Statistical modelling and Learning vs Machine Learning

'Causal modelling vs Predictive modelling'
'Health Data Science Short Course'
'University of Kwazulu-Natal'

Innocent Maposa (PhD)
Stellenbosch University
Faculty of Medicine and Health Sciences
Department of Global Health
Division of Epidemiology & Biostatist

25 August 2024

Introduction

Fundamental concepts

The goal

Innocent Maposa (PhD)
Stellenbosch University
Faculty of Medicine and Health Sciences
Department of Global Health
Division of Epidemiology & Biostatistics

- The purpose of statistics is to summarize data and quantify uncertainty around the *statistics*.
  - Descriptive
  - Inference including $\rightarrow$ Hypothesis testing, p-values, confidence intervals $\rightarrow$ Generalization
    - Bivariate
    - Regression (inference on the parameters)
- Predictive modelling mainly aims to find a *function* which can predict unseen outcomes based on new *feature* inputs with high **accuracy**.
- First we lay the foundational thoughts and philosophy in the *learning goals and processes*

The Causal framework

- The *Hunter* → the mental model → the chances of success

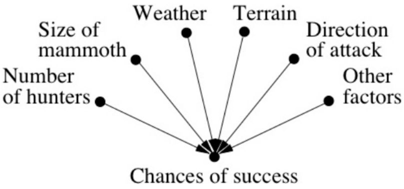


Figure 1: *Why do we observe a success? : credit: Judea Pearl*

- The human mental models are always seeking to address this question: WHY?**[ref-Pearl]**.
    - Causal modelling and inference is all about taking this question seriously
    - Understanding the mechanism of occurrence led to human progress over centuries!

Causal Framework

Innocent Maposa (PhD)
Stellenbosch University
Faculty of Medicine and Health Sciences
Department of Global Health
Division of Epidemiology & Biostatistics

- The ladder of causation

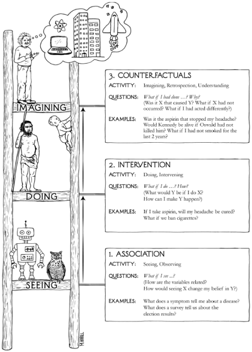


Figure 2: *Three levels of causation? : credit: Judea Pearl*

Association level

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Figure 3: *Three levels of causation? : credit: Judea Pearl*

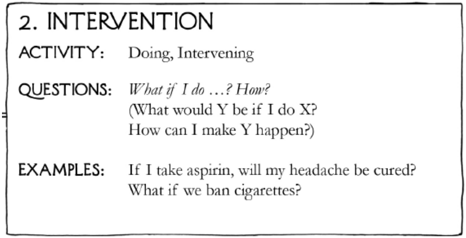
- *How would seeing X change my belief in Y?*
- Can rephrase to: *How would seeing X influence my understanding of Y?*
  - By observing X, can I say something about unobserved Y?
    - Under what circumstances (assumptions)?
- Challenges with this level of evidence includes *BIAS* - confounding, selection, mediation, moderation,

Association

Figure 4: *Correlation challenges*

- Causation as a limit for correlation (*causation always implies correlation*) - Pearson

Intervention Level



2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

Figure 5: *Three levels of causation? : credit: Judea Pearl*

- *How can I make Y happen?*
  - In other words, can I do something to influence the outcome of interest?
- Study design elements are optimized to *minimize (eliminate) bias*

Counterfactual

3. COUNTERFACTUALS

ACTIVITY:	Imagining, Retrospection, Understanding
QUESTIONS:	*What if I had done ...? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES:	Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Figure 6: *Three levels of causation? : credit: Judea Pearl*

- *Was it X that caused Y? What if X had not happened? What if I had acted differently?*
  - These are high level questions that cannot be answered by just seeing and observing.
  - Most statistical paradigms that rely on learning from data are limited at this level

Law of Regression

- **The average regression of the offspring to a constant fraction of their respective mid-parental deviations, which was first observed in the diameters of seeds, and then confirmed by observations on human stature, is now shown to be a perfectly reasonable law which might have been deductively foresee[ref-Galton]**
  - The introduction of *regression* as a principle that can help us understand relationship
    - be they causal or *associational*
- There are two goals in analysing the data:
  - Explanation (Information): To extract some information about how nature is associating (relating) the response variables to the input variables
  - Prediction: To be able to predict what the responses are going to be to future input values[ref-Breiman]

What is regression?

