Outline

This is supposed to be the outline for the actual project report.

Introduction

Research Question

Which variables have an impact on the probability of a Pokemon being captured?

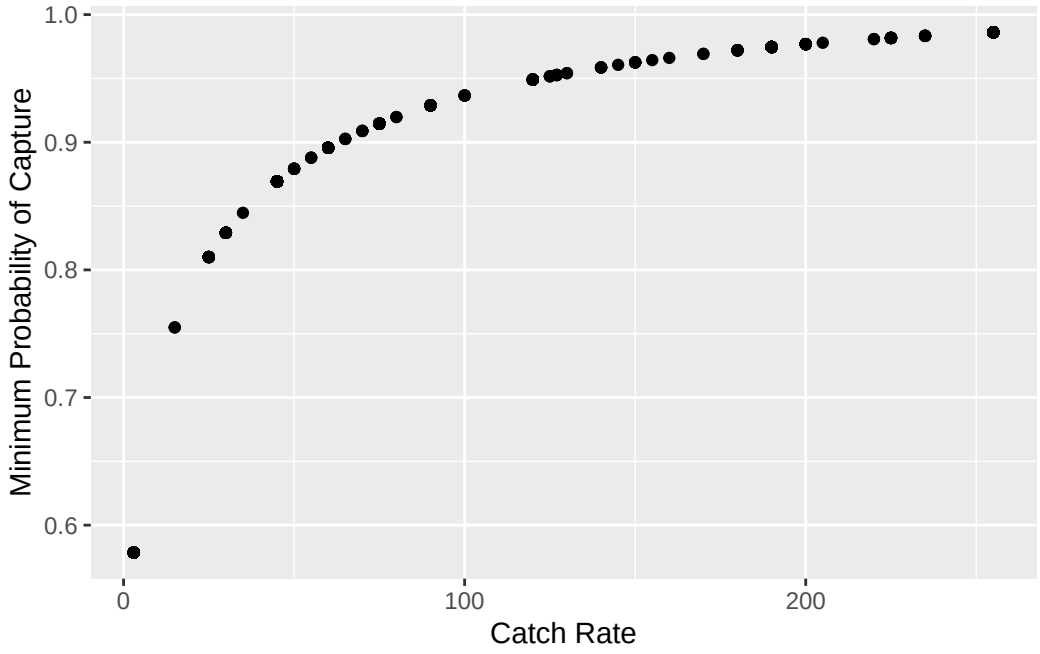
Goals

Now the actual probability of a Pokemon being captured is a function of health percentage, capture rate, and some modifiers. Given a fixed health percentage and fixed modifiers, the probability of capture p_c is

$$p_c = 1 - \left(1 - a r_c^{1/4}\right)^3$$

Where r_c is the capture rate, and a is some constant based on health percentage and modifiers.

In order to visualize the relationship, we can use the following plot.



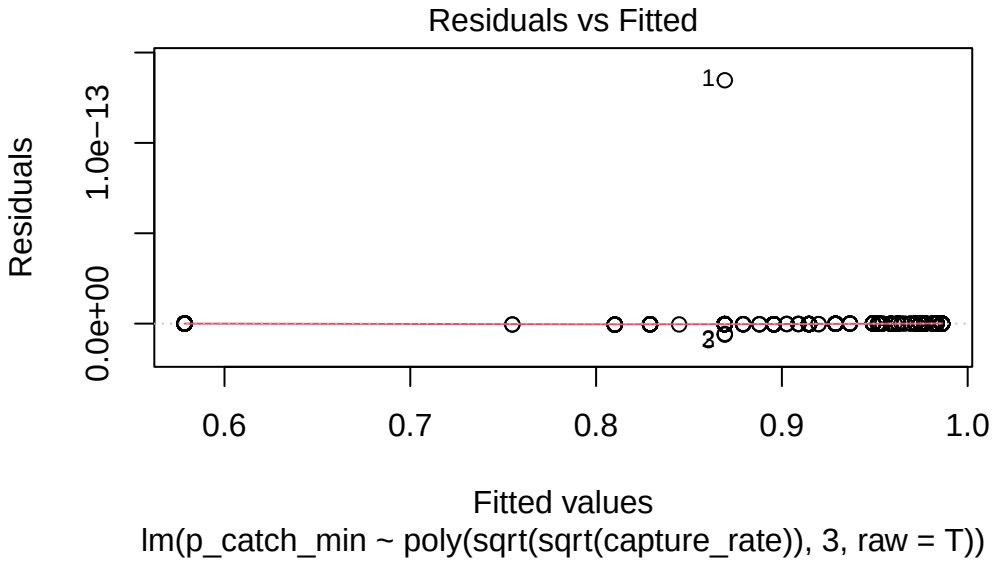
This can be shown in the following diagnostics for a model with the formula.

