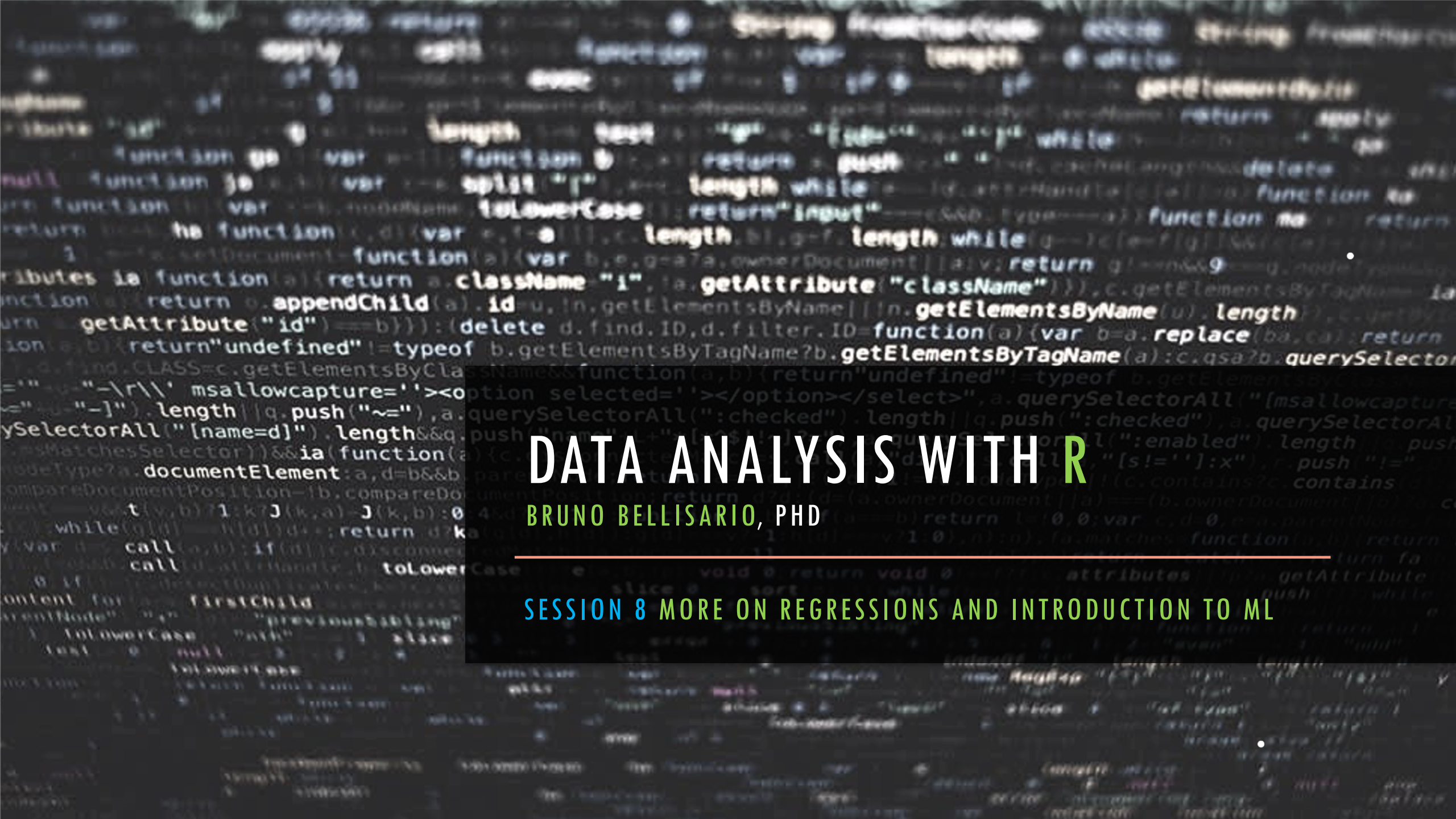


DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 8 MORE ON REGRESSIONS AND INTRODUCTION TO ML

MODEL SELECTION THE AKAIKE INFORMATION CRITERION AIC

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

→The number of independent variables used to build the model.

