Team Identification Number

16926

M3 Challenge 2023
Executive Summary

Title

Exec Summary

Keywords: Keywords, More Keywords

Contents

1 Q1: The Road Ahead

1.1 Defining the Problem

In Problem 1, we were tasked with producing a short-term predictive model for e-bike sales. More specifically, we were asked to develop projections for total sales volume 2 and 5 years into the future respectively.

This is a time-series forecast problem, where we predict a single variable - e-bike sales - based on a single input - time. We will take a general approach of regressing particular equations against the existing data. Multiple types of equations will be used and the most promising model will be chosen.

1.2 Assumptions

Assumption 1: There will be no major legislative changes, governmental campaigns and/or 'black swan' (i.e., highly unpredictable and consequential) world events that significantly impact the market for e-bikes within the next five years.
Justification: in practice, it is impossible to account for rare or extreme events within the constraints of a mathematical model; the implications of such events cannot be predicted with accuracy.

Assumption 2: The market for e-bikes in the European Union behaves comparably to that of the United Kingdom; therefore, British and European sales can be considered to be in direct linear proportion.
Justifications:

a) Of the data provided for European sales, several figures appear to include sales made in the UK (CITE EBICYCLES.COM). Therefore, UK consumer behaviour is partially accounted for even within the larger dataset.
b) E-bicycles have only begun gaining traction as a mode of transport in relatively recent years; as a result, UK-specific consumption data is largely unavailable to the public.
c) To a large extent, the UK and EU follow similar urban planning practices that include pedestrian walkability and bicycle access. In other terms, city layouts support the practical use of e-bikes. For this reason, population-scaled EU predictions can be considered appropriate substitutes for UK-specific predictions. By contrast, most American cities use car-centric design, frequently involving longer commute distances and poor bike access. This renders the United States hostile to the adoption of e-bikes in a way that the EU and UK are not. For this reason, we chose to exclude the US from our analysis, instead focusing on the UK and EU.

1.3 Variables

See table 1.1:

Variable	Definition
y_i	(actual) e-bike sales in year i
$\hat{y}_i$	predicted e-bike sales in year i
z	description

Table 1.1: Variables in the Model

1.4 Models

Linear

A linear regression is performed first due to its simplicity and ability to help pick more complex models.

Linear regression is an approach used to model a linear relationship between an independent variable x and a dependent variable y by finding the slope of the trend and initial value (y when x is 0). It is used to represent existing data and predict future values; linear models are used both for interpolation and extrapolation. In this case, the model will be fit to existing data and used to predict future sales of e-bikes. A linear growth function takes the general form of (1.1):

$$f(x) = \alpha x + \beta \tag{1.1}$$

where α is the coefficient, or growth rate; and β is the y-intercept, or initial value. The values of α and β are "optimized" using an algorithm to model a given dataset with the minimum "error".

One of the most common and simplest methods used to calculate the coefficient and intercept of the regression line is the Ordinary Least Square (OLS) *optimization* method. In short, OLS minimizes the Square-Error for each point against a given linear function by adjusting the function's parameters, which in the end produces optimal parameters for a equation in the form of a linear line of best fit.

The Square-*Error* function, which is what OLS *optimizes*, is simply a summation of the squares of the difference between actual values and predicted values, over all data points (1.2):

$$E = \sum (y - \hat{y})^2 \tag{1.2}$$

Because the predicted values $\hat{y}$ for a linear model is modelled as $\alpha x + \beta$, the Square-Error function for a linear model can be more specific (1.3):

$$E = \sum (y - (\alpha x + \beta))^2 \tag{1.3}$$

OLS calculates the values of α and β which minimize S in the summation above. Unlike the generic differential method described above, OLS is specialized for linear functions and can calculate the optimal parameters in one stop, using summation ratios. The coefficient α, or the linear trend of the dataset can be calculated with (1.4):

$$\alpha = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \tag{1.4}$$

where n is the number of data points.

After calculating the slope of the trend, the intercept β, is calculated by (1.5):

$$\beta = \bar{y} - \alpha \bar{x} \tag{1.5}$$

OLS was applied to the given data set to obtain the coefficient α and the y-intercept β - 222.6 and 446810, respectively - which corresponds to the following linear equation:

$$\hat{y}_i = 222.6i - 446810 \tag{1.6}$$

where i is the year.

Linear regression is useful in relation to the problem as it is simple to interpret and portray, allowing the prediction of data to be accurate during interpolation. However, if the data to be predicted is outside the range, i.e. predicting future e-bike sales, extrapolation may be inaccurate due to a false assumption of the trend. Furthermore, if the variables plotted provide a non-linear relationship, a linear regression line may inaccurately represent and predict values, which is the case in the data provided. Statistical error of the linear regression model against existing data shows a good but not perfect accuracy (Table 1.2).

