Title

Frank G DeVone

December 12, 2019

Introduction

Non-linear functions allow us to examine non-linear relationships between our variables within the framework of traditional linear regression. We will focus in this analysis on MARS (Multivariate Adaptive Regression Splines), which produces splines in a step-wise matter. To test this method we observe the Pokémon dataset compiled on Kaggle []. Pokémon is a video game series that's been ongoing since 1993, and contains a pre-constructed and artificial statistics on each of its 801 imaginary creatures. This dataset full of designed relationships is a fun and useful dataset that will allow us to explore and learn MARS.

MARS Background

Multivariate Adaptive Regression Splines functionally work in a similar manner to traditional splines. Note figure n []; we form our splines around knots at point t. We can represent this in the basis functions:

[]

The key feature of MARS is that in addition to acting as a traditional covariate step-wise it also tests different values of t for possible splines. It accomplishes this through two passes of the data:

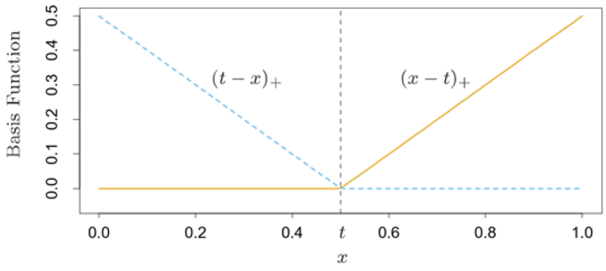


Figure 1: Knots

The Forward Pass: This part is similar to our traditional forward step-wise regression, only in addition to new and existing terms it also considers possible basis functions or a product of two. We begin with just the intercept and we then add terms in an attempt to minimize the residual sum-of-squares. Note that this pass alone creates a large model that is prone to over fitting.

The Backward Pass: After the forward pass is done we begin deleting terms based on the smallest increases in residual squared error when we do so. Where we define λ as the number of predictors in the model, we create

a best fitting model for each possible value of λ; from the intercept model to the full number of terms found in the forward pass. Once that process is complete we select the value of λ via a generalized cross-validation criterion:

[]

Note that "M(λ) is the effective number of parameters in the model: this accounts both for the number of terms in the models, plus the number of parameters used in selecting the optimal positions of the knots."[] Our final model is the one that minimizes this generalized cross-validation criterion. From our perspective the final model would be similar to something we could produce with a GAM, though the processes as we described above to get the model are quite different. They are easier to interpret then most non-linear functions of a similar nature and (critical to our analysis) can handle categorical as well as numerical data. Though MARS is capable of measuring three dimensional relationships as interactions (note product of two terms in the forward pass), for our analysis we will strictly observe two dimensional splines. Early checks into including these interaction splines found that they didn't improve our models much, were hard to compute, and drastically increased computation time (alreadt a bit of an issue).

Pokémon Background

A Pokémon is a creature within the fictional world of the video games with extraordinary natural and sometimes supernatural abilities. Our dataset contains data on the eight hundred and one Pokémon species that were know prior to November 2019. A "Generation" of Pokémon refers to the games in the series. The original one hundred and one Pokémon were from the first series of games released in 1993. When the sequels to those games were released, they introduced an additional one hundred Pokémon. This pattern continues up until generation 8 with the amount of Pokémon added varying from game to game. We elect to exclude the most recent generation, as information on that generation has yet to be reliably compiled. The purpose of the game is both the collection of all available Pokémon, working your way through the story and world with them, and using them in a turn-based battle system against the environment, other players, and computer-controlled opponents. Pokémon have 'stats' that determine their effectiveness in these battles and react with the 'stats' of the opposing Pokémon. Attack deals physical damage and is mitigated by the opposing Pokémon's defense. Special attack deals special damage (think flames and electricity) and is mitigated by the opposing Pokémon's special defense. Damage is dealt to a Pokémon's hp (health points) and Pokémon can no longer participate in battle when hp reaches zero. The final 'stat' is speed, the Pokémon with the highest speed attacks first. These six stats add up to a Pokémon's "Total Stats" which with be the

initial focus of our analysis as we attempt to predict it with the Pokémon's other characteristics. Within the games themselves these stats are calculated with the following:

(2)

(3)

'Base' as listed above are consistent across all Pokémon of the same species. However, IVs, EVs, nature, and level are characteristics the individual Pokémon is "born" with or can be trained by the player. As such our focus will only be on the base stat for a Pokémon species. There are additional characteristics that are consistent across Pokémon species, and we're using these as our potential covariates. The only ones related to real world comparisons are height and weight. Additional characteristics are fictional and will have to be explained. Understanding these characteristics is not critical to our model as we focus mostly on prediction, however their descriptions will be listed below:

Generation: As previously mentioned we are observing 8 generations of Pokémon. Conveniently each generation happens on a different "region" in the game's fictional world. As such I'm choosing to include it as a categorical variable as an indicator of the Pokémon species' region of origin.