→The maximum likelihood estimate of the model (how well the model reproduces the data).

→The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.

MODEL SELECTION THE AKAIKE INFORMATION CRITERION AIC

AIC is most often used for model selection. By calculating and comparing the AIC scores of several possible models, you can choose the one that is the best fit for the data.

When testing a hypothesis, you might gather data on variables that you aren't certain about, especially if you are exploring a new idea. You want to know which of the independent variables you have measured explain the variation in your dependent variable.

A good way to find out is to create a set of models, each containing a different combination of the independent variables you have measured. These combinations should be based on:

→Your knowledge of the study system – avoid using parameters that are not logically connected, since you can find spurious correlations between almost anything!

→Your experimental design – for example, if you have split two treatments up among test subjects, then there is probably no reason to test for an interaction between the two treatments.

MODEL SELECTION THE AKAIKE INFORMATION CRITERION AIC

Once you've created several possible models, you can use AIC to compare them.

Lower AIC scores are better, and AIC penalizes models that use more parameters.

If two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model.

MODEL SELECTION THE AKAIKE INFORMATION CRITERION AIC

AIC determines the relative information value of the model using the maximum likelihood estimate and the number of parameters (independent variables) in the model.

$AIC = 2K-2Ln(L)$

K is the number of independent variables used and L is the log-likelihood estimate (a.k.a. the likelihood that the model could have produced your observed y-values).

To compare models using AIC, you need to calculate the AIC of each model. If a model is more than 2 AIC units lower than another, then it is considered significantly better than that model.

LINEAR REGRESSION REVISITED

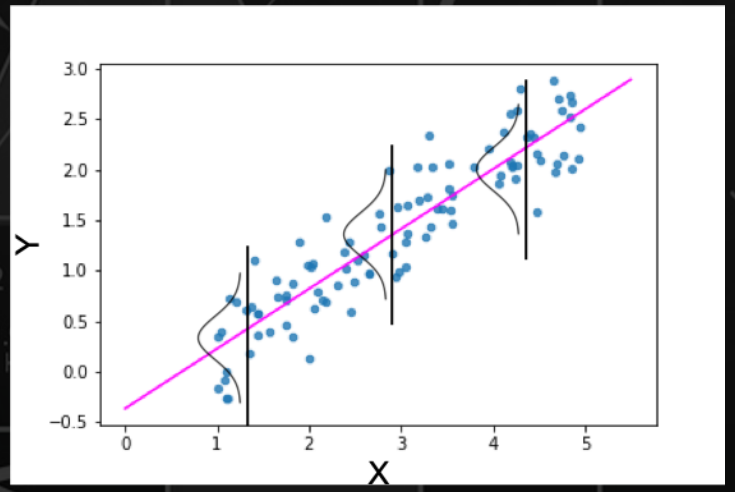
→Linear regression is used to predict the value of continuous variable y by the linear combination of explanatory variables X.

