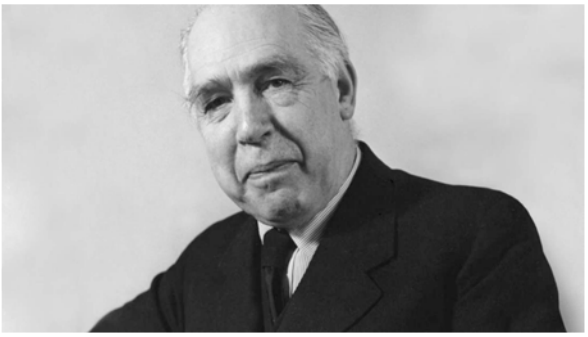
- Identification of objects within the images
  - Intersection over Union (IoU)



183

184

185



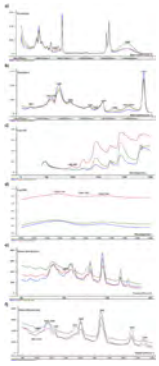
"Prediction is very difficult, especially about the future."

(Niels Bohr, 1885-1962)

186

Dairy Cow Feed Intake Prediction Using Milk MIR

milk sample



mid-infrared (MIR) spectroscopy

dry matter intake

Dorea, J. R. R., Rosa, G. J. M., Weld, K. A. and Armentano, L. E. (2018) Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science* 101: 5878-5889.

187

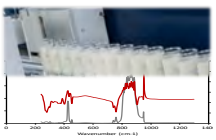
Experimental Data

- Improve intake predictions
- Hard to measure in practical conditions – Feed efficiency
  - 310 cows from 5 trials
  - 1276 observations of DMI, behavior (visit duration), milk yield, BW, milk spectra
  - Milk spectra: 1060 wavelengths

+

DairyCOMP 305

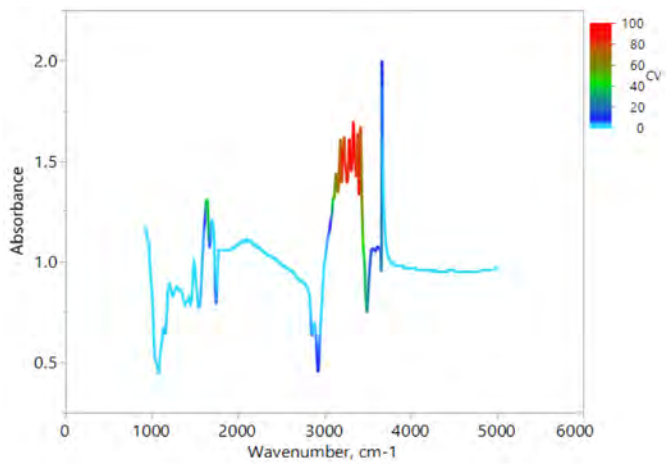
+



188

Milk Mid-infrared Spectra

- Milk spectra: 1060 wavelengths
- CV > 1%: 362 wavelengths



189

Markov Blanket

- Dimension reduction techniques
- Bayesian Network; Markov Blanket (MB):

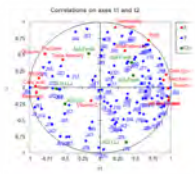
- MB of a variable X is the smallest set MB(X) containing all variables carrying information about X that cannot be obtained from any other variable
- In a DAG, this is the set of all parents, children, and spouses of X.
- Milk spectra MB: 33 wavelengths

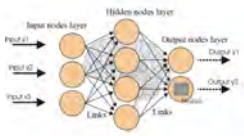
190

Data Analysis; Models

- Approaches: Partial least squares (PLS) and Artificial neural network (ANN)

1) Milk yield, BW0.75, DIM
2) Milk yield, BW0.75, DIM, and 362 WL
3) Milk yield, BW0.75, DIM, and 33 WL (MB)
4) Milk yield, BW0.75, DIM, Fat, Protein + Lactose
5) Milk yield, BW0.75, DIM, 33 WL, Visit duration
6) Milk, DIM, and 33 WL (MB)
7) 362 WL (WL)
8) 33 WL (MB)

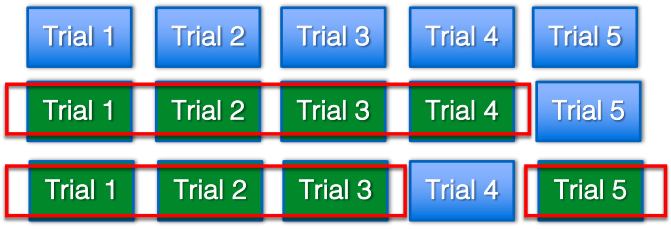




191

Data Analysis; Model Validation

- Validation: Independent datasets

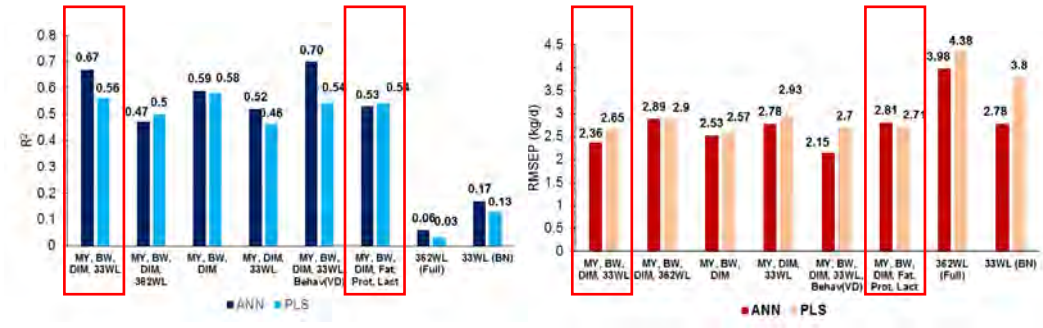


⋮

5 times

192

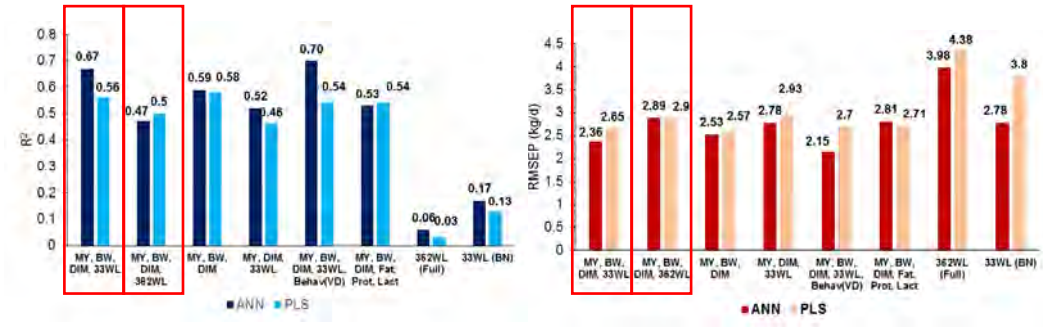
Results



- Milk components vs raw spectra: better performance with ANN

193

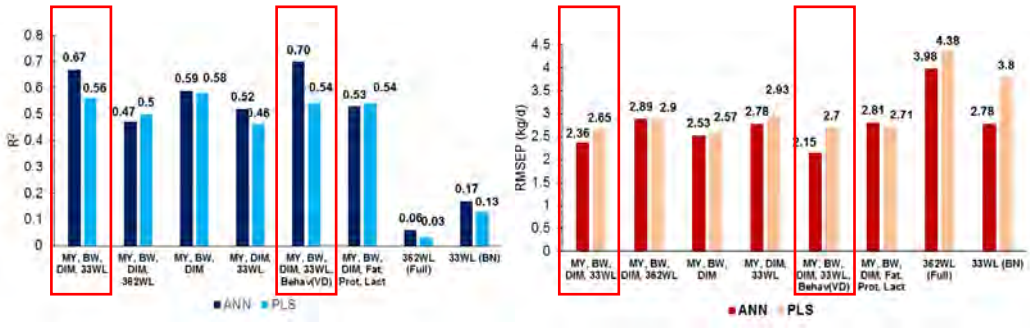
Results



- Variable selection through MB improved model performance, decreasing RMSEP

194

Results



- Model including MY + DIM + BW + Milk spectra (33 WL; BN) + Behavior (VD) presented accurate and precise predictions

195

Conclusions

- ANN on reduced WL set (with BN) improved prediction quality
- Superiority of ANN indicates potential nonlinear relationships between DMI and WL
- Superiority of models including raw spectra compared with milk components (fat, protein, and lactose) indicates that other unknown milk compounds may be important
- Validation of model predictions should be carefully conducted

196

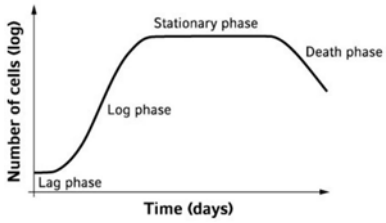
Machine Learning

- Introduction, Big Data Analytics
- Artificial Neural Networks
- Support Vector Machines
- Decision Trees
- Kernel regression, RKHS

