University of Auckland
Department of Accounting & Finance

FINANCE 788: Research Essay

Author: Connor McDowall
Supervisor: Dr Paul Geertsema

August 24, 2021

Abstract

Acknowledgements

Paul Geertsema

Declaration of Contribution

Contents

List of Figures

List of Tables

1 Introduction

2 Literature Review

Insert Research Intent

3 Research Intent

Insert Research Intent

4 Theory

4.1 Return predictability

Return predictability underlies asset pricing theory. **Insert**

4.2 Modelling, loss, and optimisation

We summarize the theory surrounding predictive modelling, loss functions, and optimisation algorithms. These functions train models by comparing predictions to realized observations using optimisation algorithms to minimize the loss function. We examine a linear model as our predictive model (1). Mean square error (2) and Gradient Descent (GD) are basic examples of a loss function and optimisation algorithm, respectively.

$$\hat{y} = mx_i + b \tag{1}$$

$$f(y, (mx_i + b)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + b))^2 \tag{2}$$

Firstly, gradient descent takes the partial derivatives of the loss function, with the respect to the parameters in our predictive model. In our example, equations 3 and 4 are the partial derivatives for the mean square error loss function.

$$\frac{\partial f(y, (mx_i + b)}{\partial m} = \frac{1}{n}\sum_{i=1}^{n} -2x_i(y_i - (mx_i + b))^2 \tag{3}$$

$$\frac{\partial f(y, (mx_i + b)}{\partial b} = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - (mx_i + b))^2 \tag{4}$$

Secondly, the algorithm explores epochs, using a learning rate to update parameters to move in the opposite directions of the partial derivatives until settling in a local minima. This extrema is the optimisation of the loss function, quantifying the accuracy of the predicative model. Ordinary Least Squares (OLS) regressions is an extension of the linear model prevalent in asset pricing.

4.3 Ordinary Least Squares (OLS)

The OLS regression is the most prominent statistical model in asset pricing theory. Rosenfeld n.d. contributes an OLS summary. The composition of the true OLS model includes four components. Firstly, $\mathbf{X}$, an n x k matrix of k independent variables for n observations. Secondly, $\mathbf{y}$, an n x 1 vector of observation on the dependent variable. Thirdly,

ϵ, an n x 1 vector of unexplained error. Lastly, θ, a k x 1 vector of parameters to be estimated.

$$y = X\theta + \epsilon \tag{5}$$

4.3.1 Criteria for estimation

The criteria to obtain the parameter estimate ($\hat{\theta}$) relies on the minimisation of the sum of squared residuals (6). We highlight the observed residuals (e) are distinct from unexplained disturbances (ϵ). Equation 7 derives residuals by taking the difference between observations based on parameter estimates.

$$\sum e_i^2 \tag{6}$$

$$e = y - X\hat{\theta} \tag{7}$$

Expanding the quadratic $e^T e$ after substituting in equation 7 leads to the alternative expression of the sum of squared residuals in equation 8. Minimizing the sum of square residuals requires taking the partial derivative of equation 8 with respect to the estimated parameters (equation) using matrix differentiation (9). It is imperative X has full rank where all vectors in the matrix are linearly independent, validating both the presence of a positive definite matrix and minimum.

$$e^T e = y^T y - 2\hat{\theta}^T X^T y + \hat{\theta}^T X^T \hat{\theta} X \tag{8}$$

$$\frac{\partial e^T e}{\partial \hat{\theta}} = -2X^T y + 2X^T X\hat{\theta} = 0 \tag{9}$$

We find the expression for the Ordinary Least Squares (OLS) estimator (13) after rearranging equation 9 to normal form, utilizing inverse matrices to form identity matrices, and simplifying.

$$2X^T X\hat{\theta} = 2X^T y \tag{10}$$

$$(X^T X)^{-1}(X^T X)\hat{\theta} = (X^T X)^{-1} X^T y \tag{11}$$

$$I\hat{\theta} = (X^T X)^{-1} X^T y \tag{12}$$

$$\hat{\theta} = (X^T X)^{-1}(X^T y) \tag{13}$$

$$\tag{14}$$

Therefore, we can use the OLS estimator to make predictions with OLS (15).

$$\hat{y} = X^T \hat{\theta}$$

4.3.2 Properties of OLS Estimators

There are six key properties in addition to the satisfaction in minimizing the summation of squared residuals.

1. The residuals are uncorrelated with the observed values of X i.e., $X^T e = 0$.
2. The sum of the residuals is zero i.e., $\sum e_i = 0$.
3. The sample mean of the residuals is zero i.e., $\bar{e} = \frac{\sum e_i}{n} = 0$.