→In the univariate case, linear regression can be expressed as follows

$$\mu_i = b_0 + b_1 x_i$$

$$y_i \sim \mathcal{N}(\mu_i, \varepsilon)$$

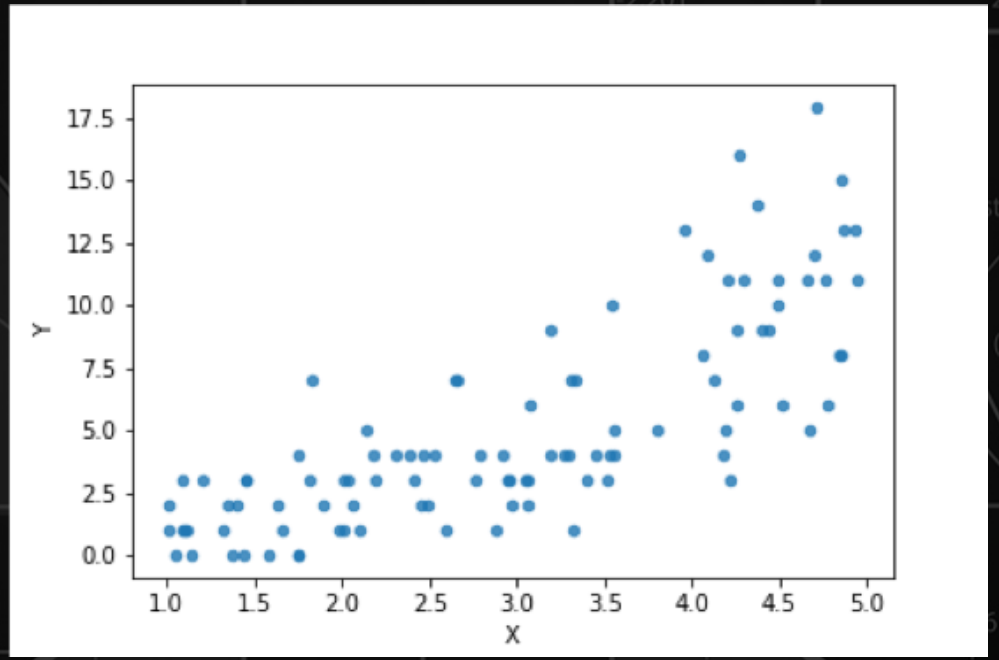
→Notice this model assumes normal distribution for the noise term. The model can be illustrated as follows;



POISSON REGRESSION

→So linear regression is all you need to know? Definitely not. If you'd like to apply statistical modelling in real problems, you must know more than that.

→For example, assume you need to predict the number of defect products (Y) with a sensor value (x) as the explanatory variable. The scatter plot looks like this.



POISSON REGRESSION

There are several problems if you try to apply linear regression for this kind of data.

→The relationship between X and Y does not look linear. It's more likely to be exponential.

→The variance of Y does not look constant with regard to X. Here, the variance of Y seems to increase when X increases.

→As Y represents the number of products, it always has to be a positive integer. In other words, Y is a discrete variable. However, the normal distribution used for linear regression assumes continuous variables. This also means the prediction by linear regression can be negative. It's not appropriate for this kind of count data.

GENERALIZED LINEAR MODELS

GLM

→Generalized linear model (GLM) is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution like Gaussian distribution.

→General refers to the dependence on potentially more than one explanatory variable, v.s. the simple linear model:

GENERALIZED LINEAR MODELS

GLM

There are three components in generalized linear models.

- Linear predictor
- Link function
- Probability distribution

Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_i)$$

Probability distribution

GENERALIZED LINEAR MODELS

GLM

→Linear predictor is just a linear combination of parameter (b) and explanatory variable (x).

→Link function literally “links” the linear predictor and the parameter for probability distribution. In the case of Poisson regression, the typical link function is the log link function. This is because the parameter for Poisson regression must be positive (explained later).