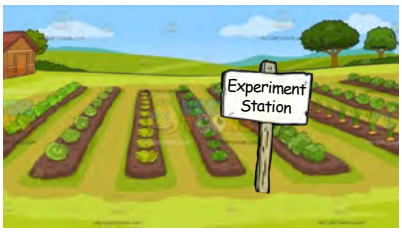
197

Statistics and Machine Learning

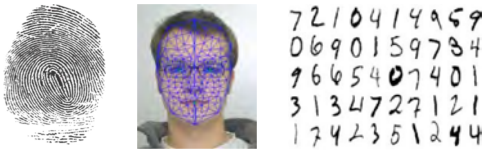
Parametric Models



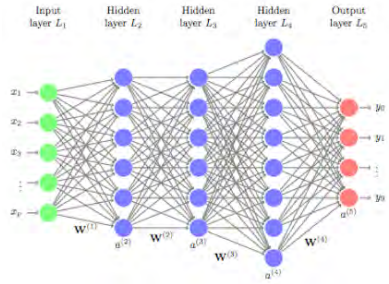
Experimental Design



Pattern Recognition Techniques



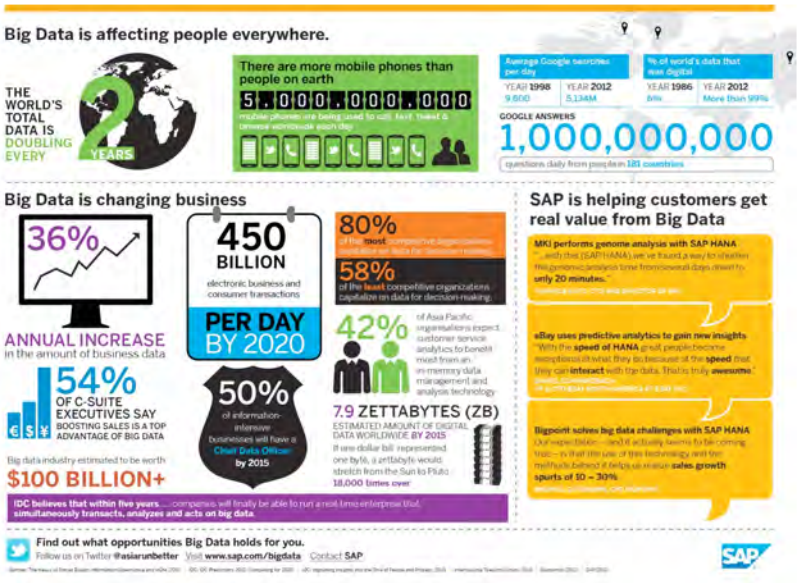
Predictive Analytics



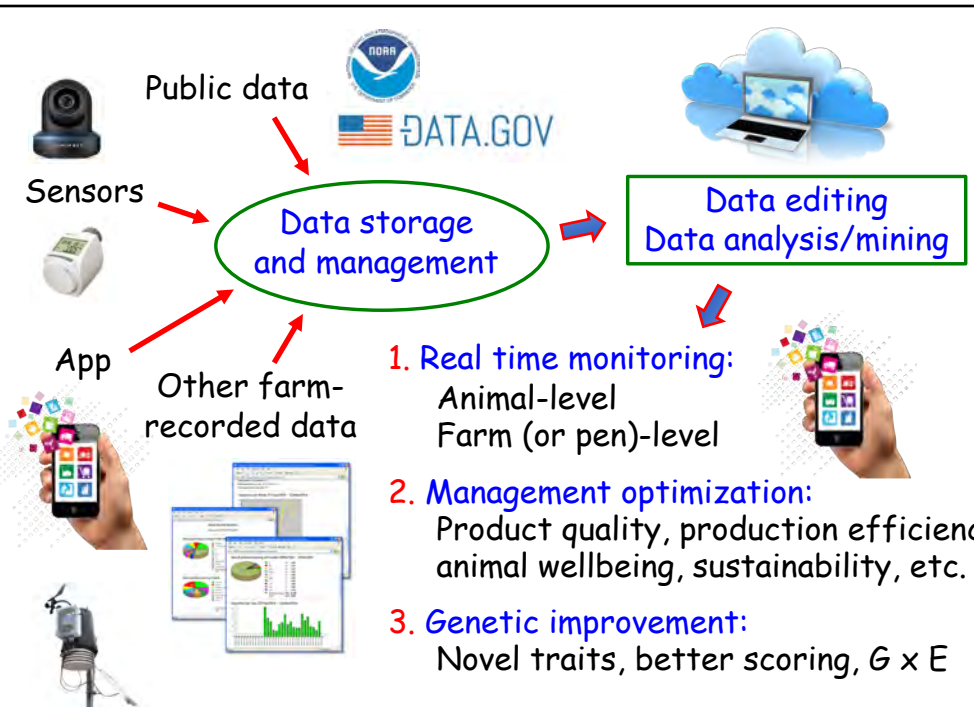
Deep Learning

198

Data, Data Everywhere

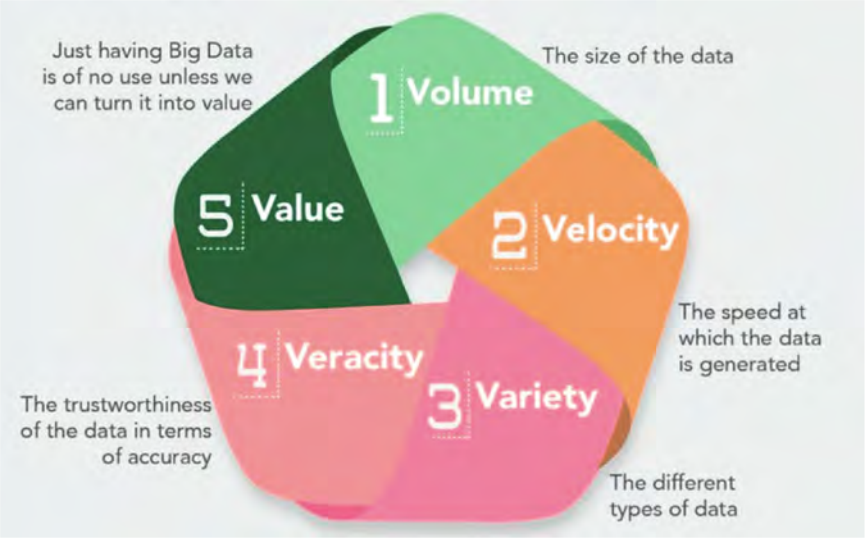


199

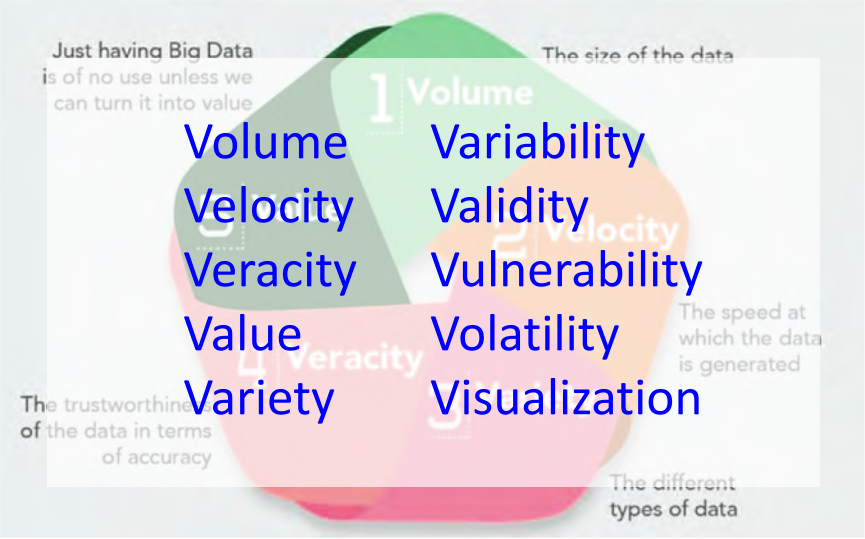


200

The 5 Vs of Big Data



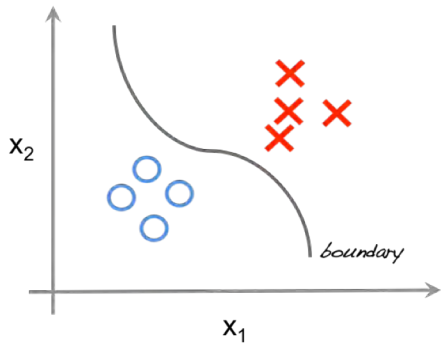
The 5 Vs of Big Data (or more...)



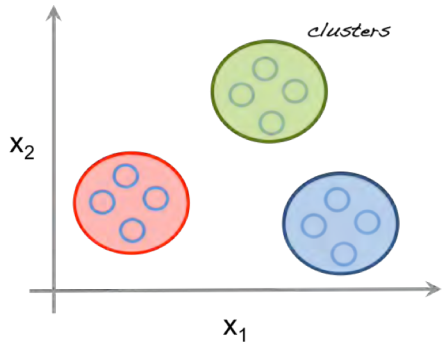
Volume	Variability
Velocity	Validity
Veracity	Vulnerability
Value	Volatility
Variety	Visualization

Supervised and Unsupervised Methods

Supervised learning



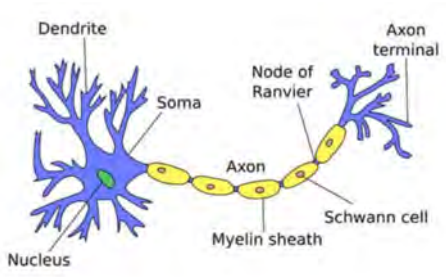
Unsupervised learning

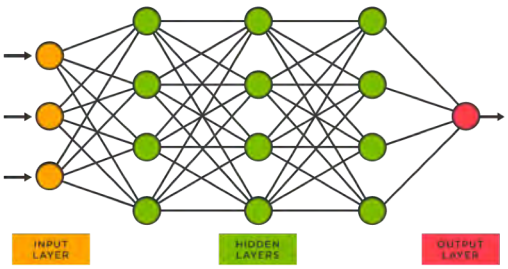


203

Artificial Neural Networks

- Nonlinear regression technique inspired on how the brain works:

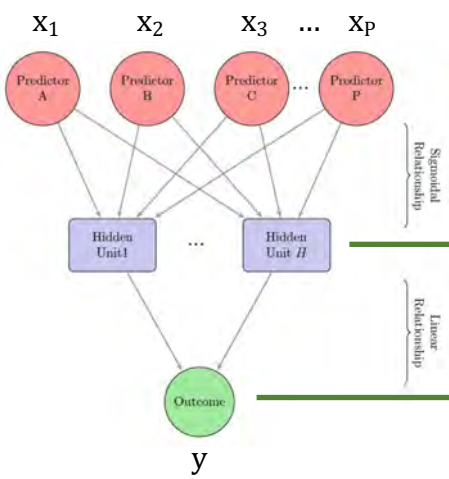




204

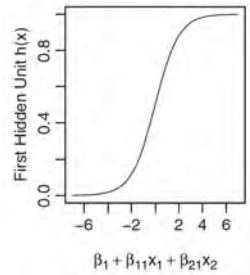
Artificial Neural Networks

- Example of NN with single hidden layer:

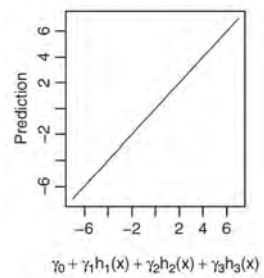


$$h_k(\mathbf{x}) = g\left(\beta_{0k} + \sum_{i=1}^{P} x_j \beta_{jk}\right)$$

Activation function: $g(u) = \dfrac{1}{1+e^{-u}}$
(e.g. sigmoidal)



$$f(\mathbf{x}) = \gamma_0 + \sum_{k=1}^{H} \gamma_k h_k$$



205

Artificial Neural Networks

- P predictors → H(P + 1) + (H + 1) parameters
- Parameters usually initialized to random values, and then specialized algorithms (e.g. back-propagation) are used to minimize the sum of squares of residuals
- NN tend to over-fit → strategies to avoid over-fitting include 'early stopping', and 'weight decay' (regularization similar to ridge regression)

Optimization using: $\displaystyle\sum_{i=1}^{n} (y_i - f_i(x))^2 + \lambda \sum_{k=1}^{H} \sum_{j=0}^{P} \beta_{jk}^2 + \lambda \sum_{k=0}^{H} \gamma_k^2$

(predictors should be on the same scale → $x_j^* = \dfrac{x_j - x_j}{s_j}$)

206

Artificial Neural Networks

- NN depicted before refers to a single-layer feed-forward network (Perceptron)
- Variations include multiple hidden layers, loops going both directions between layers, Bayesian approach, etc.
- Choice of NN architecture includes number of hidden units per layer, activation function (linear, sigmoid, hyperbolic tangent – Tanh, Rectified Linear Unit – ReLU, etc.)
- Model fitting strategies: average results of multiple NN with different starting values, pre-filter predictors with strong collinearity

207

Artificial Neural Networks

- Example: Cross-validated RMSE profiles for single hidden layer NN with sizes ranging between 1 and 13 hidden units, and three different weight decay values ($\lambda = 0.00, 0.01, 0.10$)

208

Artificial Neural Networks

- Response variable y (predictand): single or multiple outputs; continuous, binary, or multi-category variable (C classes → C binary columns of dummy variables)
- For classification, an additional nonlinear transformation is used on the combination of hidden unites, for example the *softmax* transformation:

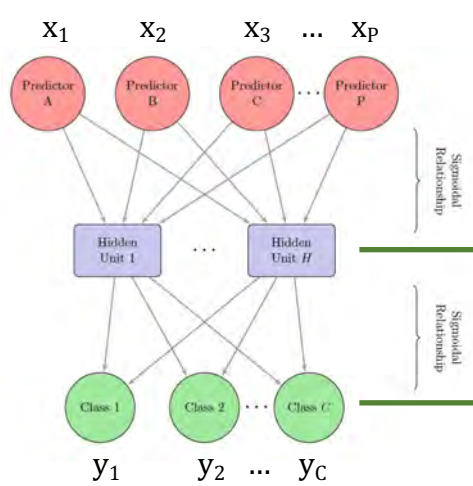
$$f^*_{i\ell}(x) = \frac{e^{f_{i\ell}(x)}}{\sum_l e^{f_{il}(x)}}$$

where $f_{il}(x)$ is the model prediction of the lth class and the ith sample

209

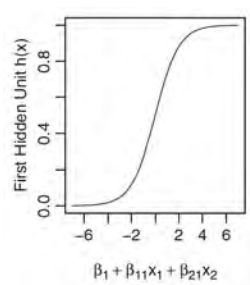
Artificial Neural Networks

- Example of NN for classification:

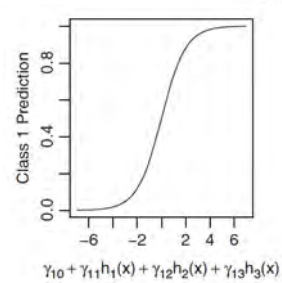


$$h_k(\mathbf{x}) = g\left(\beta_{0k} + \sum_{i=1}^{P} x_j \beta_{jk}\right)$$

Activation function: $g(u) = \dfrac{1}{1+e^{-u}}$
(e.g. sigmoidal)



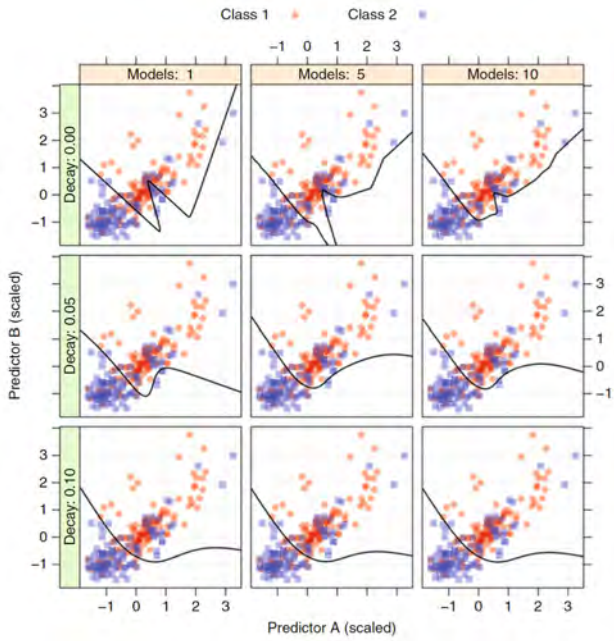
$$f^*_{i\ell}(x) = \frac{e^{f_{i\ell}(x)}}{\sum_l e^{f_{il}(x)}}$$



210

Artificial Neural Networks

- Example: Illustration of model averaging effect with different amounts of eight decay; models included three hidden units

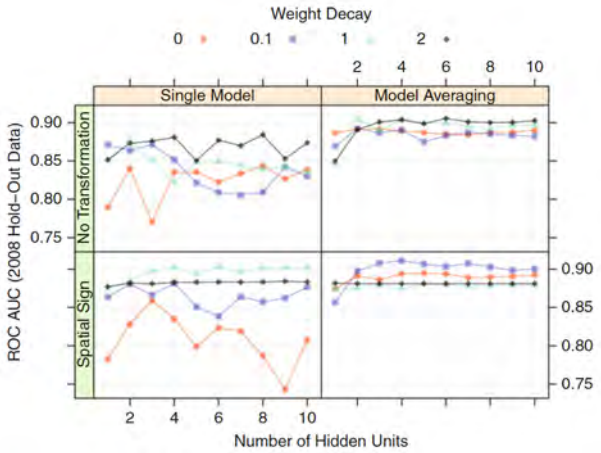


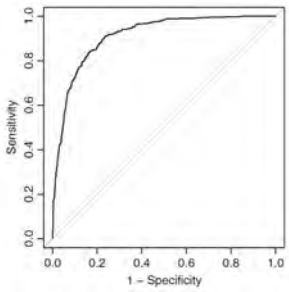
(Kuhn and Johnson, 2016)

211

Artificial Neural Networks

- Example: Effect of data transformation (spatial sign transformation) and model averaging on tuning parameter profiles



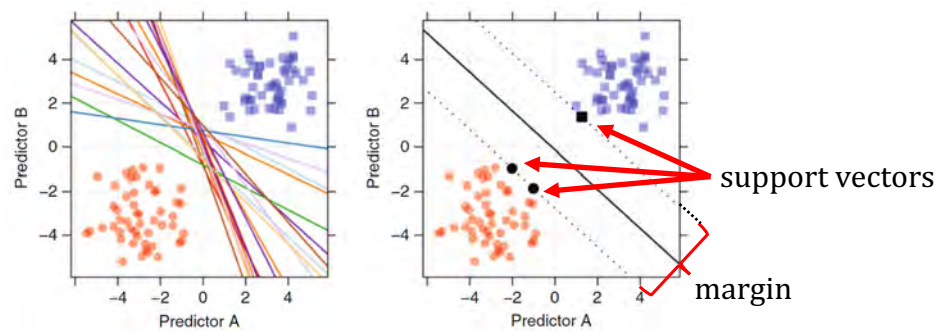


Area under the ROC curve for a model averaged network with the spatial sign transformation

212

Support Vector Machine

- Linear classification boundary, maximum margin classifier



Left: A data set with completely separable classes. An infinite number of linear class boundaries would produce zero errors. Right: The class boundary associated with the linear maximum margin classifier.