4. The regression hyperplane passes through the means of observed values i.e., $\frac{e}{=}\frac{y-X\theta}{n} = 0$. Since $\bar{e} = 0$ assumed, it is implied $\bar{y} = \bar{x}\bar{\theta}$.
5. The residuals are uncorrelated with the predicted y i.e., $\hat{y} = X\hat{\theta}$, $\hat{y}^T e = (X\hat{\beta})^T e = b^T X^T e = 0$
6. The mean of $\hat{y}$ for the sample will equal the mean of the y.

4.3.3 The Gauss-Markov Theorem

However, OLS makes Gauss-Markov assumptions about the true model to make inferences regarding β from $\hat{\beta}$. The intention of the Gauss-Markov Theorem, conditional on the below assumptions, states the OLS estimator is the best linear, unbiased, and efficient estimator:

$$y = x\beta + \epsilon \tag{15}$$
$$E[\epsilon|X] = 0 \tag{16}$$
$$E(\epsilon\epsilon^T|X) = \Omega = \sigma^2 I \tag{17}$$
$$\epsilon|X \ N[0, \sigma^T I] \text{ (hypothesis testing)} \tag{18}$$

- X is an n x k matrix of full rank
- X must be generated randomly, or fixed, by a mechanism uncorrelated to disturbances.

Equation 16 implies $E(y) = X\beta$ as no observations of the independent variables convey any information about the expected values of the disturbances. Equation 17 captures homoskedasticity and no autocorrelation assumptions. Additionally, The theory underlying Ordinary Least Squares informs the common practice in minimising of the sum of least squares when evaluating prediction performance. The mathematical tractability, in accordance with the aforementioned assumption, frame our thinking surrounding the derivation of custom loss functions.

4.4 Hedge portfolios

Table 1

Variable	Description	$MSE(y, \hat{y})$	$HP(y, \hat{y})$
θ	Est/Train	$\hat{\theta}_{MSE}$	$\hat{\theta}_{HP}$
λ	Validation	$\hat{\lambda}_{MSE}$	$\hat{\lambda}_{MSE}$

Table 1: Objective (MSE: Mean Square Error, HP: Hedge Portfolio)

The formation of hedge portfolios rely on monotonic functions. These functions both preserve or reverse a given ordered set. We rank the cross-sections of portfolio returns using variations in monotonic functions to assign weights and form hedge portfolios.

$$R(y_{i,t}) \tag{19}$$

The ranking function (R($y_{i,t}$)) and thresholds (u,v) form subsets of long and short portfolios.

$$L = \{y_{i,t}|R(y_{i,t}) \geq u\} \tag{20}$$
$$S = \{y_{i,t}|R(y_{i,t}) \leq v\} \tag{21}$$
$$0 < u < 1 \tag{22}$$
$$0 < v < 1 \tag{23}$$
$$u > v \tag{24}$$

These truth sets inform the construction of time-series hedge portfolios. The first set of time-series hedge portfolio equations assumes equal weighting in long and short portfolios through dividing each subset (L,S) by their cardinality.

$$H_t = \frac{1}{|L|}\sum_{i \epsilon L} y_{i,t} - \frac{1}{|S|}\sum_{i \epsilon S} y_{i,t} \tag{25}$$
$$\tag{26}$$

Our aim is to re-configure the loss function to maximise the objective functions when deriving custom loss functions. Subsequently, this enables the derivation of objective functions ex-post transaction costs.

5 Data

Expand: Dataset implies, use this dataset (Jensen, Kelly, and Pedersen, 2021)

6 Methodology

6.1 Project organisation

GOCPI adopted Data Science best practice, as described by Wilson et al Wilson et al., 2016. Although these practices are mostly reserved for data science projects, their principles are suitable for product development and version control. All data and results were saved regularly and reproducible. The retention of data in all forms received high levels of attention. Project files were synched continuously to Google Drive Google LLC, 2020. Git Linus Torvalds, 2020 was used to manage version control for GOCPI's source code, data, documentation and results. Git stores a complete history of versions using Git hashes. These hashes are strings unique to each state of the publicly available GOCPI repository[1]. Git hashes enabled the discretisation of GOCPI's development over time, enabling the accessibility and recollection of all previous states given a unique git hash. This functionality enabled reproducibility, error correction and the ability to revert to previous models.

