Stat 220 Lab on Log and Polynomial Transformations

1 Introduction

You are part of a research team focusing on the biology and ecology of possums. The team has collected a dataset that includes multiple physical measurements from various possums. The objective is to provide a tool to predict the age of these creatures. Do we need to measure the possum in a variety of ways, or perhaps is there just one or two characteristics that can be enough to estimate age. Understanding the aging patterns could provide valuable insights into their lifecycle and how they interact with their environment.

Your role as a data scientist is to employ statistical techniques to build the best model possible. In this lab, you will exclusively explore the impact of log and polynomial transformations on predictive modeling.

Data Acquisition and Initial Exploration

Initial understanding of the dataset is crucial for any analytical tasks.

1. Download the dataset from the following link: `https://richardson.byu.edu/220/possum.csv`

2. Load the dataset into Python and perform a preliminary exploration to understand its structure.

Variable	**Description**
age	Age of the possum
hdlngth	Head length in cm
skullw	Skull width in cm
totlngth	Total length in cm
taill	Tail length in cm
footlgth	Foot length in cm
earconch	Ear conch length in cm
eye	Distance from medial canthus to lateral canthus of right eye in cm
chest	Chest girth in cm
belly	Belly girth in cm

Table 1: Data dictionary for the possum dataset.

Log and Polynomial Transformations

1. Explore a log transformation of the target variable

2. Apply log transformations to variables that exhibit non-linear relationships with age.

3. Apply polynomial transformations for variables that seem to have polynomial relationships with age.

4. Fit initial models using these transformed variables. Note the impact of the transformations on metrics like R^2 and p-values.

Model Selection and Validation

1. Develop multiple regression models, iteratively refining them based on p-values and out-of-sample metrics.

2. Use train-test split for out-of-sample validation.

3. Determine the best model based on these out-of-sample metrics.

Deliverables

There is one deliverables for this lab:

1. A well-commented Python script or notebook containing all the code to produce the requested results.

2. A recommendation of the final model with justification as to why it is the best model you discovered.