213

Support Vector Machine

- Let two outcome classes (A and B) coded as $y = -1$ and $y = +1$, and predictors $\mathbf{x}_i = (x_{i1} + x_{i2} + \cdots + x_{iP})^T$
- $D(\mathbf{x})$: decision value; if $D(\mathbf{x}) > 0$ → class A, otherwise class B
- New sample: $\mathbf{u}$ → $D(\mathbf{u}) = \beta_0 + \sum_{j=1}^{P} \beta_j u_j$

→ $D(\mathbf{u}) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i^T \mathbf{u}$ (written as a function of the data)

Notice: Due to the dot product , predictors should be centered and scaled, i.e. $x_j^* = \frac{x_j - \bar{x}_j}{s_j}$)

214

Support Vector Machine

- Completely Separable Classes:

$$D(\mathbf{u}) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i^T \mathbf{u}$$, with:

$$\begin{cases} \alpha_i = 0 \text{ , for samples not on the margin} \\ \alpha_i > 0 \text{ , for support vectors} \end{cases}$$

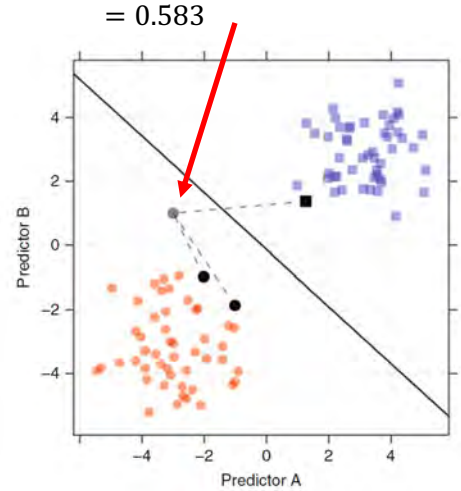
(Support vector: black points in the Figure)

	True class	Dot product	y_i	α_i	Product
SV 1	Class 2	−2.4	−1	1.00	2.40
SV 2	Class 1	5.1	1	0.34	1.72
SV 3	Class 1	1.2	1	0.66	0.79

($\beta_0 = -4.372$)

new sample

$D(\mathbf{u}) = -4.372 + 2.4 + 1.72 + 0.79$
$= 0.583$



215

Support Vector Machine

- Not Completely Separable Classes: new formulation with a cost on the sum of the training set points that are on the boundary or on the wrong side of the boundary
- Nonlinear Classification Boundaries: "Kernel Trick"

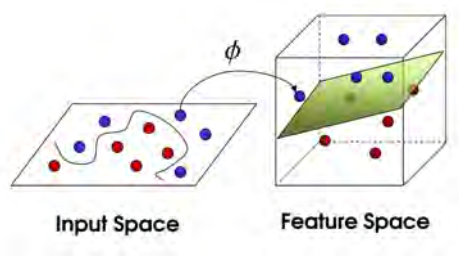
$$\rightarrow D(\mathbf{u}) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{u})$$

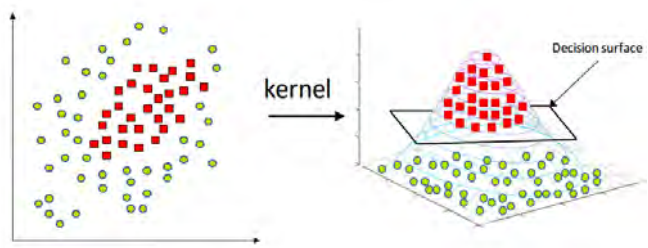
- Kernel Function:
  - Linear: $K(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T \mathbf{u}$
  - Polynomial: $K(\mathbf{x}, \mathbf{u}) = (\text{scale}(\mathbf{x}^T \mathbf{u}) + \mathbf{1})^{\text{degree}}$
  - Radial basis function: $K(\mathbf{x}, \mathbf{u}) = \exp(-\sigma \|\mathbf{x} - \mathbf{u}\|^2)$
  - Hyperbolic tangent: $K(\mathbf{x}, \mathbf{u}) = \tanh(\text{scale}(\mathbf{x}^T \mathbf{u}) + 1)$

216

Support Vector Machine

- Kernel Trick:





217

Support Vector Machine

- The choice of the Kernel function parameters and the cost value should be tuned to avoid over-fitting
- Other extensions: multiple classes, estimation of class probabilities, specialized Kernels, etc.
- SVM originally developed for classification, later extended to regression (support vector regression)

218

Support Vector Machine

- Support Vector Regression
- Common technique: ϵ-sensitive regression (robust regression)

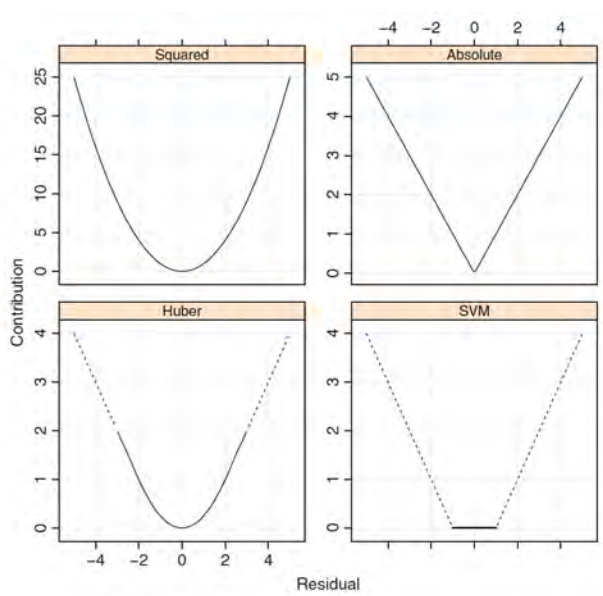
SSE: $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (sensitive to outliers)

Huber function: $\sum_{\hat{\varepsilon} \le \epsilon} (y_i - \hat{y}_i)^2 + \sum_{\hat{\varepsilon} > \epsilon} |y_i - \hat{y}_i|$

Support Vector Regression: $\sum_{\hat{\varepsilon} > \epsilon} |y_i - \hat{y}_i|$ (only look at outliers…)

219

Support Vector Machine



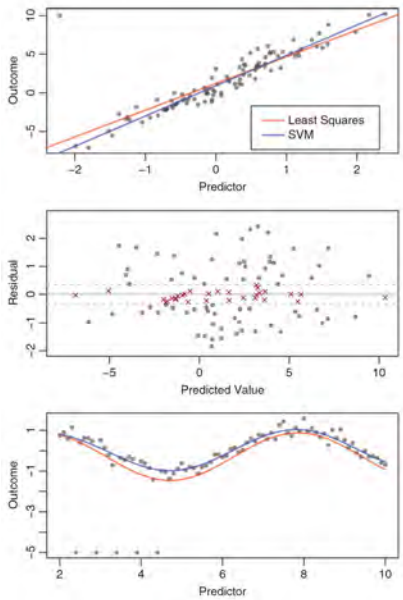
The relationship between a model residual and its contribution to the regression line for several techniques.

For the Huber approach, a threshold of 2 was used while for the support vector machine, a value of $\epsilon = 1$ was used.

(Kuhn and Johnson, 2016)

220

Support Vector Machine



The robustness qualities of SVM models

Top: a small simulated data set with a single large outlier is used to show the difference between an ordinary regression line (red) and the linear SVM model (blue)

Middle: the SVM residuals versus the predicted values (the upper end of the y-axis scale was reduced to make the plot more readable). The plot symbols indicate the support vectors (shown as grey colored circles) and the other samples (red crosses). The horizontal lines are $\pm\epsilon = 0.01$

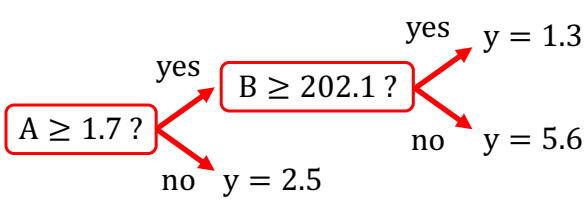
Bottom: A simulated sin wave with several outliers. The red line is an ordinary regression line (intercept and a term for sin(x)) and the blue line is a radial basis function SVM model

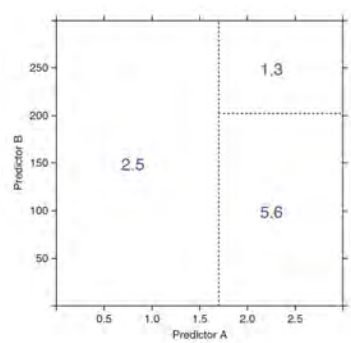
221

Decision Trees

- Tree-based Models: consist of one or more nested *if-then* statements

if $A \geq 1.7$ then
  if $B \geq 202.1$ then $y = 1.3$
  else $y = 5.6$
else $y = 2.5$





Example of the predicted values within regions defined by a tree-based model

222

Decision Trees

- Basic Regression/Classification Trees:
- Partition the data into smaller, more homogeneous groups in terms of the response y, by determining:
  - the predictor to split on and value of the split
  - the depth (or complexity) of the tree
  - the prediction equation in the terminal nodes
- There are many algorithms for constructing regression/classification trees, for example the Classification and Regression Tree (CART)

223

Decision Trees

- CART starts with the entire data set S, and finds the predictor and split value that partition the data into two groups (S_1 and S_2) such that SSE is minimized:

$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

- Then, within each sub set, the method proceeds with additional partitions
- In classification, the partition seeks more 'pure' sets, i.e. sets containing a larger proportion of one class in each node; measures such as Gini index and cross entropy are generally used; $\mathrm{Gini} = p_1(1 - p_1) + p_2(1 - p_2)$

224

Decision Trees

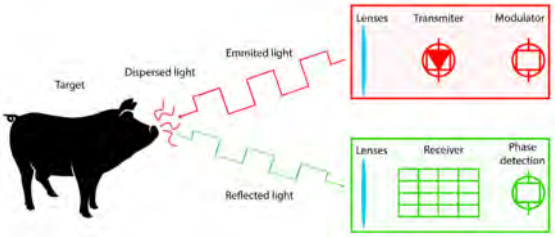
- Bagging (bootstrap aggregation):
  - Generate m bootstrap samples
  - Construct a tree model for each bootstrap
  - Average m prediction for any new sample
- Drawback of bagging: 'tree correlation'
- Random Forests: similar approach to bagging, but trees constructed for each bootstrap sample use $k < P$ randomly selected of the original predictors
- Boosting: ensemble of weak classifiers, which are trained by increasing weights of incorrectly classified samples at each iteration. Algorithms include AdaBoost, and Stochastic Gradient Boosting

225

Data Streaming Example: Computer Vision Systems

226

Depth Sensors (3D Cameras)



Time of Flight
(ToF)

Light Detection and Ranging
(LiDAR)

227

Real-Time Monitoring: Growth and Development in Pigs

➢ Periodic measurements:

- Direct assessment of animals growth
  - Assess intra-group variability
  - Optimal management (e.g. precision nutrition)
- Prohibitive
  - Labor and cost
  - Animal welfare (stress)
  - Scale within pen: expensive, requires periodically cleaning and calibration

228

Prediction of Pig Weight

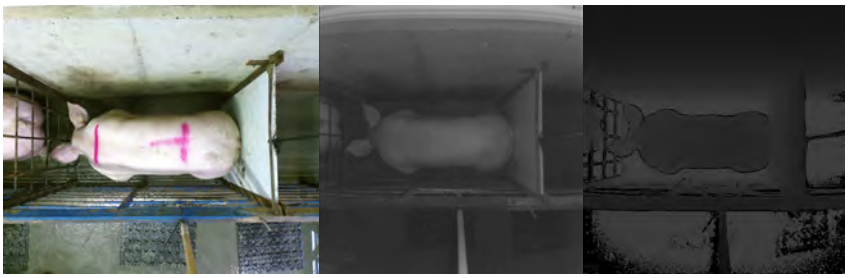
- Data on 655 pigs
- Boars and gilts from three commercial lines
- Weight across different ages (Scale EziWeigh5i, ste ±1%)
- Pigs were not fasting

Fernandes AFA, Dórea JRR, Fitzgerald R, Herring W and Rosa GJM (2019) A novel automated system to acquire biometric and morphological measurements, and predict body weight of pigs via 3D computer vision. *J. Anim. Sci.* 97: 496–508.