6.1.1 Version Control

Git, hosted by GitHub, provided a comprehensive set of version control technologies. These technologies provided a range of benefits. Firstly, Git is excellent at providing and supporting collaborative functionalities. The master version of a project is accessible for all who have access to the repository. Each contributor could create custom copies of branches through pull requests on the master branch. Contributors could commit changes to custom branches and push these changes to the master branch through push requests. The product manager could review these push requests, approving suitable requests to integrate changes to the master branch. Collaborative efforts were possible with commit messages describing the contributions from each contributor. This project had one contributor. Git ensured the histories of code, work and authors are stored. The descriptive nature of the commit log ensured an accurate journal is kept.

6.1.2 Folder Structure

GOCPI maintained the file folder structure recommended in Wilson et al Wilson et al., 2016. Project organisation was paramount as the modelling of energy systems involves integrating a range of optimisation models, data files and documents. Wilson et al's recommendations were appropriate as data science projects require similar organisational rigor. Subsequently, file management and structure was most efficient and comprehensive. **GOCPI** is the root directory of this project and contains several sub directories: bin, data, doc, src and results. The **bin** sub directory contained external scripts and compiled programmes related to the GOCPI project. The **data** sub directory contained all raw data associated with the project. This data included energy statistics, energy

[1] https://github.com/CMCD1996/GOCPI

balance datasets, partitioned geographies, standardised optimisation models and TIMES modelling frameworks. The **doc** sub directory stored GOCPI's user guides, academic resources, research reports and project deliverables. The **results** sub directory contained the output from optimisation simulations and processed data to display on dashboards and websites to inform investment and policy decisions. The **src** sub directory stores the source code for preparing raw data, partitioning sets of geographies with varying granularities and the GOCPI python package available to download using PyPI[2] and install using pip[3]. All files were continuously backed up using Google Drive.

6.1.3 Python

Python 3.7 was the primary coding language for the GOCPI project. GOCPI's objective is to enable any user to design and model their own energy system to inform investment and policy decisions. The intention is to empower users to discuss energy investment and policy decisions made by public and private parties. Additionally, GOCPI intends to reduce misinformation regarding energy policies and help assess the feasibility of meeting the International Energy Agency's Sustainable Development Scenario Agency, 2019. Python is omnipresent, widespread in software development. Python's language design makes the language highly productive and simple to use. Python can hand off computationally straining tasks to C/C++ and has first-class integration capabilities with these two languages. The language also has a very active and supportive community Medium, n.d. In addition, Python is the most popular coding language on the planet defined by the PYPL PopularitY of Programming Language Index. As at August 2020, Python had 31.59% of all language tutorial search instances on Google PYPL, n.d. Python has many useful packages for creating the GOCPI package such as NumPy, Scikit-learn, os, csv and Pandas. Programming is quick due to Python's dynamic nature. The language is also open-source with no cost. Subsequently, Python was the best language to ensure the GOCPI model is accessible for many users to use and extend.

6.1.4 Package Management

The Anaconda package management platform for Python Anaconda, Inc., 2020 was the chosen coding environment. Anaconda is a well defined, free platform with known versions of python packages such as matplotlib, numpy and pip. The use of this environment ensured both reproducibility and consistency across infrastructure. Although this project required no collaboration, the use of Anaconda will inform future developers on how to manage collaborative processes, especially for packages which are less well-maintained. Anaconda allows you to create custom environments which was necessary for creating scalable linear optimization problems to express energy systems. Pip is Python's default package manager and is included in the Anaconda package. Pip was used to install and update packages for python not available on Anaconda such as twine and the custom GOCPI package developed for this project.

6.1.5 Excel

It is important users are comfortable with using the GOCPI model. Energy modelling can be quite complex. The modelling process must be transparent to inform users how to build

[2]https://pypi.org/
[3]https://pypi.org/project/pip/

their own models. Excel is ubiquitous across academic and professional communities. Excel's omnipotence makes the software well-suited for describing the components of the GNU Mathprog energy system model. The **GOCPI OseMOSYS Structure.xlsx** file describes the sets, paramaters, constraints and objective function of a scalable energy system model. The User may toggle statement sets, parameters and constraints to adjust the complexity of the model. The model file was imported to a text file. However, data related to these energy systems was stored using Python dictionaries, lists and NumPy arrays. This Python formulation was later transcribed to a text file. Excel is best for two dimensional variables or data stored in Codd-Boyce relational databases Arenas, 2009. The majority of parameters in energy systems were three or more dimensions. Therefore, Excel was not suitable to store these parameters. Python dictionaries, lists and NumPy arrays were preferred alternatives.