- The goal of regression is to model the relationship between the response (outcome or target) variable Y and predictor(s) variable(s) X using the form

$$Y = f(X) + \epsilon$$

- where the function f describes the functional form of the relationship between variables and ϵ accounts for error. This relationship can qualitatively be thought of in different ways:
  - response = deterministic + random
  - response = signal + noise
  - response = model + unexplained
  - response = prediction + error
- Linear and generalized linear models make strong assumptions about the data generating process ie the structure of this model and restricts $f(X)$ to linear functions of X ie $Y = X\beta + \epsilon$.

Causal Modelling

Classical Statistical Modelling

- Classical statistical modelling refers to practices aiming to conduct model validation, and thus, statistical inference on one or several quantities of interest eg distributions, model parameters, errors etc.
- With inference, the goal is to estimate $\hat{\beta}$ that estimates the true quantity β
  - This true quantity is assumed to exist independently of the statistical model[ref-Daoud2023Statistical]
- These models are aimed at explaining relationships between variables as main focus, prediction is of little interest
  - *Fundamental to scientific enquiry*

Classical Statistical Modelling

- Scientific methods consist of cycles of deductively formulating a hypothesis from substantive theory, testing this hypothesis in a model and against the data, and then revising the theory based on empirical results.
- The requirement of testing substantive theories through an interpretable statistical model is one of the appeals for classical statistical modelling[ref-Daoud2023Statistical].

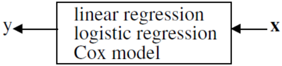


Figure 7: *Statistical Causal Modelling : credit: L.Breiman*

- Model validation: generally uses some form of goodness-of-fit tests and residual examination.

Predictive Modelling

Predictive modelling

- Sometimes referred to as *algorithmic modelling*, entails practices defining a procedure f, that generates accurate predictions, $\hat{\mathbf{Y}}$, about an event (outcome), $\mathbf{Y}$.
  - by accurate, we mean, predictions that are as similar as possible to the true event that f has not yet encountered.
- A procedure is an *algorithm*, or a *function*, that takes some input $\mathbf{X} = x$, operates on this input $f(\phi(x))$, and then produces $f(x) = \hat{y}$ where $\phi(x)$ are features derived from $\mathbf{X}$ and may include polynomials, interactions, etc.
  - Kernels
- The main goal is prediction and optimizing prediction function is key!
  - modern machine learning methods heavily rely on expanding the feature space in order to improve predictive accuracy

Predictive modelling

- Under predictive modelling framework and related assumptions, the relationship between X and Y may or may not be causal.
- The overarching goal is to develop a model f that operates on data inputs, producing the best possible predictions $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ that f has not observed yet.
- Absence of causal reasoning is a major limitation - however, according to Pearson, *causation is "sorely the conceptual limit to correlation or association"*.

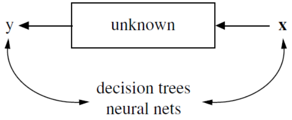


Figure 8: *Predictive Modelling : credit: L.Breiman*

- Model validation - Measured by predictive accuracy

Comparisons of the frameworks

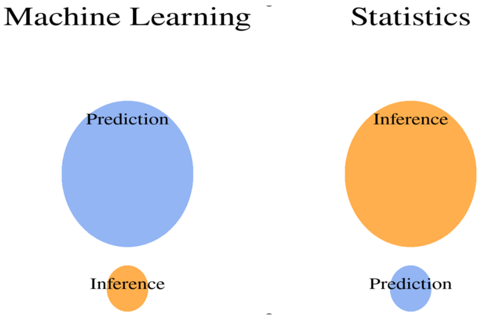


Figure 9: *Model framework goals*

Comparisons

	Data modeling culture (DMC)	Algorithmic modeling culture (AMC)
Exemplifying question	What is the causal relationship between food supply and famines?	How well can famines be predicted from available data?
Goal	Estimating unbiased parameters for causal estimation, to populate the magnitudes of the edges of a directed acylic graph (DAG).	To develop and train an algorithm f for accurate prediction.
A key assumption	Assuming a DAG, a stipulated and interpretable statistical model such as $y_i = c_0 + \beta w_i + e_i$ produces unbiased estimates of the true causal quantity β.	The algorithm f can produce accurate predictions of Y from data source, D.
Limitation	Although the parametric model is interpretable, its statistical structure may be a poor representation of the causal system.	Although f produces accurate predictions, the model is a black-box restricting causal interpretations.
Quantity of interest	$\hat{\beta}$	$\hat{Y}$