229

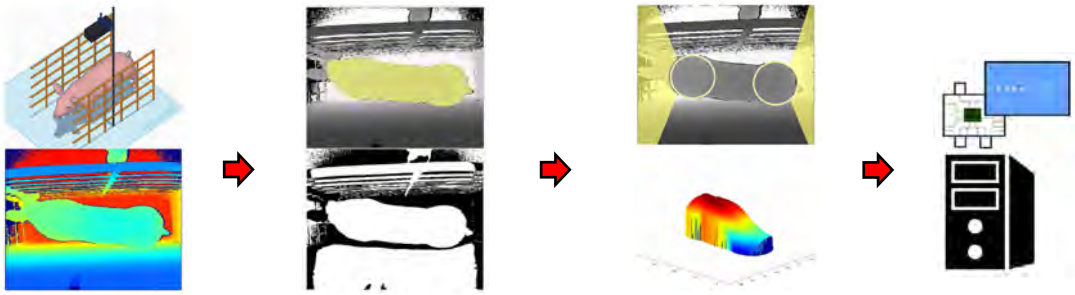
Data Acquisition

- Sensor positioned on top of the area before to the scale
- Pigs were contained under the sensor for a variable amount time
- Kinect V2 sensor (Microsoft)
- BW and multiple images acquired from each animal



230

Example of a Computer Vision System Framework



A. Image acquisition

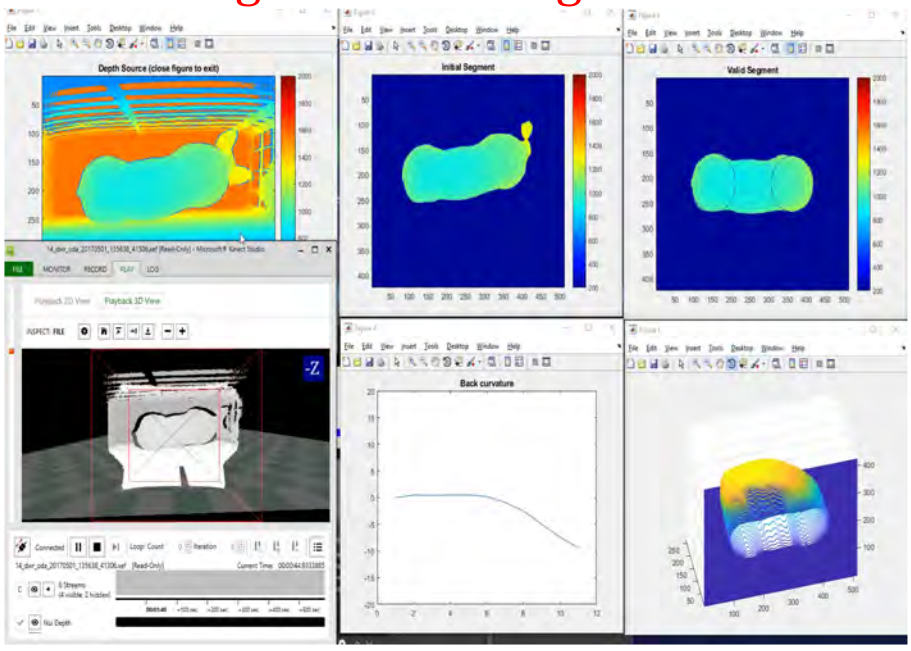
B. Image Analysis
- Thresholding
- Binarization

C. Image Processing
- Image segmentation
- Feature extraction

D. Data Analysis
- Data normalization
- Model fitting
- Validation and tuning
- Prediction

231

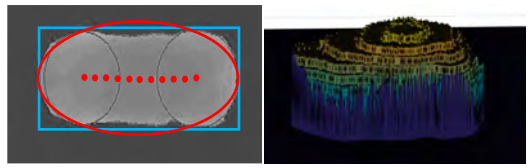
Segmentation Algorithm

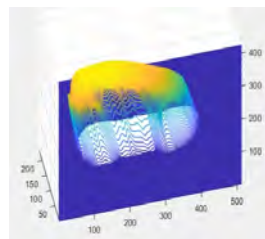


232

Features Extracted

- Feature extraction:
  - Body measurements:
    - Area
    - Volume
    - Length
    - Width
    - Height
  - Shape descriptors:
    - Eccentricity
    - Back curvature linear coefficient
    - Polar Fourier Descriptors



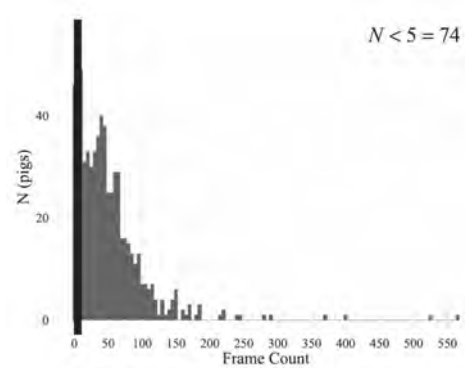


(MATLAB, Release 2017b)

233

Image Selection

- Variables from a random image
- Image with max area
- Image with max length
- Image with max volume
- Average across all images
- Median across all images
- Truncated average removing 20% of data for each animal
- Truncated average of the subset on 3rd quartile



234

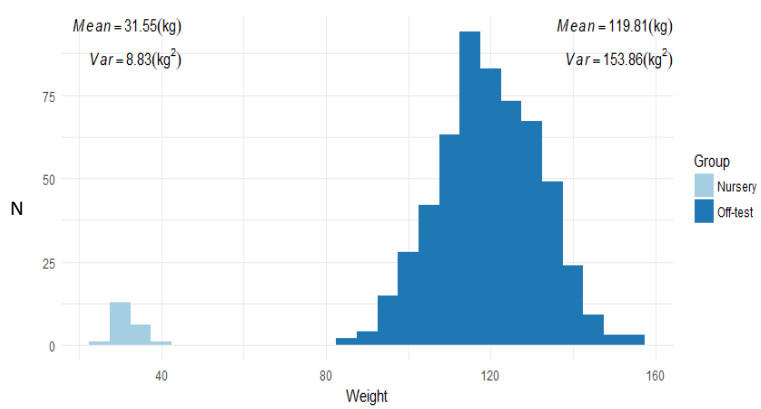
Statistical Analyses

Linear model:

- For all the reduced datasets 10 permutations on a 5-fold cross-validation were used to assess the quality of the predictions
- Stepwise regression with AIC as model selection criterion (*stepAIC* function of MASS package)
- R environment

235

Results

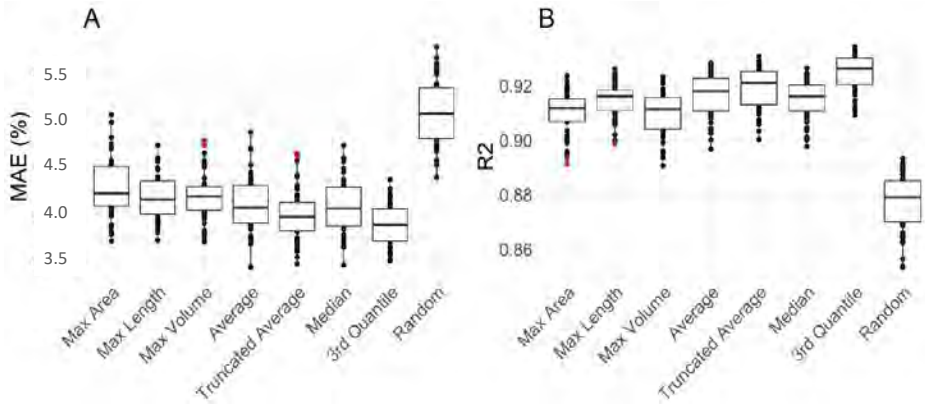


Histogram of live body weight (kg) distribution for nursery and off-test pigs with relative means and variation

236

Results

- Analysis including nursery data

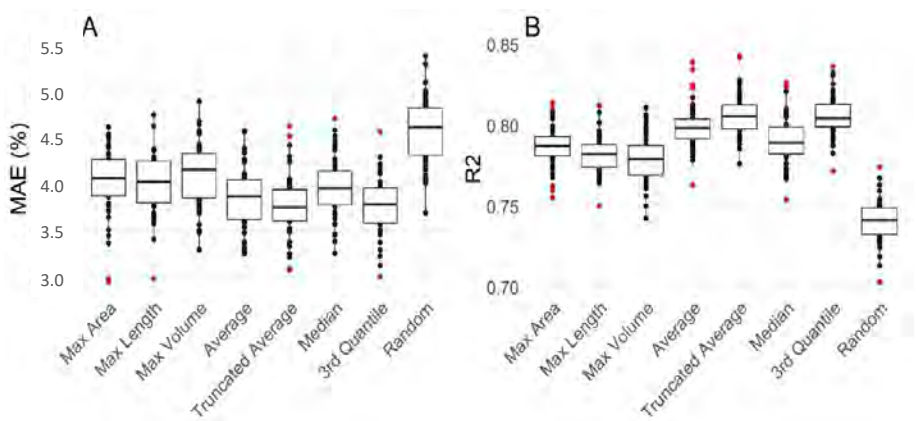


A) Box plots for Mean absolute error (MAE) as percentage of the average body weight. B) Coefficient of determination (R^2) of the different models on the test data across the cross validation.

237

Results

- Analysis without nursery data



A) Box plots for Mean absolute error (MAE) as percentage of the average body weight. B) Coefficient of determination (R^2) of the different models on the test data across the cross validation.

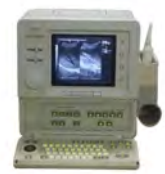
238

Improving Prediction of Pig Body Weight and Body Composition

- Body weight
- Body composition traits:

  Muscle depth (MD) and back fat (BF)

Trait	Mean	Standard deviation
BF, mm	6.03	1.47
MD, mm	65.07	6.19
BW, kg	119.97	12.43



Aloka SSD 500

Fernandes AFA, Dórea JRR, Valente BD, Fitzgerald R, Herring W and Rosa GJM (2020) Comparison of data analytics strategies in computer vision systems to predict pig body composition traits from 3D images. *Journal of Animal Science* 98:skaa250.

239

Data Mining Approaches

Prediction models:

- Multiple Linear Regression (LM)
- Partial Least Squares (PLS)
- Elastic Network Regression (EN)
- Artificial Neural Network (ANN)

} Input: Image features (MASS, pls, glmnet, H2O)

- Deep Learning Image Encoder (DL) } Input: Raw 3D images

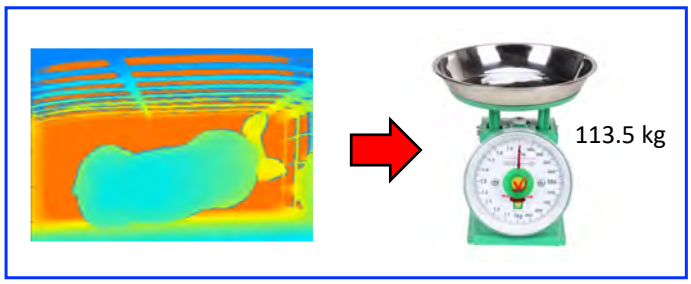
NN architectures: 1-3 hidden layers, 5-100 nodes/layer, activation functions: rectified linear unit (ReLU) or max-out, dropout rate 20-80%, loss functions: Gaussian and Huber, L1 and L2 regularizations, learning rate and time decay

Model comparison:

- 5-fold CV: mean absolute error (MAE), mean absolute scaled error (MASE), root mean square error (RMSE), R^2

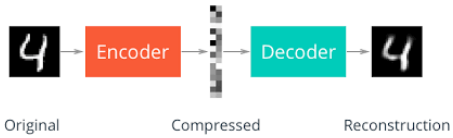
240

Deep Learning Image Encoder



241

Deep Learning Image Encoder



- TensorFlow machine learning library; Python (version 3.7)
- Network architectures: input layer, encoder blocks, fully connected layers, and output layer
- Input layer: 3D image and camera focal length
- Encoder blocks: convolutional block, followed by a max-pooling layer with a 2 by 2 window and a strider of the same size
- Convolutional blocks: convolutional layer with a 3 by 3 window, batch normalization layer, and ReLU activation function layer
- Fully connected layers had L1 and L2 regularization, dropout rate of 50%, and leaky ReLU activation function
- DL architectures varied on size of the input image, number of encoder blocks, and number of nodes on the fully connected layers

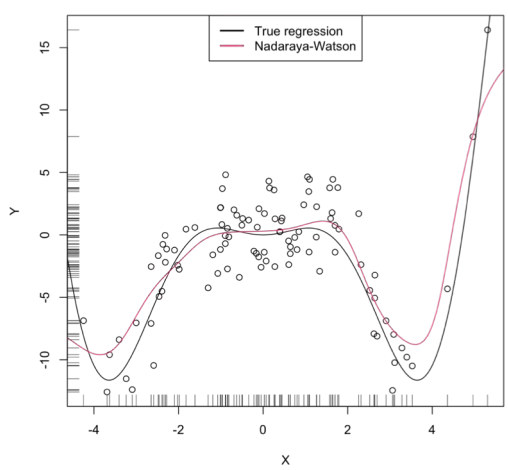
242

Predictive Performance of each Model for each Trait

Trait	Model	MAE	MASE	RMSE	R^2
BW, kg	LM	4.81	4.00	6.61	0.73
	PLS	4.62	3.85	6.48	0.74
	EN	4.55	3.79	6.39	0.75
	ANN	5.00	4.16	6.83	0.70
	DL	**3.26**	**2.69**	**4.56**	**0.86**
MD, mm	LM	4.10	6.30	5.16	0.35
	PLS	4.36	6.67	5.37	0.30
	EN	4.12	6.32	5.12	0.31
	ANN	4.61	7.07	5.77	0.21
	DL	**3.28**	**5.02**	**4.34**	**0.50**
BF, mm	LM	1.15	18.83	1.43	0.12
	PLS	1.13	18.95	1.40	0.10
	EN	1.08	18.00	1.35	0.16
	ANN	1.20	19.69	1.52	0.10
	DL	**0.80**	**13.56**	**1.11**	**0.45**

243

Kernel Regression



244

Kernel Regression

- Let: $y_i = E[y_i|\mathbf{x}_i] + \varepsilon_i = g(\mathbf{x}_i) + \varepsilon_i$, where y_i $(i = 1,2,\dots,n)$ is the response variable, $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T$ is the vector of explanatory variables (covariates), and $\varepsilon_i \sim^{iid} (0, \sigma^2)$ is the model residual
- Conditional expectation function: $g(\mathbf{x}_i) = \frac{1}{p(\mathbf{x})} \int y p(\mathbf{x}, y) dy$
- Consider a nonparametric kernel estimator of the p-dimensional density of the covariates (Silverman 1986):

$$\hat{p}(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

where $K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$ is the kernel function, h is a smoothing parameter, and $\mathbf{x}$ is the "focal point" value

245

Kernel Regression

- $\hat{p}(\mathbf{x})$ is a p-dimensional density function so that the kernel function must be positive and $\int_{-\infty}^{\infty} \hat{p}(\mathbf{x}) d\mathbf{x} = 1$, so that:

$$\frac{1}{nh^p} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) d\mathbf{x} = 1 \rightarrow \int_{-\infty}^{\infty} \frac{1}{h^p} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) d\mathbf{x} = 1$$

- Similarly (and assuming a single h), $p(\mathbf{x}, y)$ can be estimated as:

$$\hat{p}(\mathbf{x}, y) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

where $K\left(\frac{y_i - y}{h}\right)$ is also a kernel function.