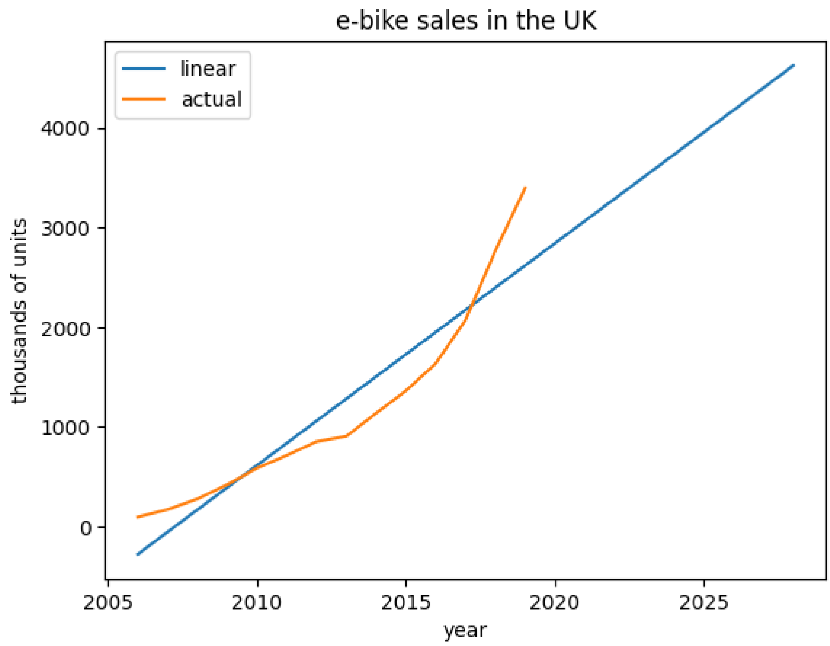

Figure 1.1: Regressed Linear Model

Metric	Value
MSE	108671
RMSE	329.65
MAE	268.82
PCC	0.9386
R^2	0.8810
CVAR	867138

Table 1.2: Linear Model Statistics

Exponential

The second approach decided upon was exponential regression, which models the non-linear relationship between an independent variable x and dependent variable y, where the rate of change of y with respect to x at a given point is proportional to the quantity itself. This choice was made based on visual indicators of the data given's trend, where the line graph exhibited a possible exponential curve. Exponential growth functions take the general form of (1.7):

$$f(x) = \alpha^x \tag{1.7}$$

where α is the exponential growth factor.

However, in order to fit such a model to arbitrary (non-normalized) values, two additional parameters have to be added to allow for displacement translations of the function on both axis. This gives:

$$f(x) = \beta(\alpha)^{x+a} + b \tag{1.8}$$

where a and b allows for offsets in the x and y axis, respectively.

An optimization can be made here - the equation can be rearranged to only have 3 parameters yet still be able to fit to any scale and offset of values:

$$f(x) = B^{b(x+a)} + c \tag{1.9}$$

where B, the base, can be any positive constant, and the function is parameterized by a, b, and c.

a and b performs a linear transformation on the input, while c performs a translation on the output. The equation in this form expresses in the relationship in purely in terms of translations. This reduction in parameters allows for a higher efficiency when regressing the function to the dataset programmatically. In the actual regression, 2 was used for the value of B.

Unlike linear regression, trying to calculate optimal parameters to an exponential model would technically require calculus concepts such as partial derivatives. However, a programmatic approach was taken, and the function was "blindly" (without accounting for its algebraic structure) regressed using gradient descent (Alg. 1) from derivative estimates. The `curve_fit()` optimizer from the Python SciPy library can blindly optimize any non-linear function to a dataset. It is able to optimize unknown functions by performing gradient estimates (1.10) of the error function with respect to the parameters using the basic definition of the derivative (at a given point, x):

$$G = \frac{\Delta f(x)}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{1.10}$$

where G is the gradient of function f at point x. Δx is set to a very small value to increase accuracy for sensitive functions.

After being able to calculate gradients of the error function at any point of the model function, a gradient descent algorithm (1) can be deployed to iteratively minimize the error function. The parameters are adjusted based the gradients of the error function:

$$\beta_j \longleftarrow \beta_j - \alpha \frac{E(x, \beta + \Delta\beta_j) - E(x, \beta)}{\Delta\beta_j} \tag{1.11}$$

where parameter β_j is adjusted based the error function E's gradient - the parameter changes in to the opposite direction of the gradient in order to find the minimum of the error function. α, the learning rate, is usually a very small value to prevent the parameters from changing too much at once.