Figure 10: *Central practices of two statistical cultures:credit: L.Breiman*

Optimization functions

- For statistical causal models
  - *Loss function* $+$ *penalty*
  - $L + \lambda \sum_{j=1}^{p} \beta_j^2$ where L is the log loss function for generalized linear models and λ parameter controls how much emphasis is given to the penalty term. The higher the λ value, the more coefficients in the regression will be pushed towards zero.
- Generally, we optimize the function based on the observed variables
- For predictive models, we optimize *featurised or kernelized loss functions*
  - high dimensional
    - number of features
    - interactions etc

Two class example

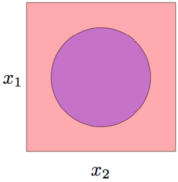


Figure 11: *Two class geometric problem:credit:D Rosenberg, NYU*

- With linear feature map $\phi(X) = (X_1, X_2)$ and linear models, no hope to separate the classes
- With appropriate nonlinearity $\phi(X) = (X_1, X_2, X_1^2 + X_2^2)$, simple.
- Example Video

Decision boundary in higher dimension

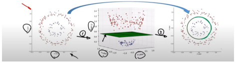


Figure 12: *Two class geometric problem:credit:D Rosenberg, NYU*

- The kernel trick optimizes expressiveness and hence prediction accuracy
- A kernel $\phi(X_i, X_j)$ is a function that quantifies the similarities between observations by summarizing the relationship between every single pairs in the training set.

Examples

Innocent Maposa (PhD)
Stellenbosch University
Faculty of Medicine and Health Sciences
Department of Global Health
Division of Epidemiology & Biostatistics

Data

```
## # A tibble: 5 x 8
##     L_SEP L_ethnicity cancer hba1c sample    id   BME deprived
##     <dbl>       <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>
## 1  1.11       -0.218       0  9.31      1     1     0        5
## 2 -0.206       2.29        1 10.7       1     2     1        3
## 3  1.22       -0.0640      1 11.1       1     3     0        5
## 4  0.0993     -0.692       1  9.68      1     4     0        3
## 5  1.51       -1.38        0  9.30      1     5     0        5
```

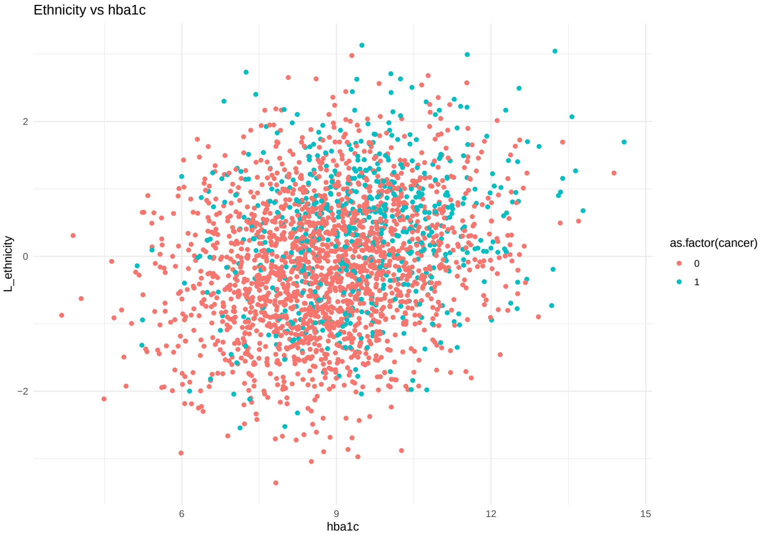
```
##      Variable    N   Mean Std. Dev.  Min Pctl. 25 Pctl. 75  Max
## 1       L_SEP 2500 -0.012         1 -3.7     -0.7      0.7  3.5
## 2 L_ethnicity 2500 -0.024         1 -3.4     -0.7      0.7  3.1
## 3      cancer 2500   0.25      0.43    0        0        0    1
## 4       hba1c 2500      9       1.5  3.7      7.9      9.9   15
## 5      sample 2500   0.75      0.43    0        1        1    1
## 6          id 2500   1250       722    1      626     1875 2500
## 7         BME 2500   0.25      0.43    0        0        1    1
## 8    deprived 2500      3       1.4    1        2        4    5
```

