Linear Regression - Cumulative Lab

In this cumulative lab you'll perform a full linear regression analysis and report the findings of your final model, including both predictive model performance metrics and interpretation of fitted model parameters.

Objectives
You will be able to:

- Perform a full linear regression analysis with iterative model development
- Evaluate your final model and interpret its predictive performance metrics
- Apply an inferential lens to interpret relationships between variables identified by the model

Your Task: Develop a LEGO Pricing Algorithm

Business Understanding

You just got hired by LEGO! Your first project is going to be to develop a pricing algorithm to help set a target price for new LEGO sets that are released to market. The goal is to save the company some time and to help ensure consistency in pricing between new products and past products.

The main purpose of this algorithm is predictive, meaning that your model should be able to take in attributes of a LEGO set that does not yet have a set price, and to predict a good price. The effectiveness of your predictive model will be measured by how well it predicts prices in our test set, where we know what the actual prices were but the model does not.

The secondary purpose of this algorithm is inferential, meaning that your model should be able to tell us something about the relationship between the attributes of a LEGO set and its price. You will apply your knowledge of statistics to include appropriate caveats about these relationships.

Data Understanding

You have access to a dataset containing over 700 LEGO sets released in the past, including attributes of those sets as well as their prices. You can assume that the numeric attributes in this dataset have already been preprocessed appropriately for modeling (i.e. that there are no missing or invalid values), while the text attributes are simply there for your visual inspection and should not be used for modeling. Also, note that some of these attributes cannot be used in your analysis because they will be unavailable for future LEGO products or are otherwise irrelevant.

You do not need to worry about inflation or differences in currency; just predict the same kinds of prices as are present in the past data, which have already been converted to USD.

Requirements

1. Interpret a Correlation Heatmap to Build a Baseline Model

You'll start modeling by choosing the feature that is most correlated with our target, and build and evaluate a linear regression model with just that feature.

2. Build a Model with All Relevant Numeric Features

Now, add in the rest of the relevant numeric features of the training data, and compare that model's performance to the performance of the baseline model.

3. Select the Best Combination of Features

Using statistical properties of the fitted model, the ***sklearn.feature_selection*** submodule, and some custom code, find the combination of relevant numeric features that produces the best scores.

4. Build and Evaluate a Final Predictive Model

Using the best features selected in the previous step, create a final model, fit it on all rows of the training dataset, and evaluate it on all rows of the test dataset in terms of both r-squared and RMSE.

5. Interpret the Final Model

Determine what, if any, understanding of the underlying relationship between variables can be determined with this model. This means you will need to interpret the model coefficients as well as checking whether the assumptions of linear regression have been met.