Project 1

Ryan Fischbach

2/20/2021

Abstract

The ability to predict the height of Star Wars characters is incredibly important to our client, who owns a hotel and rents rooms of different heights. A model to help them to predict the heights of guests would allow them to accommodate different species and make the allocation of rooms easier. In this report, we discuss the process of estimating the height of an individual in centimeters using their species and mass in kilograms. Data was collected on different Star Wars characters and was used to build linear models to predict height. The steps in the process and comparison of results from different models are discussed.

Part 1: Data Cleaning

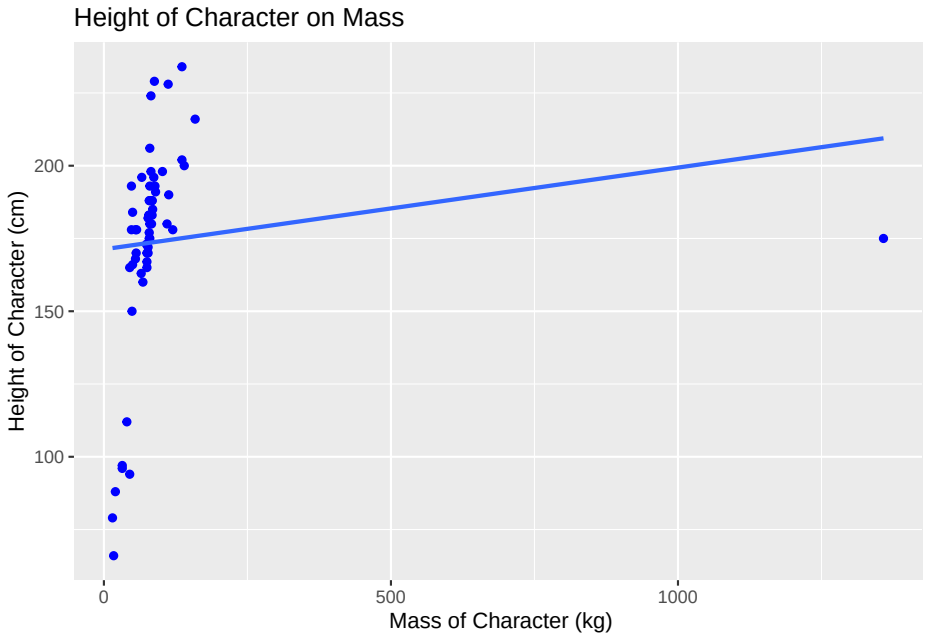
To transform this initial dataset into data that can be used to generate a model, the following steps were taken to clean the data.

1. First, the columns of interest were kept (predictors and target), with the rest removed.
2. Secondly, NA values were removed from this subset to facilitate model building. NA values did not provide any information that can be used by models.

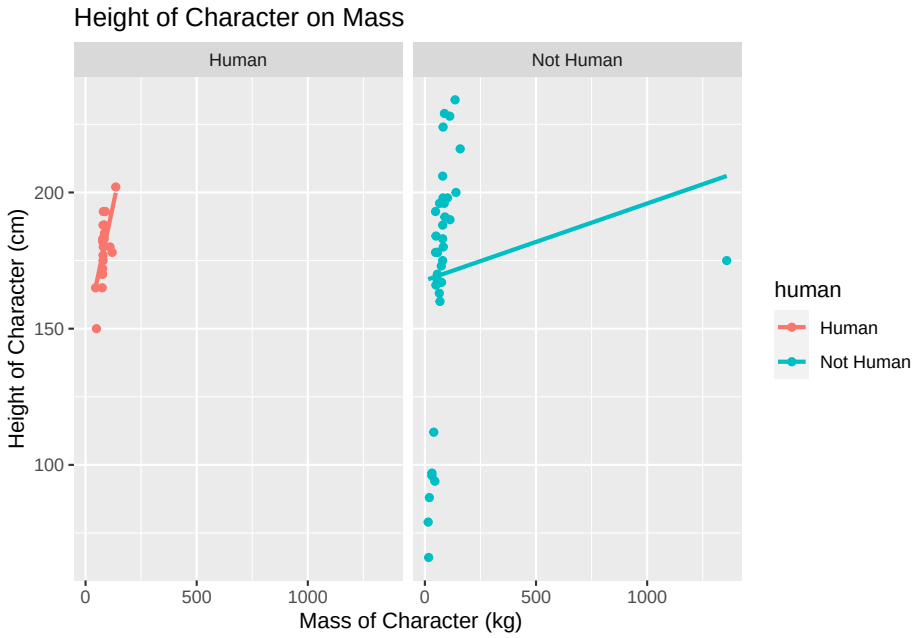
As a result of this data cleaning, the dataset had 59 observations, each with a mass, height, and species. This cleaned data set was then stored and could be used.