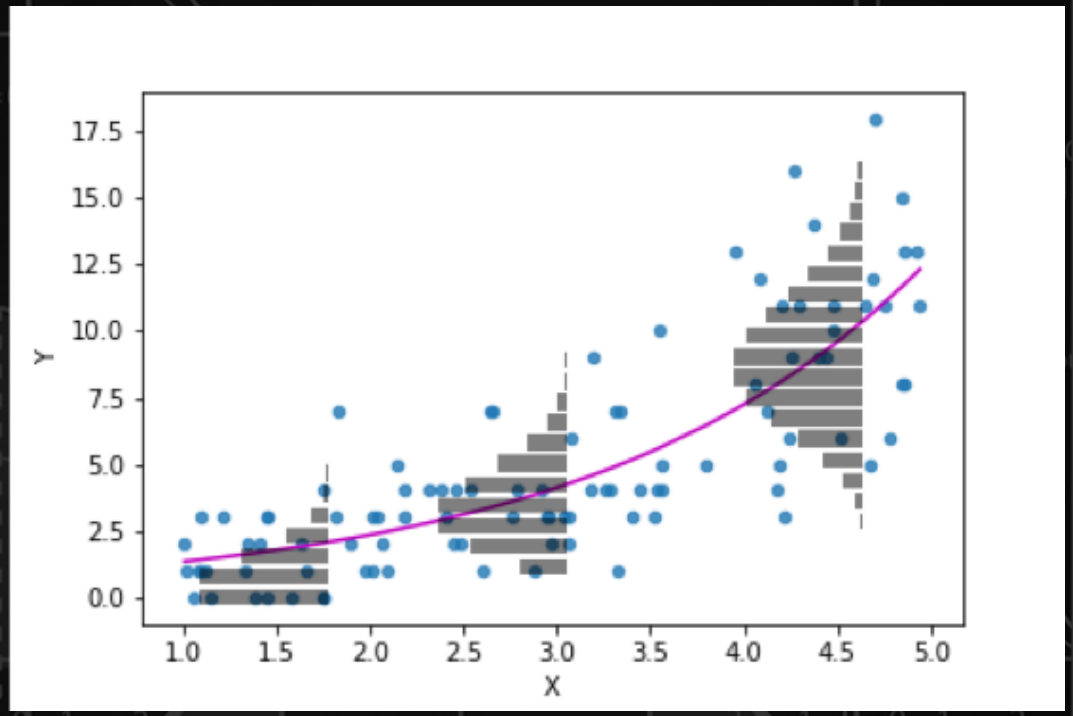
→The last component is the probability distribution which generates the observed variable y. As we use Poisson distribution here, the model is called Poisson regression.

GENERALIZED LINEAR MODELS

GLM

If we apply such a kind of regression on the previous dataset, the result should look like this.

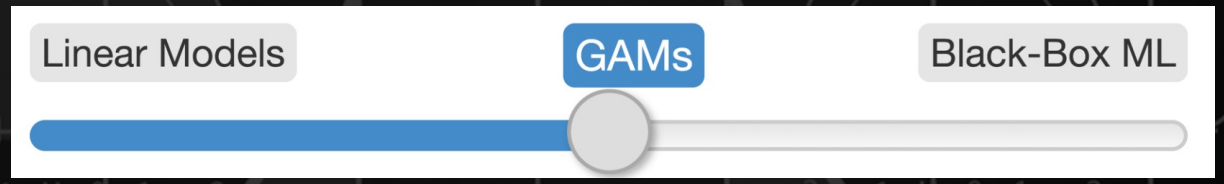
The magenta curve is the prediction by Poisson regression.



GENERALIZED ADDITIVE MODELS GAM

→Generalized additive models (GAM) offer a middle ground between simple models, such as linear regression, and more sophisticated machine learning models like neural networks that usually promise superior prediction performance to simple models.

→Many data in the environmental sciences do not fit simple linear models and are best described by “wiggly models”, also known as Generalised Additive Models (GAMs).



GENERALIZED ADDITIVE MODELS

GAM

→In linear regression, we model outcome y as a function of 2 inputs X_1 and X_2

$$y = \beta_1 X_1 + \beta_2 X_2 + u$$

→In GAM, the βX_i is replaced by $f(X_i)$, where f() can be any arbitrary nonlinear functions.

→In other words, GAM is composed of a sum of smooth functions f() on the input.