Typing: There are 18 Pokémon types that a Pokémon. It is possible for them to have a single type, or two types, but no more then that. Note figure n for the listed types and some examples of Pokémon that fit certain paradigms. These interact with the attacks a Pokémon can learn which are also typed. For example, the electric type move thunderbolt will deal two times the damage to a Pokémon with water typing but will only deal half damage to Pokémon with rock typing. In addition, if the Pokémon that used the move is an electric type it will deal one and a half times more damage. When a Pokémon has two types one is listed first; but from a gameplay prospective there is no difference between the two. Regardless we'll include the first as "type 1" and the second as "type 2." We also create an additional "type 2" as "none" for Pokémon with only a primary typing.

Legendary Status: Some Pokémon are designated as legendary and are designed to have greater stats then normal. As such we make a binary variable to indicate if a given Pokémon species is legendary. Evolution: For some Pokémon species; when certain conditions are met, Pokémon can evolve into (usually) strictly better versions of themselves. To measure this phenomenon, we created a numerical variable. A fully evolved Pokémon will be three, a Pokémon that can evolve once will be two, and a Pokémon that can evolve two more times will be one. Like legendaries, there are 'baby' Pokémon that are inherently designed to be weaker and must be evolved in order to reach their potential, these are given a 0.

Base Egg Steps: Within the game you can receive eggs that, given time, will hatch into Pokémon. Time is judged by how many in game steps the player makes, when the threshold is met the egg will hatch. This is consistent across Pokémon species. Base Happiness: Individual Pokémon have happiness levels that increase and decrease dependent on player actions. However, each species has a baseline level of happiness they begin with.

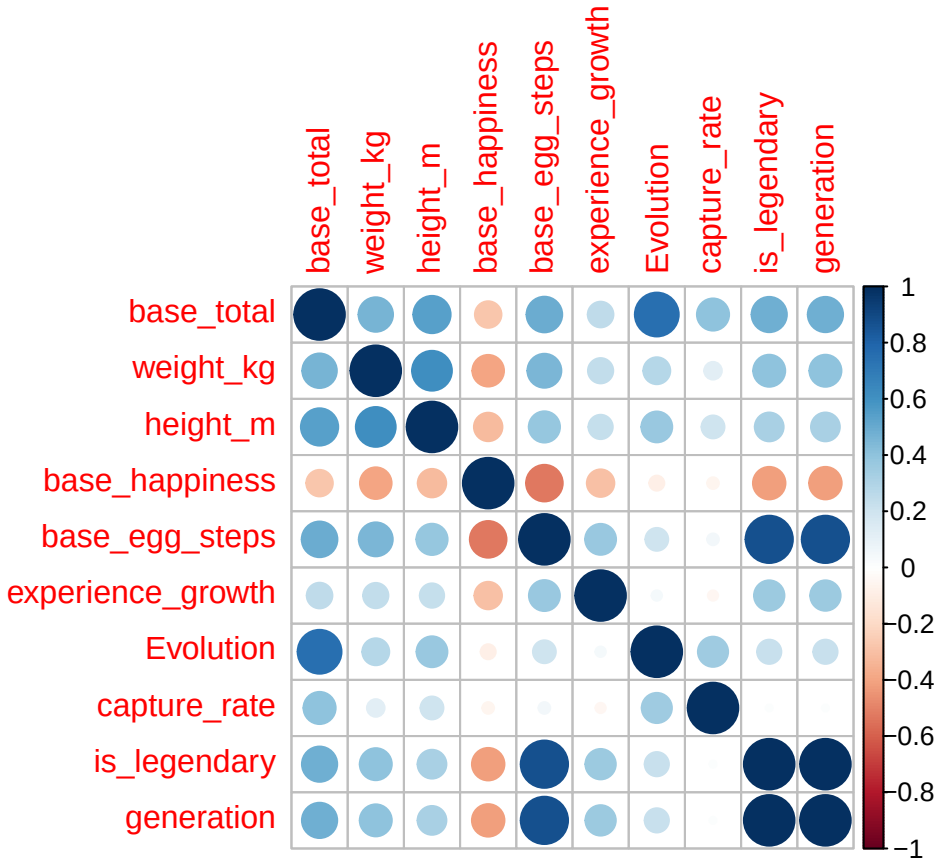
Experience Growth: Recall in Equation n the "level" component. This is a value increased as a Pokémon is used in battle. However, the amount of experience required to increase "level" varies between Pokémon species, typically with stronger Pokémon taking longer to level up.