Describe data

Characteristic	N = 2,500[1]
L_SEP	-0.01 (-0.69, 0.70)
L_ethnicity	-0.01 (-0.74, 0.68)
cancer	613 (25%)
hba1c	8.95 (7.95, 9.93)
BME	629 (25%)
deprived	
1	509 (20%)
2	512 (20%)
3	479 (19%)
4	488 (20%)
5	512 (20%)

[1] Median (Q1, Q3); n (%)

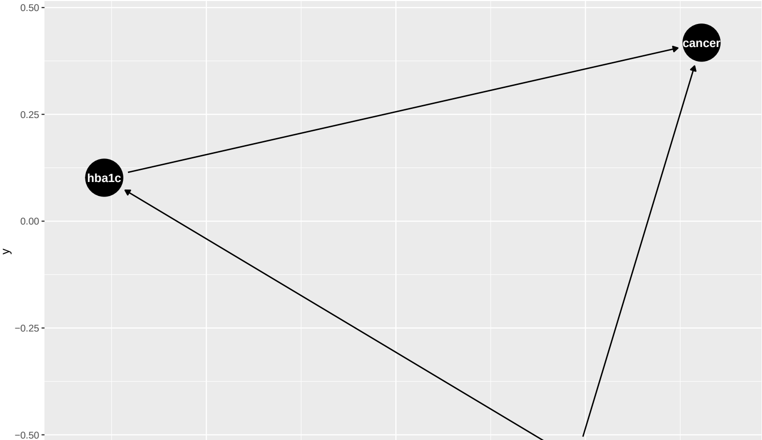
Describe data visualization



- Seems the separation problem here may be difficult

The DAG


```
##
## Attaching package: 'ggdag'
## The following object is masked from 'package:stats':
##
##     filter
```



Causal statistical model

- Question: What is the effect of hbaic on cancer?
- Unadjusted effect

Characteristic	OR[1]	95% CI[1]	p-value
hba1c	1.34	1.26, 1.43	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

Logistic regression (adjusted effect)

Characteristic	OR[1]	95% CI[1]	p-value
hba1c	1.06	0.98, 1.14	0.13
as.factor(BME)			
0	—	—	
1	1.42	1.03, 1.97	0.032
deprived	1.63	1.50, 1.77	<0.001
L_ethnicity	1.37	1.17, 1.60	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

Within sample predict

Innocent Maposa (PhD)
Stellenbosch University
Faculty of Medicine and Health Sciences
Department of Global Health
Division of Epidemiology & Biostatistics


```
##       Variable    N   Mean Std. Dev.  Min Pctl. 25 Pctl. 75  Max
## 1        L_SEP 2500 -0.012         1 -3.7     -0.7      0.7  3.5
## 2  L_ethnicity 2500 -0.024         1 -3.4     -0.7      0.7  3.1
## 3       cancer 2500   0.25      0.43    0        0        0    1
## 4        hba1c 2500      9       1.5  3.7      7.9      9.9   15
## 5       sample 2500   0.75      0.43    0        1        1    1
## 6           id 2500   1250       722    1      626     1875 2500
## 7          BME 2500   0.25      0.43    0        0        1    1
## 8     deprived 2500      3       1.4    1        2        4    5
## 9  cancer_prob 2500   0.25      0.16    0      0.1      0.3  0.8
## 10      c_pred 2500  0.079      0.27    0        0        0    1
```

Confusion Table

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2500
##
##
##                  | popdatex$cancer
## popdatex$c_pred |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |      1806 |       496 |      2302 |
##                 |     2.697 |     8.301 |           |
##                 |     0.785 |     0.215 |     0.921 |
##                 |     0.957 |     0.809 |           |
##                 |     0.722 |     0.198 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        81 |       117 |       198 |
##                 |    31.351 |    96.509 |           |
##                 |     0.409 |     0.591 |     0.079 |
##                 |     0.043 |     0.191 |           |
##                 |     0.032 |     0.047 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |      1887 |       613 |      2500 |
##                 |     0.755 |     0.245 |           |
```

Predictive modelling

Predictive Model