- So that $\int y \hat{p}(\mathbf{x}, y) dy = \frac{1}{nh^p} \sum_{i=1}^{n} \left[\frac{1}{h} \int y K\left(\frac{y_i - y}{h}\right) dy\right] K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$

246

Kernel Regression

- Let: $z = (y - y_i)/h$, so that $dy = hdz$ and

$$\frac{1}{h}\int yK\left(\frac{y_i - y}{h}\right)dy = y_i\int K(z)dz + hE[z]$$

- Assuming a proper $K(z)$ (i.e., $\int K(z)dz = 1$) and $E[z] = \int zK(z)dz = 0$, then $\int y\hat{p}(\mathbf{x}, y)dy = \frac{1}{nh^p}\sum_{i=1}^{n} y_i K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$
- Hence $\widehat{E}[y_i|\mathbf{x}_i] = \hat{g}(\mathbf{x}_i) = \frac{1}{\hat{p}(\mathbf{x})}\int y\hat{p}(\mathbf{x}, y)dy = \sum_{i=1}^{n} w_i(\mathbf{x})y_i$

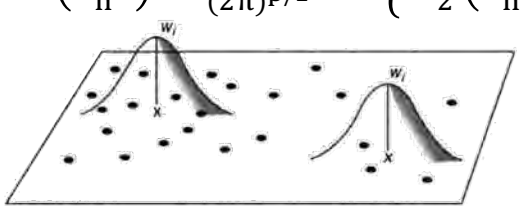
  where $w_i(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}$ are weights that depend on the choice of kernel function and smoothing parameter h

247

Kernel Regression

- Example: Gaussian kernel

$$K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) = \frac{1}{(2\pi)^{p/2}}\exp\left\{-\frac{1}{2}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)^T\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)\right\}$$



X: regression point
●: data point

Note: h controls the decay rate; smaller h implies more abruptly decrease of $w_i(\mathbf{x})$, i.e. more 'local' regression

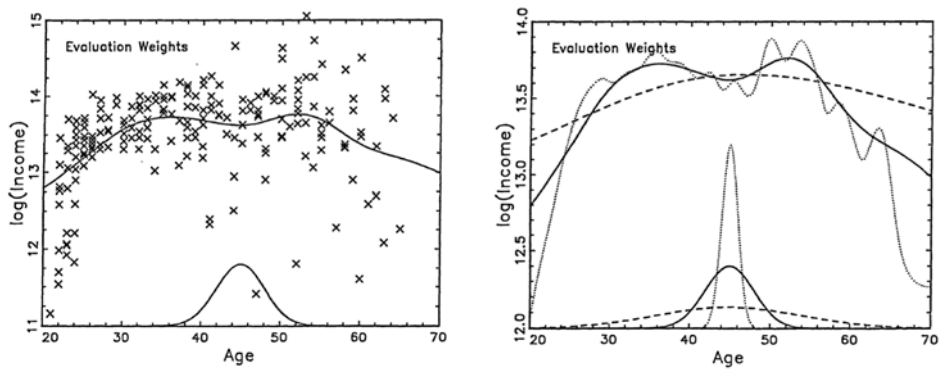
- Specific case: Additive regression model (Hastie & Tibshirani 1990)

$$g(\mathbf{x}_i) = \sum_{j=1}^{p} E[y_i|x_{ij}] = \sum_{j=1}^{p} g_j(x_{ij}) \quad \text{(no interactions)}$$

248

Kernel Regression

- Example: Regression of log-income and age of 205 people

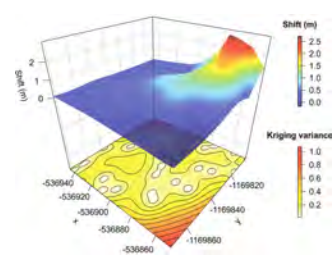


Scatter plot and smooths for earning power data using kernel $N(0, 1)$. Window widths are represented by curves: solid curves, $h = 3$; dotted curves, $h = 1$; dashed curves, $h = 9$ (Chu and Marron 1991)

249

Reproducing Kernel Hilbert Spaces

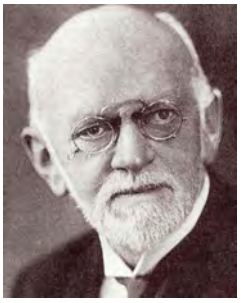
- Statistical models based on reproducing kernel Hilbert spaces (RKHS) have been useful for regression (e.g., Wahba 1990), classification (e.g., Vapnik 1998), and smoothing in highly dimensional problems.
- Examples of application can be found in spatial statistics (e.g. 'Kriging'; Cressie, 1993), scatter-plot smoothing (e.g. smoothing splines; Wahba, 1990), genetics and genomics (Gianola et al. 2008; de los Campos et al. 2009), etc.
- RKHS regression is connected with many other statistical approaches, such as additive models, splines, and mixed models.



250

Reproducing Kernel Hilbert Spaces

- Reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions in which point evaluation is a continuous linear functional.
- A Hilbert space is a vector space equipped with an inner product which defines a distance function for which it is a complete metric space.



David Hilbert
(1862-1943)

251

RKHS Regression

- Regression model: $y_i = E[y_i|\mathbf{x}_i] + \varepsilon_i = g(\mathbf{x}_i) + \varepsilon_i$
- Estimation of $g(\mathbf{x}_i)$:

1) Least Squares or Maximum Likelihood: $\hat{g}(\mathbf{x}_i) = \arg\min_g l(\mathbf{y}, \mathbf{x})$ with $g(.)$ assumed known and expressed in a parametric form, and $l(\mathbf{y}, \mathbf{x})$ is the loss function, a measure of goodness-of-fit
2) Regularized regression: $\hat{g}(\mathbf{x}_i) = \arg\min_g \{l(\mathbf{y}, \mathbf{x}) + \lambda J(g)\}$, where $J(g)$ is a penalty on model complexity

252

RKHS Regression

3) RKHS regression: assumes g belongs to a Hilbert space or real-valued functions, denote as $g \in H$, and uses the square of the norms of g as penalty, i.e. $J(g) = \|g\|_H^2$, where $\|.\|_H$ denotes the norm in Hilbert space H

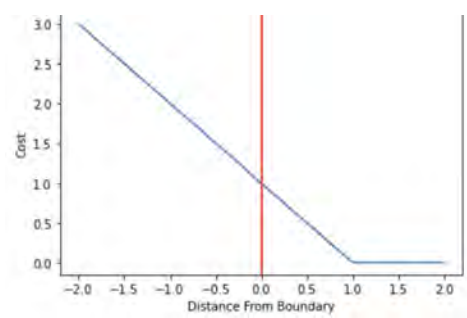
$$\hat{g}(\mathbf{x}_i) = \arg\min_{g \in H} \left\{ l(\mathbf{y}, \mathbf{x}) + \lambda \|g\|_H^2 \right\}$$

- RKHS model specification; choice of:
  - Loss function $l(\mathbf{y}, \mathbf{x})$
  - Hilbert space H
  - Smoothing parameter λ

253

RKHS Model Specification

- Standard choices of loss function: negative log-likelihood and residual sum of squares
- If the response is a binary outcome, coded as $y \in \{-1, 1\}$, and the loss function is taken to be a hinge function $l(m) = \max(0, 1 - ym)$, the problem becomes the standard support vector machine



- Smoothing parameter λ can be chosen using cross-validation, generalized cross-validation, or Bayesian methods

254

RKHS Methods and Mixed Models

- The duality between Hilbert spaces of functions and positive-definite functions is convenient, as it is easier to define a positive definite function on $\mathbf{x}$ than to define H explicitly
- Let K be an $\mathrm{n}\times\mathrm{n}$ positive definite matrix with elements $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathrm{l}(\mathbf{y}, \mathbf{x}) = (\mathbf{y} - \mathrm{g}(\mathbf{x}))^{\mathrm{T}}(\mathbf{y} - \mathrm{g}(\mathbf{x}))$ be a residual sum of squares
- Under this setting, the optimization problem can be expressed as (Kimeldorf and Wahba 1970):

$$\min_{\mathrm{g}}\{(\mathbf{y} - \mathbf{Kc})^{\mathrm{T}}(\mathbf{y} - \mathbf{Kc}) + \lambda \mathbf{c}^{\mathrm{T}}\mathbf{Kc}\}$$

where c is an $\mathrm{n}\times 1$ vector of unknown constants.

255

RKHS Methods and Mixed Models

- Solution: $[\mathbf{K}^{\mathrm{T}}\mathbf{K} + \lambda\mathbf{K}]\hat{\mathbf{c}} = \mathbf{K}^{\mathrm{T}}\mathbf{y}$
- Given that $\mathbf{K} = \mathbf{K}^{\mathrm{T}}$ and $\mathbf{K}^{-1}$ exists, premultiplication by $\mathbf{K}^{-1}$ yields:

$$[\mathbf{K} + \lambda\mathbf{I}]\hat{\mathbf{c}} = \mathbf{y}$$

- The estimated conditional expectation function is:

$$\hat{\mathbf{g}}(\mathbf{x}) = \mathbf{K}\hat{\mathbf{c}} = \mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}\mathbf{y} = \mathbf{Wy}$$

where $\mathbf{W} = \mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}$ is a projection matrix.

- Therefore, $\hat{\mathbf{g}}(\mathbf{x})$ is a weighted sum of the observations:

$$\hat{\mathbf{g}}(\mathbf{x}) = \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{ij}}\mathrm{y}_{\mathrm{j}}$$

where the weights w_{ij} are the entries of $\mathbf{W}$.

256

Bayesian Interpretation

- The solution to the optimization problem

$$\min_{g}\{(\mathbf{y}-\mathbf{Kc})^{T}(\mathbf{y}-\mathbf{Kc})+\lambda\mathbf{c}^{T}\mathbf{Kc}\}$$

- can be interpreted as a condition (given λ) posterior mean and mode of a Bayesian model with gaussian likelihood and a normal prior for the "regression coefficients" $\mathbf{c}$
- Let $\mathbf{y}=\mathbf{Kc}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\dots,\varepsilon_n)^{T}\sim N(\mathbf{0},\mathbf{I}\sigma_\varepsilon^2)$ is the vector of model residuals, and $\mathbf{K}$ is the kernel matrix, viewed as an incidence matrix for $\mathbf{c}$

$$\rightarrow\ \mathbf{y}\sim N(\mathbf{Kc},\mathbf{I}\sigma_\varepsilon^2)$$

257

Bayesian Interpretation

- Prior: $\mathbf{c}\sim N(\mathbf{0},\mathbf{K}^{-1}\sigma_c^2)$
- If and σ_ε^2 are σ_c^2 known, the density of the conditional posterior distribution of $\mathbf{c}$ is:

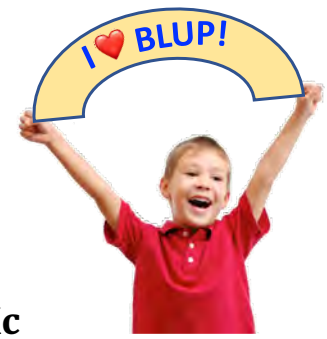
$$p(\mathbf{c}|\mathbf{K},\sigma_\varepsilon^2,\sigma_c^2,\mathbf{y})\propto\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\mathbf{y}-\mathbf{Kc})^{T}(\mathbf{y}-\mathbf{Kc})\right\}\exp\left\{-\frac{1}{2\sigma_c^2}\mathbf{c}^{T}\mathbf{Kc}\right\}$$

- This density is known to be multivariate normal with mean (mode) equal to $E[\mathbf{c}|\mathbf{K},\sigma_\varepsilon^2,\sigma_c^2,\mathbf{y}]=[\mathbf{K}+\lambda\mathbf{I}]^{-1}\mathbf{y}$, where $\lambda=\sigma_\varepsilon^2/\sigma_c^2$

258

Mixed Model

- Consider $\mathbf{y}$ centered: $\mathbf{y} = \mathbf{u} + \boldsymbol{\varepsilon}$, with $\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{K}\sigma_c^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_\varepsilon^2\end{bmatrix}\right)$
- BLUP: $\hat{\mathbf{u}} = [\mathbf{I} + \lambda\mathbf{K}^{-1}]^{-1}\mathbf{y}$, where $\lambda = \sigma_\varepsilon^2/\sigma_c^2$