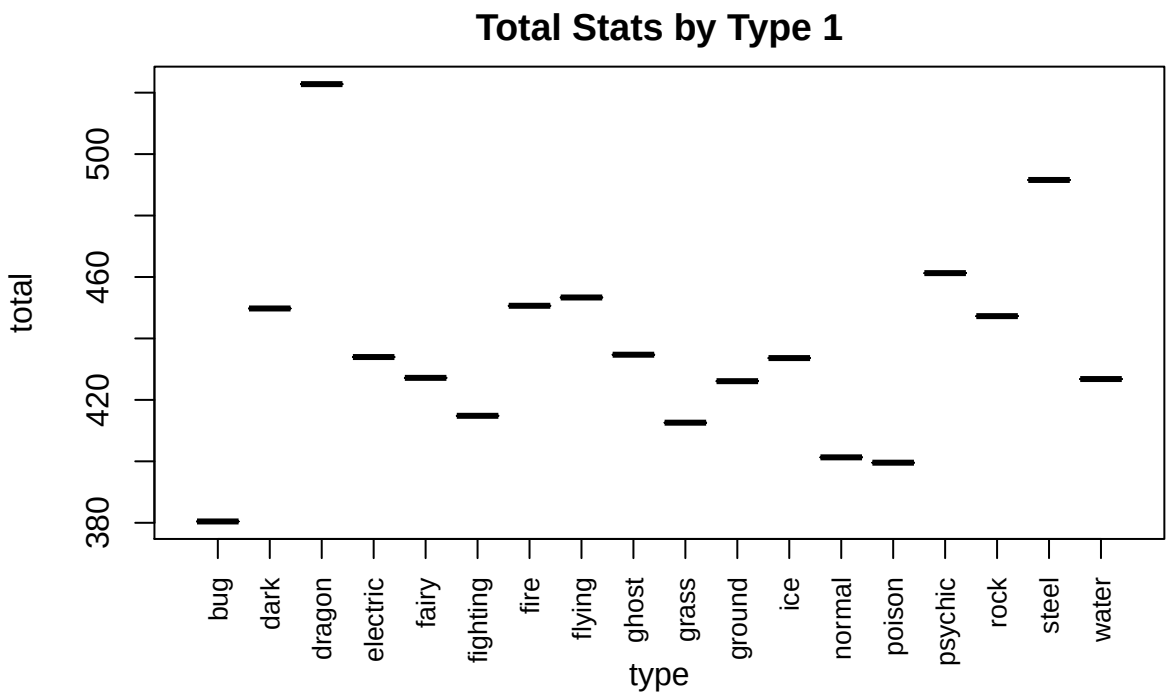
Capture Rate: There are several in game factors when attempting to capture a 'wild' Pokémon; however each has a default capture rate under the games' baseline conditions. Like experience the capture rate varies between Pokémon species, typically with stronger Pokémon being harder to capture.