Part 2: Exploratory Data Analysis

The relationship between mass (kg) and height (cm) was first investigated to determine if mass could be a good predictor of height.



The scatter plot suggests that there is a strong, positive, linear relationship between mass and height. However, there appear to be groups of observations, potentially signifying the species of the characters being grouped because of similar mass and height characteristics. Thus, we will investigate the relationship between mass and height for humans and non-humans.



The relationships between mass and height for humans and non-humans are both positive and linear, however, they appear to be slightly different. This indicates that the relationship between mass and height varies for humans and non-humans, suggesting that adding a way to differentiate between the two groups during model building should be investigated.

Lastly, an outlier was identified and investigated. A singular character was greater than 1.5 times the Inter Quartile Range + Q3 according to a box plot. This outlier was the only member of this species present in the dataset and has a very high mass. Their height was not abnormal, coming in the clump of the

majority of characters. Thus, the observation was removed because the majority of rooms would likely be able to accommodate this species and the linear relationship between height and mass was skewed from this observation.

Part 3: Comparing Two Models

From the Exploratory Data Analysis and our client's request, we will be considering two models:

Model 1: $Height = \beta_0 + \beta_1 Mass + \epsilon$

Model 2: $Height_{human} = \beta_0 + \beta_1 Mass_{human} + \beta_2 Human + \epsilon$

where Human is a dummy variable (1 if human, 0 if not human).

Based on what was seen during Exploratory Data Analysis, using a model that allows the relationship between mass and height to be different for humans and non-humans makes more sense because it can encapsulate more species-specific information and make more informed predictions. This will ultimately result in less error during the evaluation process of the model and better outcomes for the client.

We will now employ two cross-validation techniques to assess the predictive accuracy of the two models.

Validation Approach

The first approach used to validate the two models was the validation approach. The validation approach takes the data and splits it into two distinct sets: training and test. The training set is used to train the model and the test set is kept separate to evaluate it. 80% of the data was used in the training set and 20% in the test set. Once the model was trained, the Root Mean Squared Error (RMSE) was calculated to determine the overall performance of the model.

This technique is appropriate because it facilitates "stealing" some of the training data to evaluate our model. Our model is being evaluated on test data that it has never seen before, allowing for a good evaluation of performance between the two models.

Model 1 produced a RMSE of 24.65 and model 2 produced a RMSE of 24.95. From this approach, we can assess that model 1 produces slightly better performance using the RMSE metric via this approach.

k Fold Cross-Validation

k Fold Cross-Validation was the next validation technique employed to evaluate the two models. k Fold Cross Validation involves creating k folds that act as test sets. We randomly divide the data into k groups of roughly equal size. We loop through all k sets, treating all but one set as the training set and the last as the test. This is repeated in all combinations.

This method eliminates some of the drawbacks of the validation approach, namely the reliance on the split of rows in training and test data. The training and test split in the validation approach matter a lot because the process is only performed once. Thus, the process can be more random. k Fold Cross-Validation splits the data and trains it on multiple different combinations of folds, partially reducing that randomness. It can be computationally intensive, however, this data set is relatively small so that is not an issue.

Model 1 had a RMSE of 10.08 using this approach while model 2 had a RMSE of 10.04. From this, we conclude that model 2 has slightly better performance according to the RMSE using k Fold Cross Validation.

Model Comparison

k Fold Cross Validation should be used to determine the best model because it is more comprehensive and helps eliminate randomness. This should help ensure the best performance while predicting the height of guests based on their mass.

Based on this result, I would recommend using the model that captures the relationship between height and mass for humans and not humans to predict the height of guests. The average error for model 1 using k Fold Cross Validation was 10.08 cm compared to 10.04 cm for model 2. In other words, model 2 was more accurate by 0.04 cm on average.

Part 4: Fitting the Model

Having produced the better performance using k Fold Cross-Validation, model 2 was trained with the entire dataset and produced the following estimated coefficients and model:

$\widehat{Height} = 103.7763 + 0.94Mass - 1.97Human$

We are 95% confident that the true slope for mass is in the range [0.72,1.14]. In other words, we are 95% confident that the true change in height is in the range [0.72,1.16] per kg of mass added. Additionally, we are 95% confident that the true slope of the human "dummy variable" is in the range [-15.03, 11.1]. In other words, we are 95% confident that being a human changes the height in cm of the prediction by [-15.03, 11.1].

Using this final combined model with a new observation with a mass of 10 kg, the model would predict a height of 113.16 cm. This creates a residual of -72.58 (40.6-113.16 cm), suggesting we predicted that the height of this new observation would be 73.85 cm more than it was.