Algorithm 1 Gradient Descent

repeat
- $\epsilon \leftarrow E(f(x, \beta))$ ▷ *E is the error function*
- $\gamma \leftarrow \frac{\Delta E(f(x,\beta))}{\Delta\beta}$ ▷ *calculate current gradient*
- $\beta \leftarrow \beta - \alpha\gamma$ ▷ *α is the learning rate*

until ϵ is sufficiently small ▷ *ϵ is the error or residuals*
return β as the optimal parameters

The SciPy curve fit optimizer was applied to the given data to obtain the following function 1.12. The exponential model was graphed alongside the actual values and the linear model for comparison in Figure 1.2.

$$\hat{C}_i = 2^{0.023381(i-1707.690634)} + 256.024002 \tag{1.12}$$

Visually, the exponential function better aligns with the actual values; the predicted values also seem to fit with the general trend, and this is confirmed by extremely good error metrics, as show in Table 1.3. An advantage of exponential regression as a predictive model is that

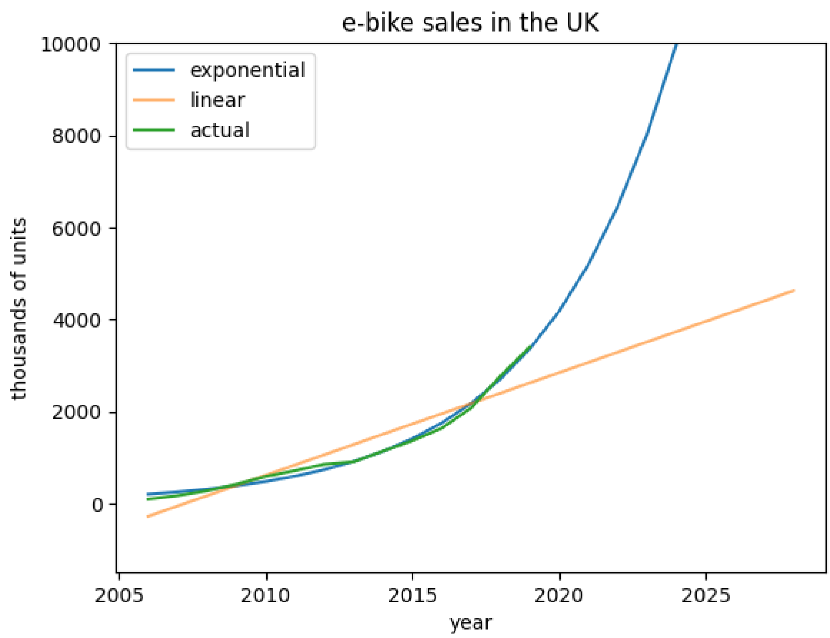

Figure 1.2: Regressed Exponential Model

Metric	Value
MSE	0.48061
RMSE	0.69326
MAE	0.56981
PCC	0.99973
R^2	0.99945
CVAR	890.45

Table 1.3: Exponential Model Statistics

it provides high quality forecasts, which increase the accuracy of predicted values during interpolation. However, a key drawback is that a large data set is necessary to carry this method out, as a reasonable amount of continuity is needed to accurately predict future values, especially during extrapolation.

1.5 Results

Results

1.6 Model Revision

Model Revision

1.7 Evaluation

Strength 1: asdf

Strength 2: asdf

Weakness 1: asdf

1.8 Technical Computing

Technical computing

2 Q2: Shifting Gears

2.1 Defining the Problem

Defining the Problem

2.2 Assumptions

Assumption 1: Statement **Justification**: blah blah

Assumption 2: Statement **Justification**: blah blah

Assumption 3: Statement **Justification**: blah blah

2.3 Variables

See table 2.1:

Variable	Definition
x	description
y	description
z	description

Table 2.1: Variables in the Model

2.4 The Model

Model

2.5 Results

Results

2.6 Model Revision

Model Revision

2.7 Discussion

Strength 1: asdf

Strength 2: asdf

Weakness 1: asdf

2.8 Sensitivity Analysis

Sensitivity Analysis

2.9 Technical Computing

Technical computing

3 Q3: Off the Chain

3.1 Defining the Problem

Defining the Problem

3.2 Assumptions

Assumption 1: Statement **Justification**: left as an exercise to the reader

Assumption 2: Statement **Justification**: blah blah

Assumption 3: Statement **Justification**: blah blah

3.3 Variables

See table 2.1:

Variable	Definition
x	description
y	description
z	description

Table 3.1: Variables in the Model

3.4 The Model

Model

3.5 Results

Results

3.6 Model Revision

Model Revision

3.7 Discussion

Strength 1: asdf

Strength 2: asdf

Weakness 1: asdf

3.8 Sensitivity Analysis

Sensitivity Analysis

3.9 Technical Computing

Technical computing

4 References

4.1 Bibliography

4.2 Program Code

Result data generated:

```
text data stuff
```

Python program code:

```
        # pass
```