```
# Encoding the target feature as factor and splitting
suppressWarnings(suppressMessages(library(caTools)))
dcancer<-popdatex %>% select(c(cancer, hba1c, BME, de
dcancer$cancer = factor(dcancer$cancer, levels = c(0,
set.seed(123)
splitdat = sample.split(dcancer$cancer, SplitRatio =

train = subset(dcancer, splitdat == TRUE)
test = subset(dcancer, splitdat == FALSE)
```

Logistic regression (try prediction out of sample)


```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  750
##
##
##               | test$cancer
##  test$c_pred |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       547 |       155 |       702 |
##              |     0.560 |     1.723 |           |
##              |     0.779 |     0.221 |     0.936 |
##              |     0.966 |     0.842 |           |
##              |     0.729 |     0.207 |           |
## -------------|-----------|-----------|-----------|
##            1 |        19 |        29 |        48 |
##              |     8.190 |    25.192 |           |
##              |     0.396 |     0.604 |     0.064 |
##              |     0.034 |     0.158 |           |
##              |     0.025 |     0.039 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       566 |       184 |       750 |
```

Support vector machine


```
##
## Call:
## svm(formula = cancer ~ ., data = train, type = "C-
##     kernel = "linear", gamma = 1)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  884
```

SVM linear

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  750
##
##
##               | y_pred
##  test$cancer |         0 | Row Total |
## -------------|-----------|-----------|
##            0 |       566 |       566 |
##              |     0.755 |           |
## -------------|-----------|-----------|
##            1 |       184 |       184 |
##              |     0.245 |           |
## -------------|-----------|-----------|
## Column Total |       750 |       750 |
## -------------|-----------|-----------|
##
##
```

SVM radial


```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  750
##
##
##                     | y_pred2
```

SVM polynomial

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  750
##
##
##              | y_pred3
##  test$cancer |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       548 |        18 |       566 |
##              |     0.391 |     6.435 |           |
##              |     0.968 |     0.032 |     0.755 |
##              |     0.775 |     0.419 |           |
##              |     0.731 |     0.024 |           |
## -------------|-----------|-----------|-----------|
##            1 |       159 |        25 |       184 |
##              |     1.204 |    19.795 |           |
##              |     0.864 |     0.136 |     0.245 |
##              |     0.225 |     0.581 |           |
##              |     0.212 |     0.033 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       707 |        43 |       750 |
##              |     0.943 |     0.057 |           |
```

Random forest

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

##
## Call:
##  randomForest(formula = cancer ~ ., data = train, proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 23.43%
## Confusion matrix:
##      0  1 class.error
## 0 1244 77  0.05828917
## 1  333 96  0.77622378
```

Random forest train performance

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1303  147
##          1   18  282
##
##                Accuracy : 0.9057
##                  95% CI : (0.891, 0.919)
##     No Information Rate : 0.7549
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7165
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9864
##             Specificity : 0.6573
##          Pos Pred Value : 0.8986
##          Neg Pred Value : 0.9400
##              Prevalence : 0.7549
##          Detection Rate : 0.7446
##    Detection Prevalence : 0.8286
##       Balanced Accuracy : 0.8219
##
##        'Positive' Class : 0
##
```

Random forest test performance

Innocent Maposa (PhD)
Stellenbosch University
Faculty of Medicine and Health Sciences
Department of Global Health
Division of Epidemiology & Biostatistics


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 525 144
##          1  41  40
##
##                Accuracy : 0.7533
##                  95% CI : (0.7209, 0.7838)
##     No Information Rate : 0.7547
##     P-Value [Acc > NIR] : 0.5534
##
##                   Kappa : 0.1787
##
##  Mcnemar's Test P-Value : 6.421e-14
##
##             Sensitivity : 0.9276
##             Specificity : 0.2174
##          Pos Pred Value : 0.7848
##          Neg Pred Value : 0.4938
##              Prevalence : 0.7547
##          Detection Rate : 0.7000
##    Detection Prevalence : 0.8920
##       Balanced Accuracy : 0.5725
##
##        'Positive' Class : 0
##
```

References I

1 Pearl J, Mackenzie D. The book of why: The new science of cause and effect, 1st edn. USA: Basic Books, Inc., 2018.

2 Galton F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 1886; **15**: 246–63.

3 Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* 2001; **16**: 199–231.

4 Daoud A, Dubhashi D. Statistical Modeling: The Three Cultures. *Harvard Data Science Review* 2023; **5**.

References II