GENERALIZED ADDITIVE MODELS

GAM

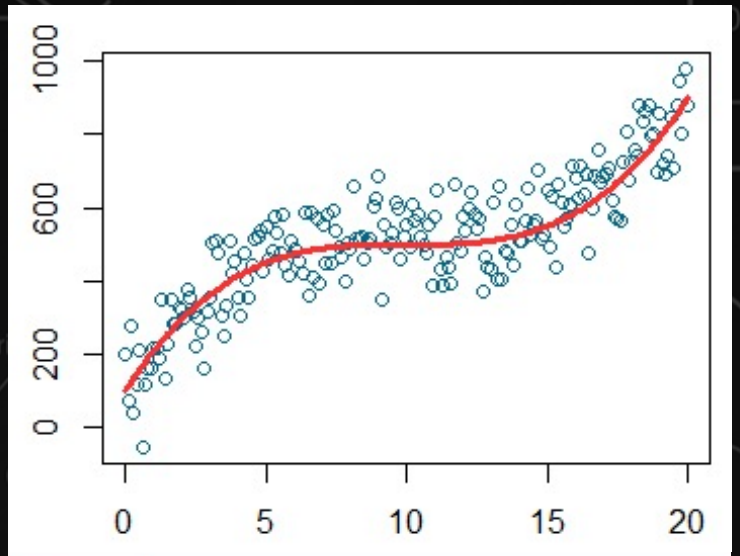
→The idea is still that each input feature makes a separate contribution to the response, and these just add up, but these contributions don't have to be strictly proportional to the inputs.

→The beauty of this approach is that, similar to what β represents in linear regression, the partial response function f() still captures the change on outcome y to change the input.

→The change in prediction depends on the initial value of X_i.

GENERALIZED ADDITIVE MODELS GAM

→A common approach to deal with nonlinear relationships in regression models involves creating polynomial features. For the predictor in question, X_i, we add terms e.g. quadratic (X_i^2), cubic (X_i^3), etc to get a better fit.



Simple Linear Regression $y = b_0 + b_1 x_1$

Multiple Linear Regression $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

Polynomial Linear Regression $y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

GENERALIZED ADDITIVE MODELS

GAM

- A common “problem” with polynomial regression is model interpretation.
- GAM encompasses this idea but includes an additional aspect: penalized estimation, where penalty terms are added to help avoid overfitting.
- Under certain circumstances, GAMs are basically an extension of generalised linear models (GLMs) with a smoothing function.

GENERALIZED ADDITIVE MODELS GAM

→Gaussian model: $y=\beta_0+x_1\beta_1+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$

→GAM model: $y=\beta_0+f(x_1)+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$

→This simply means that the contribution to the linear predictor is now some function f

BRIEF INTRODUCTION TO MACHINE LEARNING

The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.

HOW ARE STATISTICS AND MACHINE LEARNING RELATED?

Many machine learning techniques are drawn from statistics (e.g., linear regression and logistic regression), in addition to other disciplines like calculus, linear algebra, and computer science. But it is this association with underlying statistical techniques that causes many people to conflate the disciplines.

WHAT ARE THE KEY DIFFERENCES BETWEEN STATISTICS AND MACHINE LEARNING?

→The purpose of statistics is to make an inference about a population based on a sample. Machine learning is used to make repeatable predictions by finding patterns within data.

→Machine learning requires large amounts of data in order to make accurate predictions. Models are built using training data, fine tuned using a validation dataset, and are evaluated with a test dataset. All of these steps help the machine "learn."

WHAT ARE THE KEY DIFFERENCES BETWEEN STATISTICS AND MACHINE LEARNING?

→Machine learning requires no prior assumptions about the underlying relationships between the variables. You just have to throw in all the data you have, and the algorithm processes the data and discovers patterns, using which you can make predictions on the new data set. Machine learning treats an algorithm like a black box, as long it works. It is generally applied to high dimensional data sets, the more data you have, the more accurate your prediction is.

→In contrast, statisticians must understand how the data was collected, statistical properties of the estimator (p-value, unbiased estimators), the underlying distribution of the population they are studying and the kinds of properties you would expect if you did the experiment many times. You need to know precisely what you are doing and come up with parameters that will provide the predictive power. Statistical modelling techniques are usually applied to low dimensional data sets.

WHAT ARE THE KEY DIFFERENCES BETWEEN STATISTICS AND MACHINE LEARNING?

BRIEF INTRODUCTION TO MACHINE LEARNING

Machine learning algorithms are often categorized as supervised or unsupervised.

→Supervised learning - is a machine learning approach that's defined by its use of labelled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labelled inputs and outputs, the model can measure its accuracy and learn over time.

→Unsupervised learning - Uses machine learning algorithms to analyse and cluster unlabelled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are "unsupervised").