Bayesian RKHS Model

- $\mathbf{y} = \mathbf{K}\mathbf{c} + \boldsymbol{\varepsilon}$, with $\begin{bmatrix}\mathbf{c}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{K}^{-1}\sigma_c^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma_\varepsilon^2\end{bmatrix}\right)$
- Consider the following change of variable $\mathbf{u} = \mathbf{K}\mathbf{c}$
- $\mathrm{E}[\mathbf{u}] = \mathbf{K}\mathrm{E}[\mathbf{c}] = \mathbf{0}$ and $\mathrm{Var}[\mathbf{u}] = \mathrm{Var}[\mathbf{K}\mathbf{c}] = \mathbf{K}[\mathbf{K}^{-1}\sigma_c^2]\mathbf{K}^{\mathrm{T}} = \mathbf{K}\sigma_c^2$, and, as $\mathbf{u}$ is a linear function of $\mathbf{c}$, it follows that $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}\sigma_c^2)$

259

Building Kernels

- Selecting a kernel is the most critical stage in applying kernel-based algorithms
- Prior knowledge about a problem may be useful but it is not always enough for choosing a specific kernel
- Kernels can be selected using model comparison techniques, e.g. cross-validation or Bayesian methods
- In addition, bandwidth parameters control how fast the (co)variance drops as points get further apart in input space. For example, in a Gaussian kernel, $h > 0$ may be used to control how local the regression is

260

Building Kernels

- Another interesting way of generating flexible kernels is to exploit polynomial kernels with positive constants

$$\mathbf{K} = \sigma_1^2\mathbf{K}_1 + \sigma_2^2\mathbf{K}_2 + \sigma_{12}^2\mathbf{K}_1\mathbf{K}_2$$

- From a Bayesian perspective, this can be viewed as a model $\mathbf{y} = \mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_{12} + \boldsymbol{\varepsilon}$, with prior

$$\mathrm{p}(\mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_{12}) = \mathrm{N}(\mathbf{f}_1|\mathbf{0}, \sigma_1^2\mathbf{K}_1)\mathrm{N}(\mathbf{f}_2|\mathbf{0}, \sigma_2^2\mathbf{K}_2)\mathrm{N}(\mathbf{f}_{12}|\mathbf{0}, \sigma_{12}^2\mathbf{K}_1 \# \mathbf{K}_2)$$

- This is equivalent to model: $\mathbf{y} = \mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_{12} + \boldsymbol{\varepsilon}$, with prior

$$\mathrm{p}(\mathbf{f}) = \mathrm{N}(\mathbf{f}|\mathbf{0}, \sigma_1^2\mathbf{K}_1 + \sigma_2^2\mathbf{K}_2 + \sigma_{12}^2\mathbf{K}_1 \# \mathbf{K}_2)$$

261

Implementation of RKHS Regression

- The fact that any RKHS regression can be parameterized as a mixed model with specific (co)variance matrices implies that available packages for mixed model implementation can be used to perform RKHS regressions
- This choice is especially efficient in situations when there is an efficient algorithm for computing K^{-1} directly from T, e.g. in animal and plant breeding where the inverse of the relationship matrix can be built directly from pedigree information

262

Example with Genomic Prediction

- Animal Model: $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$, with $\mathbf{u} \sim N(\mathbf{0}, \mathbf{K}\sigma_u^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{K}$ is a known matrix, constructed in three different ways:

  1. Pedigree information: $\mathbf{K} = \mathbf{A}$, the additive genetic (or numerator) relationship matrix, having elements given by 2 x coefficient of coancestry between individuals
  2. Genomic information (GBLUP): $\mathbf{K} = \mathbf{G}$, the genomic relationship matrix, given by $\mathbf{G} = [2\sum p_j(1-p_j)]^{-1}\mathbf{MM}^T$
  3. Both pedigree and genomic information (ssGBLUP): $\mathbf{K} = \mathbf{H}$, where $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$

263

Example with Genomic Prediction

- Matrices **A**, **G** and **H** used to estimate additive genetic effects
- More general kernels should allow capturing non-additive effects as well
- Here, two non-linear kernels were use in the context of ssGBLUP: the averaged Gaussian kernel (AK) and the arc-cosine deep kernel (DK)

Momen, M., Kranis, A., Rosa, G. J. M. and Muir, P. (2022) Predictive assessment of single-step BLUP with linear and non-linear similarity RKHS kernels: A case study in chickens. *J Anim Breed Genet* 139: 247-258.

264

Material and Methods

- Body weight (BW) and hen-housing production (HHP), recorded on 5,500 genotyped broiler chickens
- Training (TRN) and testing (TST) sets with different genotyping rates (20, 40, 60 and 80% of birds) in 3 selective genotyping scenarios (genotyping of the youngest individuals in the pedigree, random genotyping, and genotyping based on parent average)
- Model with H matrix described as:

$$\mathbf{H} = \begin{bmatrix} \mathrm{var}(\boldsymbol{u}_1) & \mathrm{cov}(\boldsymbol{u}_1, \boldsymbol{u}'_2) \\ \mathrm{cov}(\boldsymbol{u}_2, \boldsymbol{u}'_1) & \mathrm{var}(\boldsymbol{u}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}(\boldsymbol{K} - \boldsymbol{A}_{22})\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{21}\boldsymbol{A}_{22}^{-1}\boldsymbol{K} \\ \boldsymbol{K}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{K} \end{bmatrix}$$

265

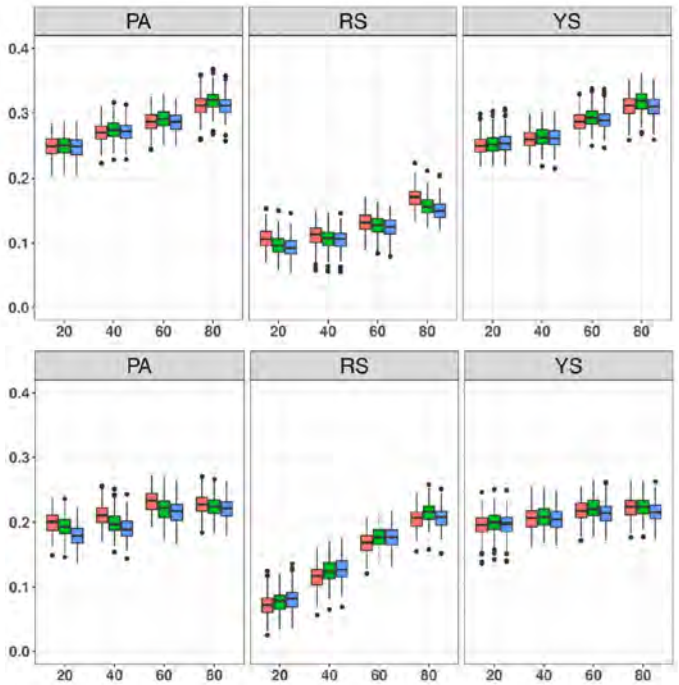
Material and Methods

- Non-linear Kernels:
  - Gaussian kernel (GK): $\mathbf{GK}_{ii'} = \exp\left(-h\frac{\|\mathbf{M}_i - \mathbf{M}_{i'}\|^2}{Q}\right)$

    where $\|\mathbf{M}_i - \mathbf{M}_{i'}\|^2$ is the Euclidean distance between the vectors of SNP markers of individuals i and i' normalized to (0, 1)
  - Arc-cosine kernel (Deep kernel, DK): similarity between two genotyped individuals given by the angle between their vectors of SNP markers $\theta_{i,i'} = \cos^{-1}\left(\frac{\mathbf{M}_i.\mathbf{M}_{i'}}{\|\boldsymbol{M}_i\|\,\|\boldsymbol{M}_{i'}\|}\right)$

    Recursive algorithm: $\begin{cases} \mathbf{AK}^{(l+1)}(\mathbf{M}_i.\mathbf{M}_{i'}) = \frac{1}{\pi}\left[\mathbf{AK}^{(l)}(\mathbf{M}_i.\mathbf{M}_i)\,\mathbf{AK}^{(l)}(\mathbf{M}_{i'}.\mathbf{M}_{i'})\right]^{\frac{1}{2}} J_t\left(\theta_{i,i'}^{(l)}\right) \\ \theta_{i,j}^{(l)} = \cos^{-1}\left\{\mathbf{AK}^{(l)}(\boldsymbol{M}_i.\boldsymbol{M}_j)\left[\mathbf{AK}^{(l)}(\boldsymbol{M}_i.\boldsymbol{M}_i)\,\mathbf{AK}^{(l)}(\boldsymbol{M}_j.\boldsymbol{M}_j)\right]^{-\frac{1}{2}}\right\} \end{cases}$

266



Boxplot of the Fisher's z-transformed predictive rank correlations using H matrices based on the genomic relationship matrix (G), averaged Gaussian kernel (AK) and deep kernel (DK) for body weight (Top Panel) and hen-house production (Bottom Panel) in the single-step GBLUP model (ssGBLUP).

267

Results and Discussion

- Prediction accuracy was influenced by the type of kernel when a large proportion of birds was genotyped
- An advantage of non-linear kernels (AK and DK) was more apparent when 60 and 80% of birds had been genotyped.
- The results indicated that AK and DK are more effective than G when a large proportion of the target population is genotyped.
- ssGBLUP with AK or DK models should perform better than G for traits with important non-additive genetic effects

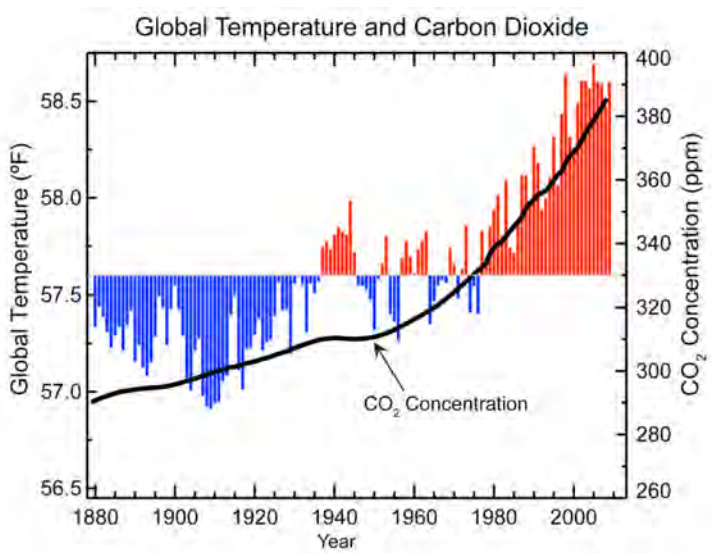
268

Correlation and Causation

- Association vs. Causation
- Confounding and Selection Bias
- Randomization
- Analysis of Observational Data
  - Propensity Score
  - Instrumental Variable
  - Bayesian Networks
- Causal Assumptions

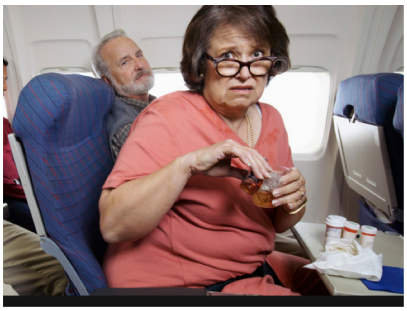
269

Prediction vs. Causal Inference



270

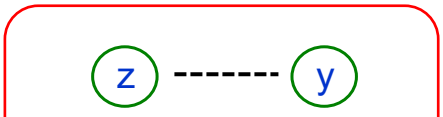
Causal Inference



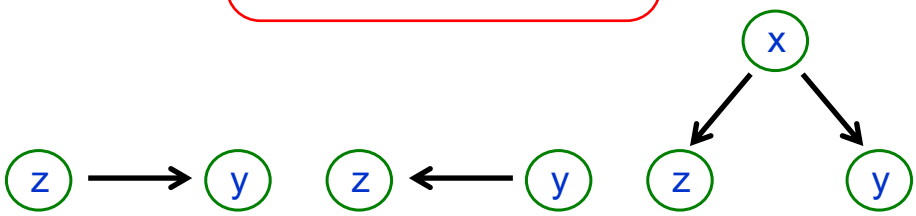
"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

271

Association vs. Causation



"z is associated with y"



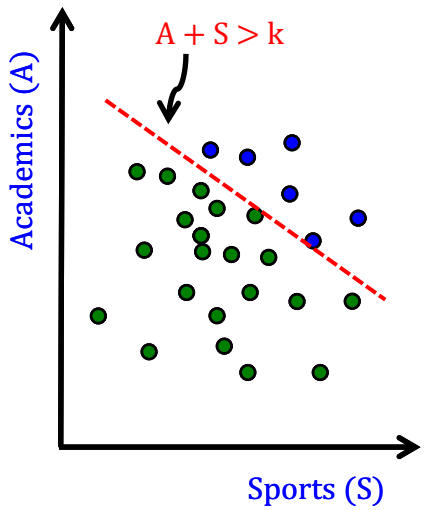
"z causes y" "y causes z" "x causes z & y"

272

Confounders

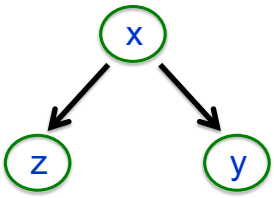
273

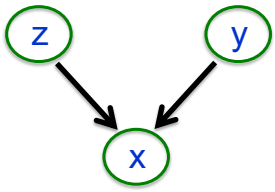
Selection Bias



274

Confounding and Selection Bias





Confounding

(x is a common cause for z and y)

Selection Bias

(z and y observed only for a subset of x values)

275

Randomized Trials

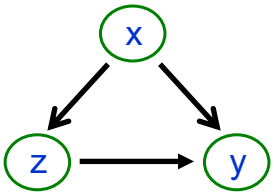
Lady tasting tea

Sir R. A. Fisher

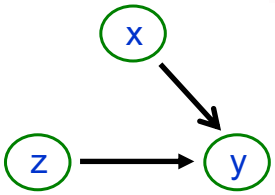
276

Randomized Experiments

- Testing the effect of z on y.