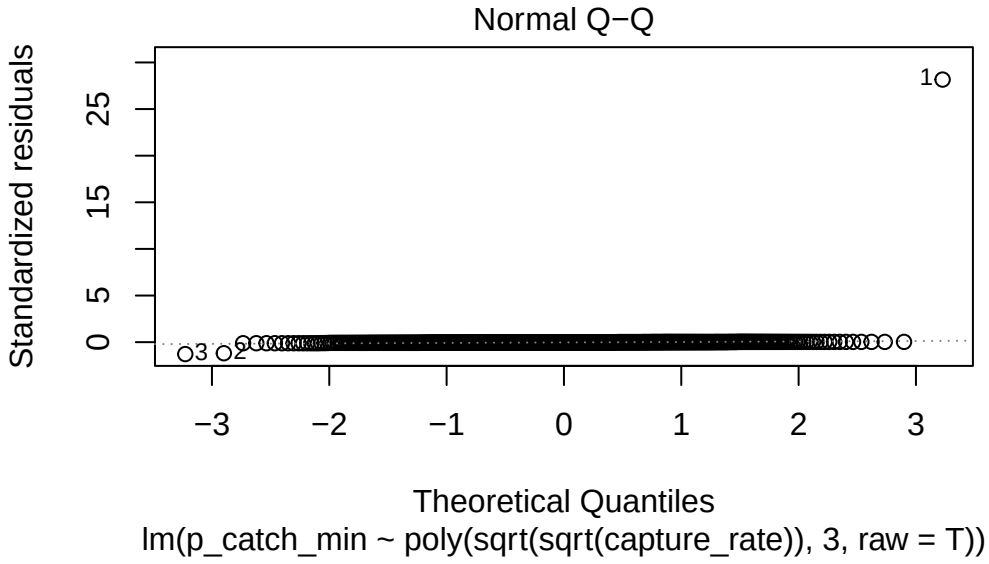
```
Call:
lm(formula = p_catch_min ~ poly(sqrt(sqrt(capture_rate)), 3,
    raw = T), data = pokemon)

Residuals:
       Min          1Q      Median          3Q         Max
-6.119e-15 -3.060e-16 -2.540e-16  2.500e-17  1.347e-13

Coefficients:
                                             Estimate Std. Error    t value
(Intercept)                                 7.536e-16  7.334e-15  1.030e-01
poly(sqrt(sqrt(capture_rate)), 3, raw = T)1  5.704e-01  9.579e-15  5.955e+13
poly(sqrt(sqrt(capture_rate)), 3, raw = T)2 -1.085e-01  3.845e-15 -2.821e+13
poly(sqrt(sqrt(capture_rate)), 3, raw = T)3  6.874e-03  4.822e-16  1.426e+13
                                            Pr(>|t|)
(Intercept)                                    0.918
poly(sqrt(sqrt(capture_rate)), 3, raw = T)1   <2e-16 ***
poly(sqrt(sqrt(capture_rate)), 3, raw = T)2   <2e-16 ***
poly(sqrt(sqrt(capture_rate)), 3, raw = T)3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.788e-15 on 796 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 1.159e+29 on 3 and 796 DF,  p-value: < 2.2e-16
```





Data Description

The dataset used was "The Complete Pokemon Dataset" from Kaggle user Rounak Banik, who scraped data from serebii.net on Pokemon up to Generation 7.

The dataset has 41 columns and 801 rows.

The relevent columns to this project are:

- `name`

  This is a column of strings. Each value is the official English name of the Pokemon.

- `capture_rate`

  This is a column of unsigned 8-bit integers used to calculate the probability of a capture.

Results

The model can be more effectively tested by splitting the data into test and train subsets.

After fitting a linear model with the training data and testing it, we get the following summary of the Residual Vector.

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.220e-16  0.000e+00  2.220e-16  1.268e-16  2.220e-16  3.331e-16
```

This seems to indicate that the model is extremely accurate. Now these models only estimate the minimum probabilities of capture for pokemon.

Another detail is that when fitting all models, one Pokemon had to be omitted. This was Minior. The reason Minior was omitted is because they have two phases. Each phase has it's own capture rate while the other stats remain the same. In its Meteorite Phase, Minior has a capture rate of 30, which is below the first quartile. When Minior is in its Core Phase, it has a capture rate of 255, which is the maximum value for capture rate. Since the capture rate can take either value, I decided to omit it.

Overall, this approximation seems to accurately predict the minimum probability of a successful capture. This is useful as the minimum value cannot decrease based on the health of the pokemon, with the exception of a faint. It can also be combined with an estimate for higher bound in order to give a helpful interval of the probability of capture.