6.1.6 IBM ILOG CPLEX Optimization

The OseMOSYS methodology (see **??**) translates energy systems into linear programming problems. A solver was required to optimise these user-defined energy systems. The IBM ILOG Optimization Studio International Business Machines Corporation, 2020, more commonly known as CPLEX, was chosen to be this solver. CPLEX solves very large linear programming problems using the Barrier Interior-point method Potra and Wright, 2000 or primal/dual variants of the Simplex Method Bronson and Costa, 2009. GOCPI's user-defined energy systems could be scaled up to model very large systems, creating large linear programming problems.

The IBM ILOG CPLEX Optimization Studio has an interface with the Python language based on a C programming interface. Subsequently, Python APIs were available to run the CPLEX solver when installed either locally or on a cloud service. The python packages are **cplex** and **docplex**. The cplex package contains classes for accessing CPLEX for the Python programming language. The Cplex class is the most important class in this package as provides methods for creating, modifying, querying, or solving optimisation problems. Docplex also enables the formulation of new linear programmes where one creates the model, defines the decision variables, sets the constraints and expresses the objective function. The user uses docplex to solve the linear programme on a local solver. Alternatively, the model can be solved on a private cloud using Decision Optimisation on Cloud service through the provision of a service url and personal API key. The CPLEX Python APIs were most attractive as provided the user with a powerful commercial solver in an accessible format.

There is a caveat to the use of the CPLEX solver. The IBM ILOG CPLEX Optimization Studio is commercial by nature and requires a license to use. Fortunately IBM have the IBM Academic Initiative IBM, n.d.-a, granting students access to commercial software for free. This commercial nature creates accessibility issues for users who are not enrolled at an academic institution or can afford to pay for the software. Accessibility issues caused by the need for commercial solvers must be addressed to enable the distribution of the GOCPI product.

6.1.7 IBM Watson Machine Learning Service

The IBM CPLEX Optimisation Cplex python API is suitable for smaller models that can be solved locally. As the model increases in complexity, the docplex Python API

did enable the ability to solve larger linear programmes. Unfortunately, IBM phased out the docplex Python API by incorporating the Decision Optimisation on Cloud services into the IBM Watson Machine Learning cloud services IBM, n.d.-b. This change occurred during September 2020. This service uses IBM Cloud to access assets through credentials, create model deployments in IBM's servers and execute jobs to solve models. The model deployments must be Python-based models with jobs specifying a payloads containing input data and output formats.

6.1.8 PyPI

PyPI[1] is the Python Package Index, a repository of software for the python programming language. This repository helps you find and install software developed by the Python community who have decided to share their work. The GOCPI package is distributed from this platform to enable as many as possible the ability to model their own energy systems to inform and question energy policy and investment. Enter command: **pip install GOCPI** in the terminal to install the package using pip package management software.

6.1.9 Code Style

The GOCPI project was developed as the GOCPI package. All development code is organised within this package. The PEP8 style for Python Code was the formatting style for development code Guido Van Rossum and Coghlan, 2001. All code was formatted with **yapf**, a formatter maintained by Google to format Python files. Standardised formatting is important as makes the code easy to read, helps optimise the code and promotes consistency. Docstrings and commenting were most important in documentation. A docstring is a Python inline comment. Each class and function has an unique docstring, a one sentence description of the function, inputs with data types and types of outputs. The Google style docstring was most appropriate because of it's readability, ease to write and consistency with the Google Style Guide. Additionally, automated documentation generators (**pdoc3, Sphinx** etc.) can parse this format to create documentation. This self-consistent code style facilitated best practise maintenance and enabled reproducibility.

6.1.10 Infrastructure

GOCPI creates scalable energy system optimisation models with complexity size dependent. Computations either took place locally on a 128 GB, four core Apple MacBook Pro or remotely using a cloud service.

6.2 Documentation

The GOCPI project is well documented to keep an accurate record of key design decisions. The commit history described in 6.1.1 was the most important form of document. Other explicit documentation methods were applied to supplement this commit history. These methods, in addition to in-code documentation, include project updates and meeting minutes nested within a project logbook.

[1] https://pypi.org/

6.2.1 Project updates

Project updates were recorded as itemized lists. Each item is a brief description of the work completed during that day, week or month. Items include, but are not limited to, completing GOCPI submodules, researching energy system statistics, building websites or writing sections of this research report. These updates were pivotal to exploring new options, monitoring progress and making decisions to drive forward development. For example, the decision to adopt the OseMOSYS methodology in favour of the TIMES modelling methodology. Project updates were transcribed to the project logbook held in this project's research compendium.

6.2.2 Meeting minutes

Project meetings took place for half an hour once a week. These meetings included discussions on energy markets, modelling methodologies, project progress and key design decisions. The minutes from these meetings accompanies project updates in the project logbook nested within the research compendium.

7 Results

8 Discussion

9 Conclusion