Causal relationship between variables



Effect of randomization applied to variable z

277

Observational Studies

- Lack of randomization due to legal, ethical, or logistics reasons
- Potential bias and confounding effects
- Example:
  Parenthood and life expectancy

278

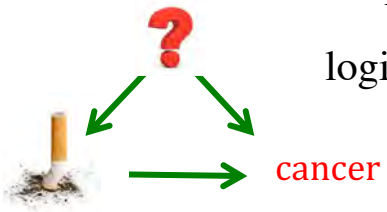
Analysis of Observational Data

- ⇨ Regression techniques with carefully chosen covariables
- ⇨ Propensity score techniques
- ⇨ Instrumental variables
- ⇨ DAGs

279

Propensity Score

- Propensity Score (PS): Conditional probability of assignment to a particular category of the causal variable given the values of the confounder set (Rosenbaum and Rubin 1983)
- Three different techniques: Matched Samples, Stratification, and Regression



$$PS_i = \Pr(\text{smoke} \mid x_i) = p_i$$

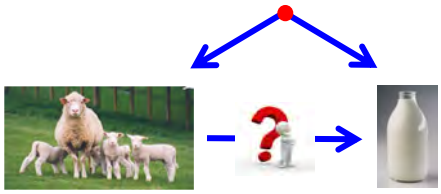
$$\text{logit}(p_i) = \beta_0 + \underbrace{\beta_1 x_{1i} + \ldots + \beta_m x_{mi}}_{\text{confounders}}$$

280

Example

Inferring the Causal Effect of Number of Lambs Born on Milk Yield in Dairy Sheep

- Association between litter size (prolificacy) and milk yield (MY) has been shown in several species: mice (Skjfervold 1976 and Knight et al. 1986), rats (Yagil et al. 1976), pigs (Auldist 1998), goats (Heyden et al. 1978)



- Potential Confounders:
  Age (parity)
  Genetics
  Year, Season, etc.

281

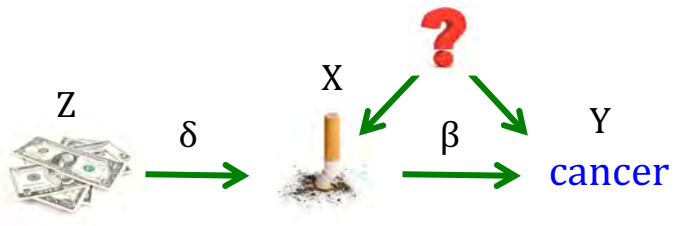
Estimated causal effect of prolificacy on MY using Propensity Scores with Matched Samples, as well as using marginal and partial regression of prolificacy on MY.

	Effect (L/lamb)	SE	95% CI
Simple Matching	20.52*	3.77	[13.13, 27.91]
Bias-corrected Matching	12.62*	3.63	[5.50, 19.74]
Marginal regression	43.93*	3.87	[36.34, 51.52]
Partial Linear regression	3.25	3.21	[-3.04, 9.56]

Ferreira VC, Valente BD, Thomas DL and Rosa GJM. Causal effect of prolificacy on milk yield in dairy sheep using propensity score. *Journal of Animal Science* 100: 8443–8450, 2017.

282

Instrumental Variable (IV)



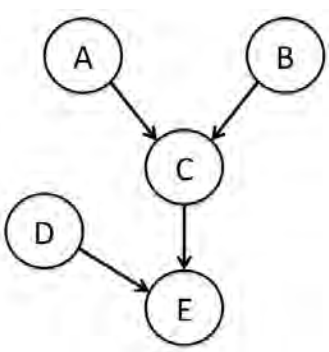
$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

$$\hat{\beta}_{IV} = (Z^T X)^{-1} Z^T Y$$

283

Bayesian Networks

- Graphic representation of a probability distribution over a set of variables → DAG



- Nodes (vertices) and arrows
- Parent, Child, and Spouse
- Joint distribution as the product of local distributions
- Markov Blanket (MB): a MB of a node is defined as the set containing its parent(s), child(ren) and spouse(s); Conditionally on its MB, a node is independent from all other nodes

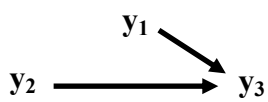
$$\Pr(A,B,C,D,E) = \Pr(E \mid C,D)\Pr(C \mid A,B)\Pr(D)\Pr(B)\Pr(A)$$

284

Inference Steps

① Structure Learning

- Score-based algorithms
- Constraint-based algorithms



$$\Lambda = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \lambda_{31} & \lambda_{32} & 0 \end{bmatrix}$$

② Parameter Estimation

$$\mathbf{y} = \mathbf{\Lambda y} + \mathbf{X\beta} + \mathbf{e}$$

Maximum Likelihood
or Bayesian Inference

285

Structure Learning

- Constraint-based algorithms
  - IC, PC – Spirtes et al. (2001)
  - Grow-Shrink (GS) – Margaritis (2003)
  - Incremental Association Markov Blanket (IAMB) – Tsamardinos et al. (2003)
  - Max-Min Parents & Children (MMPC)
- Score-based algorithms
  - Hill Climbing (HC) – Bouckaert (1995)
  - Tabu Search (Tabu)
- Hybrid structure learning algorithms
  - Sparse Candidate (SC) – Friedman et al (1999)
  - Max-Min Hill Climbing (MMHC) – Tsamardinos et al. (2006)

286

Constraint-based algorithms

- Series of conditional independence tests (parametric, semiparametric and permutation)
  - Linear correlation or Fisher's Z (continuous data; multivariate normal distribution)
  - Pearson's X^2 or mutual information (categorical data; multinomial distribution)
  - Jonckheere-Terpstra (ordinal data)

Score-based algorithms

- Different score functions
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - multinomial log-likelihood, Dirichlet posterior density (BDe) or K2 score (categorical data)

287

Example: Egg Production in Poultry

- Two strains (L1 and L2) of European Quail
- 31 traits (female quails):
  - Body weight
  - Weight gain
  - Age at first egg
  - Egg production
  - Egg quality traits

Felipe VPS, Silva MA, Valente BD and Rosa GJM. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. *Poultry Sci.* 94:772-780, 2015.

288

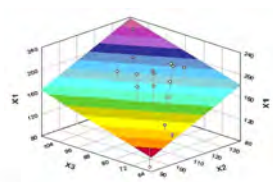
Material & Methods

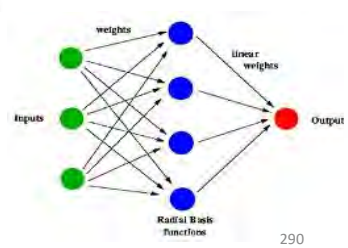
- Sample sizes (training and test sets):
  - Line 1 (90 + 90), Line 2 (102 + 103)
- Traits:
  - Weekly body weight (birth to 35 d, BW1 to BW6)
  - Weight gain (0-35 and 21-35 d, WG1 and WG2)
  - Age at first egg (AFE)
  - Egg quality traits, four time points: 125, 170, 215, 260 d
    Egg Weight – Ew, Yolk Weight – Y, Egg Shell Weight – ES
    Egg White Weight – EW, Egg Specific Gravity - DENS
  - Partial Egg Production (35-80d, EP1) and
    Total egg production (35-260d, TEP)

289

Material & Methods

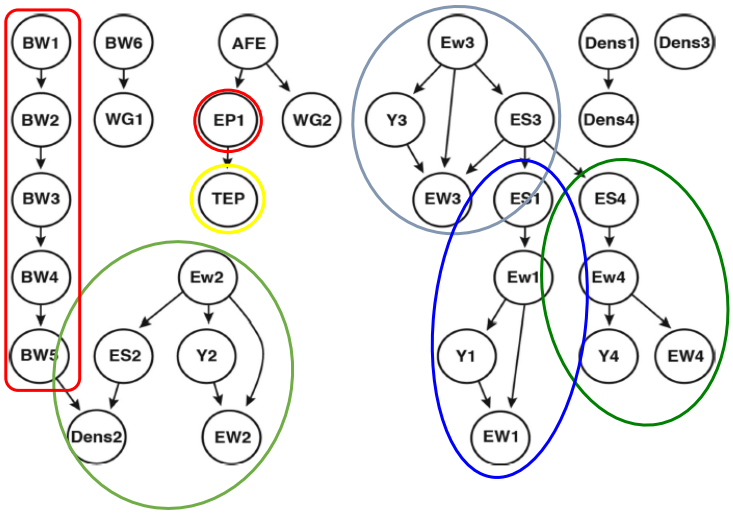
- Multiple regression analysis
  - Step-wise OLS
- Bayesian Networks
  - MB detection
- Artificial Neural Networks
  - Machine learning tool to map relationship between inputs and output

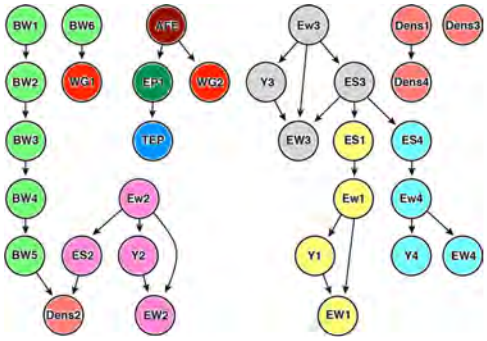




290

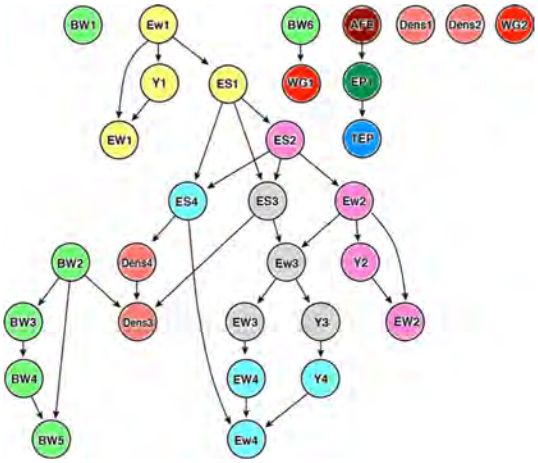
- Structure Learning (L1): Given EP1, TEP is independent from the other traits



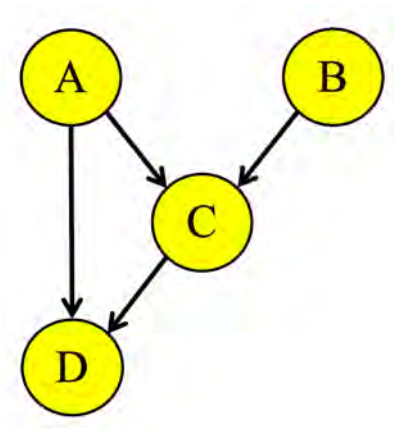


Structure for L1

Structure for L2



Causal Inference



- Arrows: Causal interpretation; consequences of intervention
- Direct, indirect and total effects
- Additional assumptions: Markov condition, faithfullness and causal sufficiency assumptions

293

Causal Inference

- Prediction of the result of an intervention (gene knockout, management decision, treatment effect)
- Estimation of causal effects:

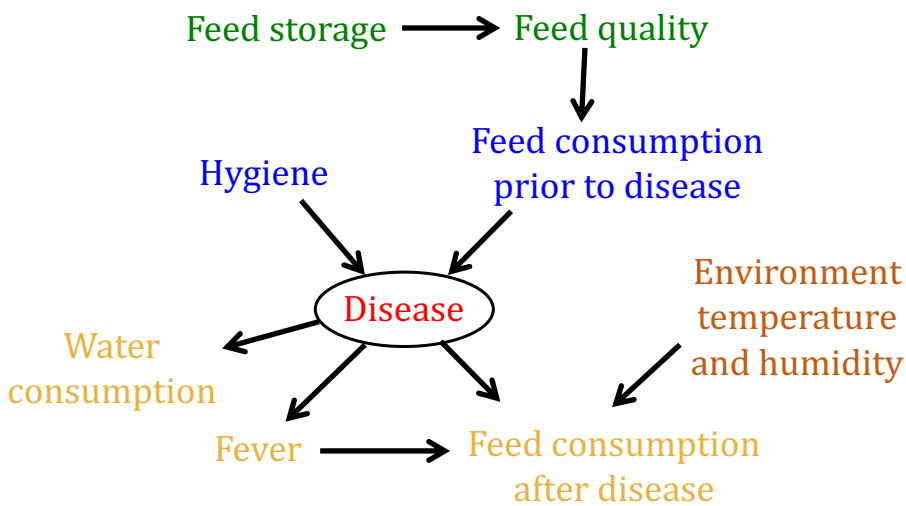
If the causal DAG is known and the distribution is multivariate Gaussian, then the causal effect (β) of X on Y can be estimated from the regression :

$$E[Y] = m + \beta X + pa(X)$$

i.e., DAG determines adjustment variables
[backdoor adjustment; Pearl (1993)]