WHAT IS DATA LABELING?

Actually, what is data?

Let's first establish a clear definition of what we mean by data. Data is simply information. Any time we have a table with information, we have data. Normally, each row is a data point. Let's say, for example, that we have a dataset of pets. In this case, each row represents a different pet. Each pet is described then, by certain features.

Ok. And what are features?

Features are simply the columns of the table. The features may be size, name, type, weight, etc. This is what describes our data. Some features are special, though, and we call them labels.

WHAT IS DATA LABELING?

Labels?

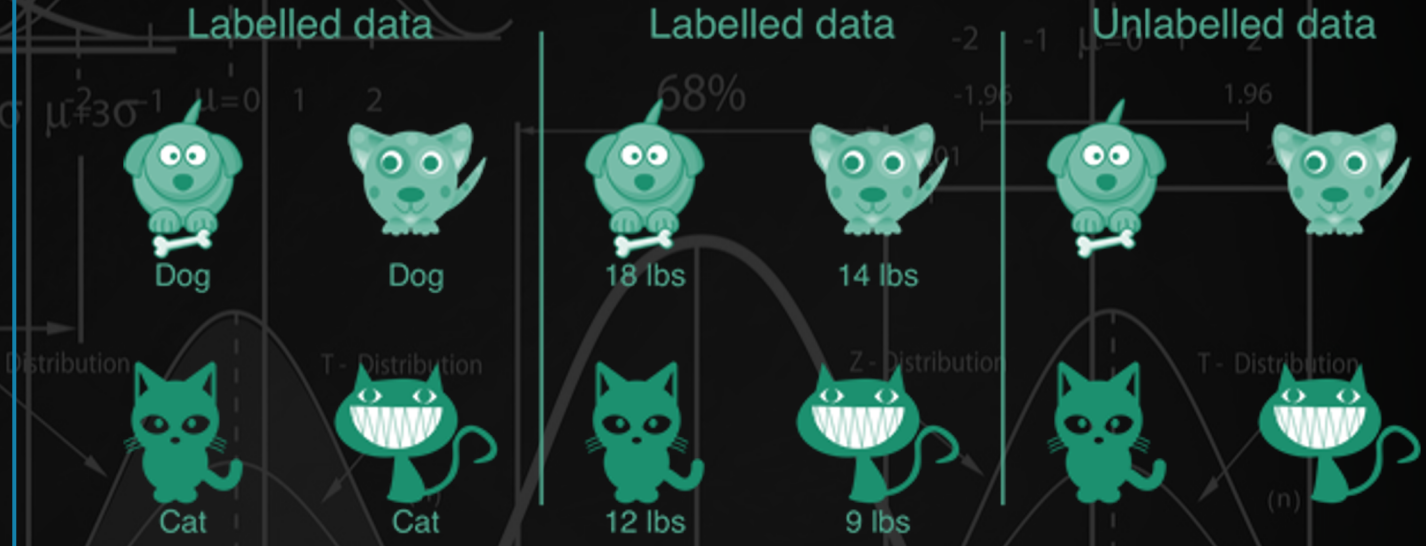
This one is a bit less obvious, and it depends on the context of the problem we are trying to solve. Normally, if we are trying to predict a feature based on the others, that feature is the label. If we are trying to predict the type of pet we have (for example cat or dog), based on information on that pet, then that is the label. If we are trying to predict if the pet is sick or healthy based on symptoms and other information, then that is the label. If we are trying to predict the age age of the pet, then the age is the label.

So now we can define two very important things, labelled and unlabelled data.

Labelled data: Data that comes with a label.

Unlabelled data: Data that comes without a label.

WHAT IS DATA LABELING?



ML APPLICATIONS IN BIOLOGY & ECOLOGY

- →Identifying gene coding regions (e.g., genomics)
- →Structure prediction (e.g., proteomics)
- →Habitat Modelling and Species Distribution
- →Species Identification
- →Food web reconstruction
- →Remote Sensing
- →Resource Management
- →Climate change