Figure 2: Pokémon Types

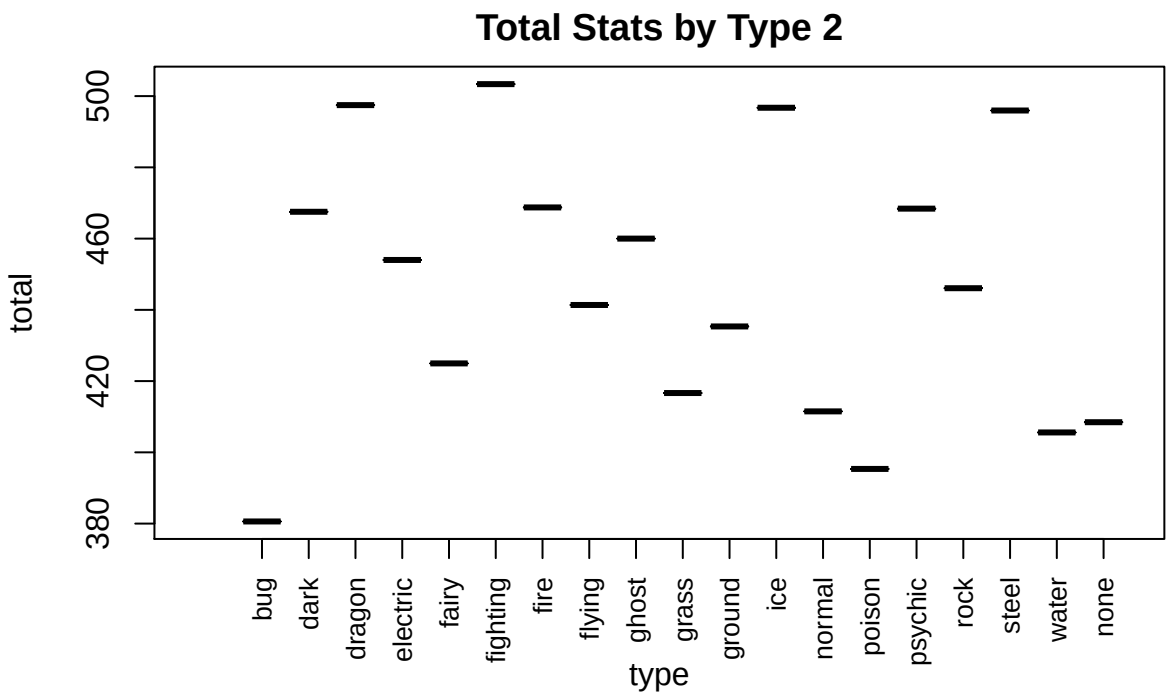
Data

As previously stated our initial analysis will focus just on base total, the total pool of a Pokémon's stats. Please note the following graphical representations of the data described in the Pokémon background:

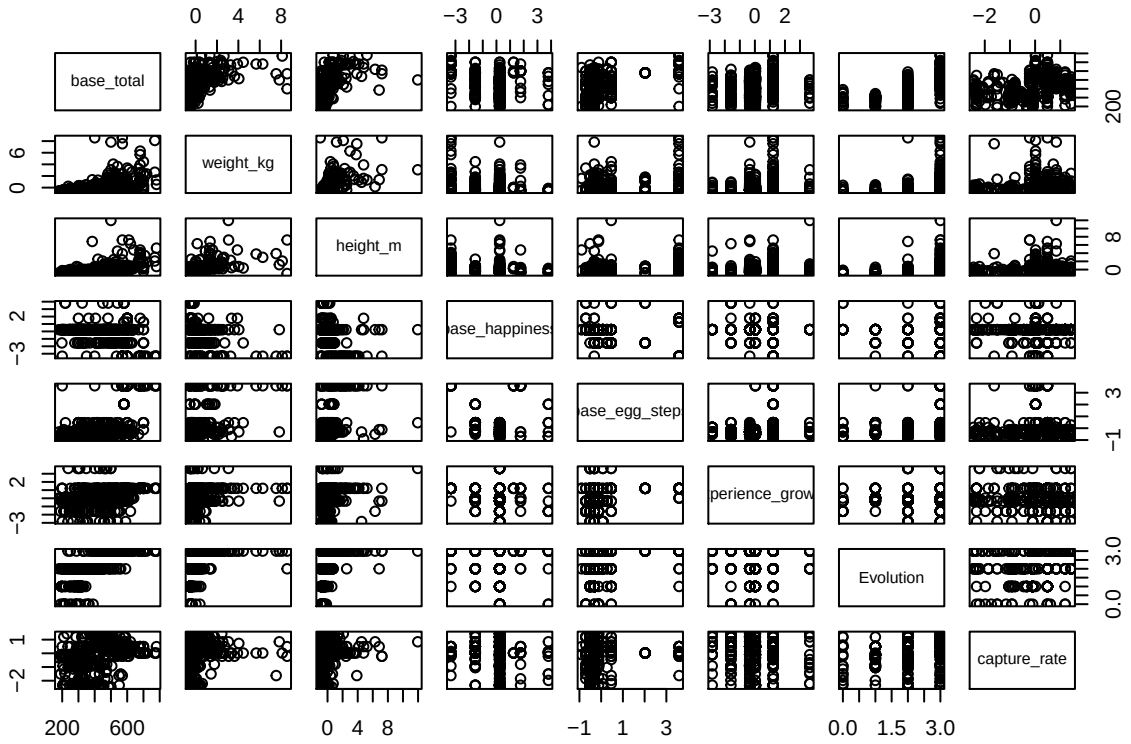

base_total
weight_kg
height_m
base_happiness
base_egg_steps
experience_growth
Evolution
capture_rate
is_legendary
generation
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



This is a test!This is a test! This is a test! This is a test! This is a test! This is a test! This is a test! This is a test! This is a test!



This is a test! This is a test! This is a test! This is a test! This is a test! This is a test!


base_total
weight_kg
height_m
ase_happines
ase_egg_step
perience_grow
Evolution
capture_rate