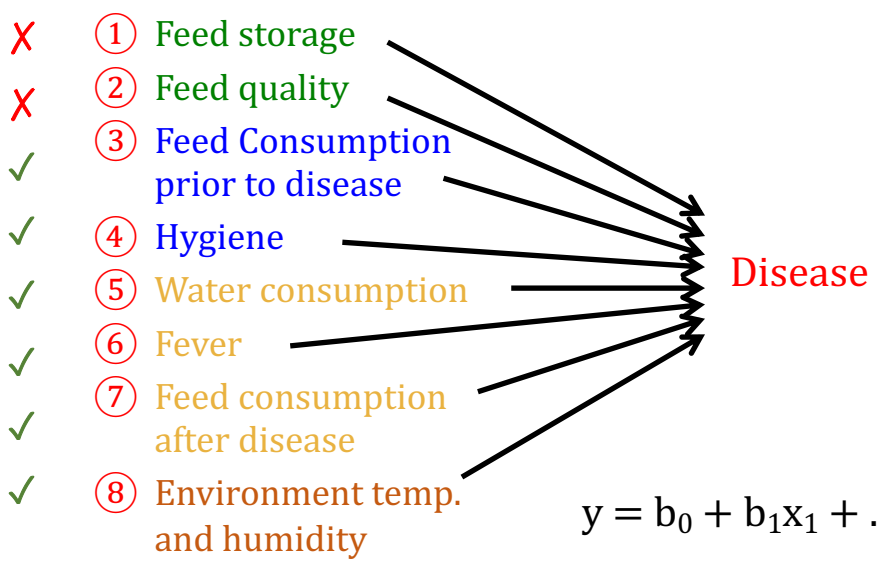
294

Fictitious Causal Network



295

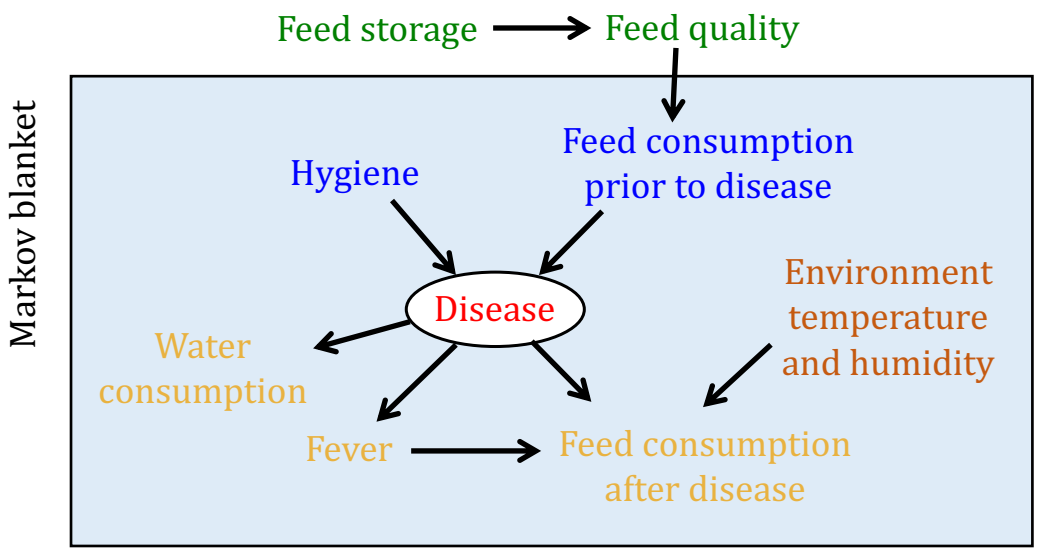
Multiple Regression Analysis



$$y = b_0 + b_1x_1 + \ldots + b_px_p + e$$

296

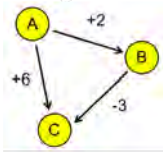
Fictitious Causal Network



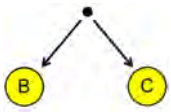
297

Causal Assumptions

- Markov condition: given its parents, a node is independent of all its non-descendants.



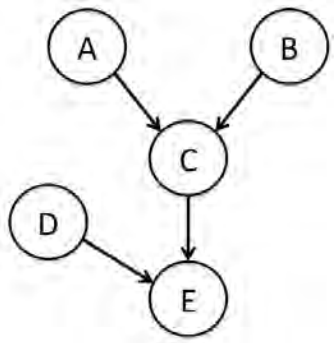
- Faithfulness: The joint distribution has all of the conditional independence relations implied by the causal Markov property, and only those conditional independence relations.
- Causal sufficiency: No pair of variables has a latent (unobserved) common cause.



The assumption of causal sufficiency is equivalent to the assumption of independence of exogenous variables. This assumption can be relaxed in structure learning — some search algorithms proposed by Spirtes et al. (1993) allow for discovery of models that are not causally sufficient. In this case, the algorithm suggests possible common causal predecessors of any pair of the measured variables.

298

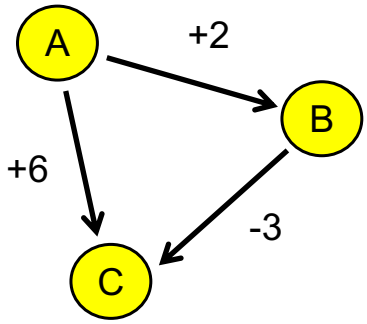
Markov Condition



$$\Pr(A,B,C,D,E) = \Pr(E \mid C,D)\Pr(C \mid A,B)\Pr(D)\Pr(B)\Pr(A)$$

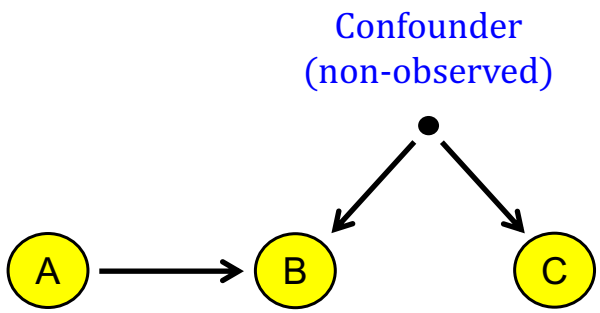
299

Faithfullness



300

Causal Sufficiency

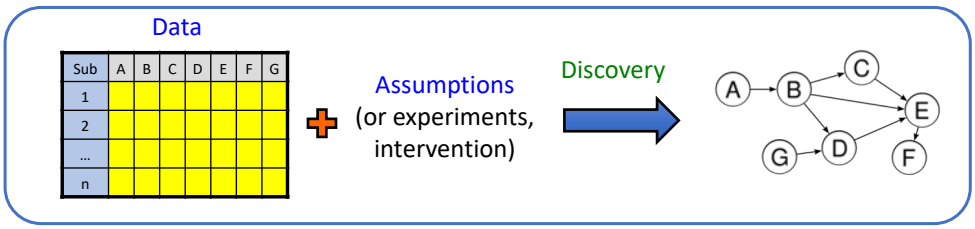


301

Causal Assumptions

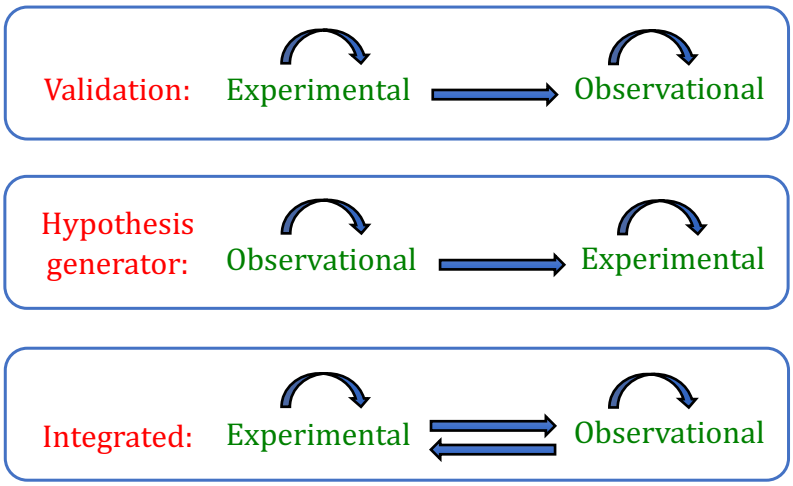
"No causes in, no causes out." (Nancy Cartwright, 1994)

Prior causal knowledge must be supplied to be able to learn new causal information.



302

Experimental and Observational Studies



303

Inferring Causal Effects from Observational Data in Agriculture

Rosa, G. J. M. and Valente B. D. Inferring causal effects from observational data in livestock. *Journal of Animal Science* 91: 553-564, 2013.

Bello, N. M., Ferreira, V. C., Gianola, D. and Rosa, G. J. M. Conceptual framework for investigating causal effects from observational data in livestock. *Journal of Animal Science* 96: 4045-4062, 2018.

304

Experimental and Observational Studies

Feature	Controlled Experiment	Observational Study
Randomization	Yes (hopefully!)	No (partially)
Sample size	Smaller	Larger
# Factors involved	Fewer	Multiple; interactions
Cost of data collection	Higher	Lower; quite often already available
Causal inference	Gold standard	Complex, but feasible (?)
Direct applicability of results to commercial settings	Not always	Yes
Prediction of field outcomes	Complex	Gold standard
Most important issues	Imperfect randomization, missing data, narrower conclusion/extrapolation	Confounding, selection bias, data size/complexity

305

Additional Topics

- Some Other Machine Learning Methods
  - Recurrent Neural Network
  - Convolutional Neural Network
  - Graph Neural Networks
- Strategies for Implementing Big Data Analysis

306

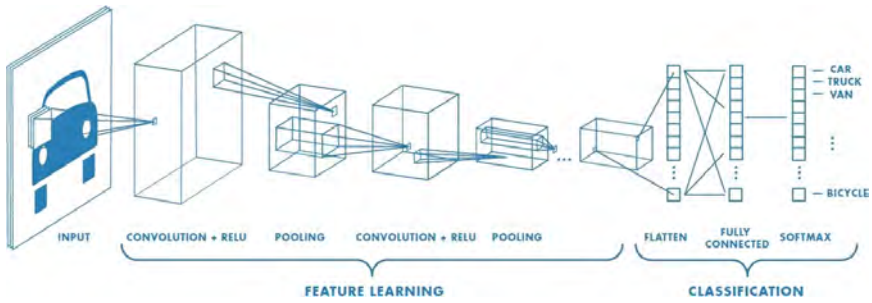
Recurrent Neural Network

- RNNs is a class of ANNs that gained popularity for time series analysis. RNNs process sequences of data by internally looping through each element of the sequence, instead of processing the whole input in a single step.
- Recurrent layers are characterized by their step function, which in the previous simple example was an activation function applied to a weighted sum of input and state features. Two other popular types of recurrent layers are Long Short-Term Memory (LSTM) layers and Gated Recurrent Units (GRUs).

307

Convolutional Neural Network

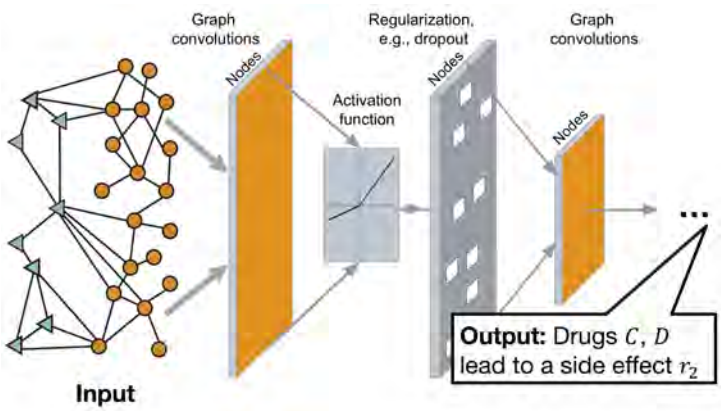
- CNN is a class of artificial neural network, commonly applied to analyze images.
- CNNs take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters.



308

Graph Neural Network

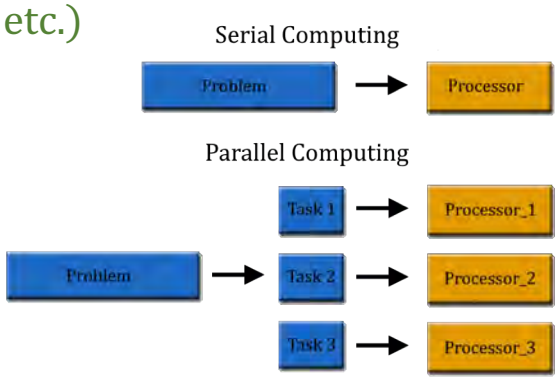
- GNN is a type of Neural Network which directly operates on the Graph structure.



309

Strategies for Implementing Big Data Analysis

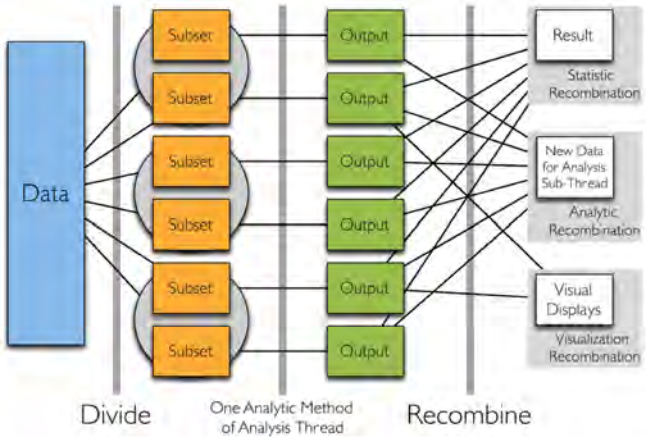
- Parallel Computing
  (easily implemented for comparison of multiple models, or different architecture of ANN, Cross-validation runs, multiple MCMC, etc.)



310

Strategies for Implementing Big Data Analysis

- Divide and Recombine (Delta-Rho)



311